SCarborSNV: Efficient phylogeny-aware single nucleotide variant detection for single cells

Christopher Oldham

January 2019

Abstract

Ongoing somatic mutations during cancer development lead to genetically distinct subclonal populations of cells within a tumour, each with a distinct subset of acquired mutations. These subclonal populations genetically diverge as new mutations occur and are subject to Darwinian selection pressures. This leads to a complex intra-tumour heterogeneity, the subclonal architecture of which is important for understanding cancer evolution and developing individualised therapies. Some approaches have used bulk DNA sequencing coupled with advanced clustering techniques to attempt to tease out this structure. Recent advances in single-cell DNA sequencing (SCS), however, have allowed new approaches such as Monovar and SCIΦ to examine this heterogeneity directly, despite the inherent low quality of the SCS data. We here present a new probabilistic algorithm, SCarborSNV, which we expect will efficiently call point mutations using aligned SCS sequencing data from a sample of multiple cells. After calling variant loci using a detailed prior, SCarborSNV uses a neighbour joining algorithm to reconstruct a phylogeny which is used to genotype individual cells. We will compare SCarborSNV with existing methods on simulated and real data and expect to show that SCarborSNV performs competitively in calling single cell genotypes.

1 Introduction

Cancer is caused by somatic genetic mutations and as a tumour grows it develops further mutations. Some of these somatic mutations are clonal or truncal mutations, which have occured in a cell that is a common ancestor of all extant tumour cells and thus affect all the cells in the tumour. Other mutations are subclonal, only a portion of the extant tumour cells share a lineage with the cell in which the mutation occured and as such only this fraction contain

A a tumour grows, some of its cells develop further somatic mutations such as single nucleotide variations (SNVs) and copy number changes. The cells descended from these further mutated cells therefore form genetically distinct subclonal populations which are subject to a Darwinian evolutionary process. At the time of tumour biopsy, then, the cells removed will represent a genetically heterogeneous sample from the leaves of a subclonal tumour phylogeny, as well as any healthy cells accidentally introduced. While bulk DNA sequencing has been coupled with

Neighbour joining, biological data, etc. ploidy changes, state of the art, infinite sites(SNV and SNP don't occur on same locus), prior values, dynamic programming, massive parallelism (avoid read collisions by having workers do lots of work between reading loci.), assumption of no polyploidy. Which parameters are reestimateable? all?

2 Methods

Our method of identifying SNVs begins with computing the likelihood of observing the sequencing data for each cell and each locus independently, given all possible underlying genotypes. From these likelihoods and phylogeny aware priors we determine the posterior probability that each locus contains a mutation:

pooling the data across cells to select only those loci of interest. The calculation of priors uses asymptotic aproximations and dynamic programming to further improve efficiency. We then use a neighbour joining algorithm to quickly reconstruct a phylogeny of the single cells. (MAYBE:) We then use this tree to re-estimate the frequencies of certain in vivo, in vitro and in silico (?) artifacts in the data, such as loss of heterozygosity (LOH) and allelic dropout (ADO). We then reconstruct a cell tree with these updated values using neighbour joining and derive from it a mutation tree (how??). We perform a search on this tree to impute missing mutations within cells (?) and reject unlikely mutations as false positives... This needs some work. Kim and Simon or Le and durbin

2.1 Preprocessing and parallelism

Compute priors for m and store somewhere where all processes can read. (shared memory, file)

Create unique identifiers for each cell. Each cell should be in own aligned BAM file.

Pile up aligned BAMs on loci.

- Use Python's PySam pileup to iterate through columns (loci).
- Workers pick up batches of 1000? loci. Since sites assumed to be independent (note: should they be?) synchronicity is not important if locus positions and cellIDs are tracked.
- Workers process batch through to genotype calculation so not all are on I/O at once.
- Return locus objects with cells' read and qual info at that locus.

Mark or discard indels and low/no coverage.

Impute welltype from germline data.

- If germline VCF file provided, impute here
- If no germline VCF, if cell consensus (likelihood threshold?) different than reference, impute
- Optionally use dbSNP as prior for germline vs somatic mutation

2.2 Cell genotype likelihoods

Calculate genotype likelihoods for each cell j at each locus i. We assume independence between sites.

Homozygous genotypes

- Let $g \in \{0, 1, 2\}$ be the unphased genotype of a locus designated by the number of non-reference alleles. For homozygous genotypes (that is, $g \in \{0, 2\}$) We generally assume reads to be independent:

$$P(D_{ij} \mid g) = \prod_{k=1}^{n} P(d_{ijk} \mid g) \tag{1}$$

  Note $D_{ij} = (\vec{r}, \vec{e})$, where $\vec{r}$ are the n reads at this nucleotide and this cell. $\vec{e}$ are the associated probabilities of read error, derived from the phred quality scores.

- Marginalizing on sequencing error:

$$P(d_k \mid g) = P(r_k,\, se \mid g) + P(r_k,\, \neg se \mid g)$$

- Since errors can occur during amplification or sequencing, we model an "intermediate allele", denoted β that is amplified from the original nucleotide with some probability of error [1]. Trivially:

$$P(r_k,\, \neg se \mid g) = P(r_k \mid \neg se,\, g)(1 - e_k) = P(\beta_k = r_k \mid g)(1 - e_k)$$

We similarly see:

$$P(r_k,\, se \mid g) = P(r_k \mid se,\, g)e_k$$

Furthermore:

$$P(r_k \mid se,\, g) = P(r_k \mid \beta_k \neq r_k,\, se,\, g)P(\beta_k \neq r_k \mid se,\, g)$$

Assuming (?) an error in sequencing the intermediate allele could produce any of the other three alleles with equal probability we find $P(r_k \mid \beta_k \neq r_k,\, se, g) = 1/3$. Since the amplification of β is unaffected by sequencing $P(\beta_k \neq r_k \mid se,\, g) = P(\beta_k \neq r_k \mid g) = 1 - P(\beta_k = r_k \mid g)$. We therefore have:

$$P(r_k, se \mid g) = e_k \frac{1}{3}\left[1 - P(\beta_k = r_k \mid g)\right]$$

- Finally the likelihood $P(D_{ij} \mid g)$ for cell at a locus for a homozygous genotype is:

$$P(D_{ij} \mid g) = \prod_{k=1}^{n}\left[(1-e_k)P(\beta_k = r_k \mid g) + e_k\frac{1}{3}(1 - P(\beta_k = r_k \mid g))\right] \tag{2}$$

Heterozygous genotypes and allelic dropout

- For the heterozygous case, we must account for the possibility of allelic dropout (ADO) [1,2]. Therefore:

$$P(D_{ij} \mid g = 1) = P(D_{ij}, \text{ADO} \mid g = 1) + P(D_{ij}, \neg\text{ADO} \mid g = 1)$$

Letting P_{ADO} be the probability of an a dropout event, this expands to:

$$P(D_{ij} \mid g = 1) = P_{ADO}P(D_{ij} \mid \text{ADO},\, g = 1) + (1 - P_{ADO})P(D_{ij} \mid \neg\text{ADO},\, g = 1)$$

In the result of an allelic dropout from a heterozygous locus, only one allele will remain after the amplification process and hence the likelihood $P(D_{ij} \mid \text{ADO},\, g = 1)$ will resemble the homozygous case. We assume allelic dropout can affect either allele with equal probability and hence:

$$P(D_{ij} \mid \text{ADO},\, g = 1) = \frac{1}{2}P(D_{ij} \mid g = 0) + \frac{1}{2}P(D_{ij} \mid g = 2)$$

For the case without allelic dropout, the form of the likelihood is identical to the homozygous case:

$$P(D_{ij} \mid \neg\text{ADO},\, g = 1) = \prod_{k=1}^{n}\left[(1-e_k)P(\beta_k = r_k \mid g = 1) + e_k\frac{1}{3}(1 - P(\beta_k = r_k \mid g = 1))\right]$$

2.3 Mutated site priors

We now focus on the prior probability of the total alternate allele count being σ at the locus under consideration: $P(\sum_j g_{ij} = \sigma) = P(\sigma)$. The majority of sites will not include a somatic SNV (sSNV); we say that any site has a prior probability λ of having a somatic SNV, which is set to 0.0001 by default [1, 2].

$$P(\sigma) = P(\sigma \mid \text{sSNV})\lambda + P(\sigma \mid \neg\text{sSNV})(1 - \lambda) \tag{3}$$

Any given sample of single cells will only represent some subtree of a full cell phylogeny. As such, when considering the case where there is a sSNV at the locus we can further break down the prior into the case where the sSNV is ancestral to all sampled cells and the case where where the SNV occurs within the subtree rooted at the most recent common ancestor (MRCA) of the cells sampled. We denote the case where the mutation occurs within this subtree as SNV_T.

$$P(\sigma \mid \text{sSNV}) = P(\sigma \mid \text{SNV}_T)P(\text{SNV}_T \mid \text{sSNV}) + P(\sigma \mid \text{sSNV},\, \neg\text{SNV}_T)(1 - P(\text{SNV}_T \mid \text{sSNV})) \tag{4}$$

Ploidy changes

It is well known that many tumor cells may exhibit aneuploidy or chromosomal abnormaities [3, 4]. For simplicity, we will disregard polyploidy and focus only on the case where loci become haploid. This sort of mutation can result in lost information regarding SNVs, as a loss of heterozygosity can lead to a locus being read as homozygous [2]. Note that this is an in vivo effect, distinct from allelic dropout which occurs in vitro during DNA ampification. We will model such occurences as a sudden switch to homozygosity, as we cannot reliably distinguish diplod homozygosity from haploidy in the genomic SCS data, which already has significantly uneven coverage and depth [5]. Let H be the event that the locus under examination has become haploid, and H_T be the case that this mutation has occured within subtree rooted at the MRCA of all sequenced cells.

$$P(\sigma \mid \text{SNV}_T) = P(\sigma \mid \text{SNV}_T, H)P(H) + P(\sigma \mid \text{SNV}_T, \neg H)(1 - P(H)) \tag{5}$$

We initially set the value of $P(H)$ to .09 (See Appendix A). In the simplest case, we consider the prior probability of an alternate allele count of σ given a mutation occured within the subtree and the locus remained diploid across all sampled cells. Assuming infinite sites, in such a scenario mutations would only be heterozygous.

$$P(\sigma \mid \text{SNV}_T, \neg H) = \begin{cases} \frac{2m-1}{2(m-1)} T(m, \sigma) & 0 < \sigma < m \\ 0 & \text{else} \end{cases}$$

Where $T(m, \sigma)$ is the prior developed by Singer, Kuipers et al. that assumes a mutation may occur on any branch of the sampled subtree with equal probability.

$$T(a, b) = \frac{\binom{a}{b}^2}{(2b-1)\binom{2a}{2b}} \tag{6}$$

Now let us consider the case where both a sSNV and a haploid event have both occured at a locus. Since the haploid mutation may have occured in the subtree or ancestral to the subtree, we model these cases seperately.

$$P(\sigma \mid \text{SNV}_T, H) = P(\sigma \mid \text{SNV}_T, H_T)P(H_T \mid H) + P(\sigma \mid \text{SNV}_T, H, \neg H_T)(1 - P(H_T \mid H))$$

SNV and loss of heterozygosity within the subtree

In the first scenario described by Equation 2.3 both a point mutation and a ploidy change have occured within the sequenced subtree. This can be split into four further subcases (Figure 1).

(a) Case 1 (b) Case 2 (c) Case 3 (d) Case 4

Figure 1: SNV and haploid event both within the subtree. (a) Point mutation happens before haploid event and mutated allele is dropped. (b) Point mutation happens before haploid event and reference allele is dropped. (c) Haploid event occurs before point mutation. Since haploid cells are modelled as becoming homozygous diploids, this case leads to only even values of alternate allele count. (d) In this case the point mutation and haploid event do not occur in the same lineage. We ignore this case as the haploid event does not affect the alternate allele count.

Considering the above three cases where a ploidy change in the subtree affects the locus alternate allele count, it is twice as likely that the point mutation should occur before the haploid event (cases 1 and 2), compared with the other temporal ordring (case 3). Therefore $P(\text{case 1 or 2}) = 2/3$ and $P(\text{case 3}) = 1/3$. This is because the cells before a haploid event, being diploid, have twice the chance of having a point

mutation at a locus than the haploid descendants of such a mutation. We also assume the refence and alternate alleles have an equal chance of being dropped in a loss of heterozygosity.

$$P(\text{case 1}) = P(\text{case 2}) = P(\text{case 3}) = 1/3$$

Continuing to assume that both a point mutation and a haploid event have occured within the sequenced subtree at the locus in question, we now have

$$P(\sigma \mid \text{SNV}_T, H_T) = \frac{1}{3}\left[P(\sigma \mid \text{case 1}) + P(\sigma \mid \text{case 2}) + P(\sigma \mid \text{case 3})\right]$$

To examine these probabilities we will use the function $T(a, b)$ described above, which given a subtree with a leaves gives the probability of a mutation affecting b of those leaves. For case 1, the loss of heterozygosity effectively deletes all alternate alleles from the cells sharing a lineage with the haploid event, and so

$$P(\sigma \mid \text{case 1}) = \frac{1-2m}{2(1-m)} \sum_{a-h=\sigma} T(m, a)T(a, h) \qquad 1 \leq a < m,\ 1 \leq h \leq a$$

While we include one normalization constant, excluding the case of an ancestral point mutation, we do allow the case where a point mutation and a loss of heterozygosity happen on the same branch of the phylogeny. Hence the support for σ in case 1 is $[0, m)$. Since we model a haploid event as a sudden switch to homozygosity, the observed allele count for case 2 is the number of cells affected by the heterozygous point mutation (a) added to the number of cells affected by the loss of heterozygosity (h).

$$P(\sigma \mid \text{case 2}) = \frac{1-2m}{2(1-m)} \sum_{a+h=\sigma} T(m, a)T(a, h) \qquad 1 \leq a < m,\ 1 \leq h \leq a$$

For case 2, the support is [2, 2m-2]. In case 3, only even allele counts can be produced as all cells carrying the point mutation are haploid, which we model as being homozygous mutated. The possible values of σ are $1 \leq \sigma \leq 2m - 2$.

$$P(\sigma \mid \text{case 3}) = \frac{1-2m}{2(1-m)} \sum_{2a=\sigma} T(m, h)T(h, a) \qquad 1 \leq h < m,\ h \geq a$$

Haploid subtree

Referreing back to Equation (2.3), we must determine the prior probability of an alternate allele count σ at a locus given that a point mutation occured within the sequenced subtree and a haploid event occured ancestral to the subtree. This would lead to all sampled cells being haploid at this locus, therefore allowing only even allele counts.

$$P(\sigma \mid \text{SNV}_T, H, \neg H_T) = \begin{cases} \frac{2m-1}{2(m-1)} T(m, \frac{\sigma}{2}) & 2 \mid \sigma,\ 0 < \sigma < 2m \\ 0 & \text{else} \end{cases}$$

Clonal and subclonal mutations

We have so far considered the case where sSNVs have been subclonal: they may affect some of our sampled cells and not others. There is some probability however that a sSNV at a given locus may be due to a mutation in a cell ancestral to all sampled cells. The majority of these ancestral mutations will affect all tumour cells: so-called clonal, truncal or public mutations [4, 6, 7]. If the sample of single cells is small enough, however, it could be the case that a subclonal mutation is common to all cells sampled.

$$P(\text{ancestral} \mid \text{sSNV}) = P(\text{clonal}) + P(\text{ancestral} \mid \text{subclonal})(1 - P(\text{clonal})) \tag{7}$$

To find the probability of a subclonal mutation affecting all sampled cells, we assume tumour subclones follow a neutral evolutionary model such that subclonal mutant allele frequencies follow a power law distribution [6]. Using an IID model for tumour cell sampling, the probability that all m cells are from a subclone

with cellular frequency $2f$ (allelic frequency $= f$) is $(2f)^m$. Similar to Williams et al. we define a probability density function for the allelic frequency of subclonal mutations proportional to the inverse of the allelic frequency.

$$P(f) = k\left(\frac{1}{f} - 2\right) \tag{8}$$

where k is a normalization constant. We define the support of $P(f)$ for subclones as $[10^{-8}, 0.5]$, as a frequency of 10^{-8} is on the order of affecting single cells and an allelic frequency of 0.5 corresponds to clonal mutations [8]. We find a value of k by integrating $P(f)$ over this support. If a subclonal mutation affects all sampled cells, then all these cells must be from the same subclone.

$$P(\text{ancestral} \mid \text{subclonal}) = \int_f P(\text{all in subclone} \mid f)P(f)\,df = \int_{10^{-8}}^{0.5} (2f)^m k\left(\frac{1}{f} - 2\right)\,df \tag{9}$$

For large enough samples ($m > 30$) the probability that all cells were sampled from a single subclone (Equation 9) becomes negligible. Since it is only a function of m these values are pre-computed for efficiency. The empirically estimated probability that any given mutation is clonal is set at $P(\text{clonal}) = 0.51$ (see appendix A) [4, 7, 9]. Having developed the probability that any given somatic mutation is ancestral to all sampled cells, we have also found the probability that a mutation has occured within the sampled subtree.

$$P(\text{SNV}_T \mid \text{sSNV}) = P(H_T \mid H) = 1 - P(\text{ancestral} \mid \text{sSNV})$$

Ancestral sSNVs

Referring Equation (4) we must also determine the prior for σ in the case that the sSNV is ancestral to all cells. Without LOH this would always result in $\sigma = m$, however we must again consider haploid events both within and ancestral to the sampled subtree.

$$P(\sigma \mid \text{sSNV}, \neg\text{SNV}_T) = P(\sigma \mid \text{sSNV}, \neg\text{SNV}_T, H)P(H) + P(\sigma \mid \text{sSNV}, \neg\text{SNV}_T, \neg H)(1 - P(H)) \tag{10}$$

If there is no haploid event, there will be no LOH and the ancestral sSNV will be heterozygous across all cells.

$$P(\sigma \mid \text{sSNV}, \neg\text{SNV}_T, \neg H) = \begin{cases} 1 & \sigma = m \\ 0 & \text{else} \end{cases}$$

In the case of a haploid event, we must consider the cases where the whole subtree is haploid and when the haploid event happens within the subtree. We also consider the alternate and reference alleles to be dropped with equal probability as above.

$$P(\sigma \mid \text{sSNV}, \neg\text{SNV}_T, H) = (1 - P(H_T \mid H))P(\sigma \mid H, \neg H_T) + P(H_T \mid H)P(\sigma \mid H_T, \text{SNV}, \neg\text{SNV}_T)$$

If both the somatic SNV and the haploid event are ancestral to the sequenced subtree, the cells will either all be haploid reference or all be haploid alternate with equal probability. Since we model haploid cells as homozygous diploid this results in only alternate allele counts of 0 or $2m$.

$$P(\sigma \mid \text{SNV}, \neg\text{SNV}_T, H, \neg H_T) = \begin{cases} \frac{1}{2} & \sigma = 0, 2m \\ 0 & \text{else} \end{cases}$$

If there is an sSNV ancestral to the sampled cells but a haploid event occured within the sampled subtree we may have any alternate allele count from 1 to $2m - 1$ depending on where in the phylogeny heterozygosity was lost and which allele was dropped.

$$P(\sigma \mid \text{SNV}, \neg\text{SNV}_T, H_T) = \begin{cases} \frac{1}{2}T(m, \sigma - m) & \sigma > m \\ \frac{1}{2}T(m, m - \sigma) & m > \sigma \end{cases}$$

2.4 Welltype site priors

We have so far considered the prior probability of alternate allele counts at a locus given that a somatic SNV has occured at that locus. The majority of loci, however, will be unaffected by sSNVs, although may still contain germline point mutations. Referring back to Equation (3) we must consider the prior probabilities of σ for sites without sSNVs. Note, however, that such a site may still be affected by anneuploidy.

$$P(\sigma \mid \neg\text{sSNV}) = P(\sigma \mid \neg\text{sSNV}, H)P(H) + P(\sigma \mid \neg\text{sSNV}, \neg H)(1 - P(H)) \tag{11}$$

In the case where there is no sSNV and no anneuploidy, we simply assume Hardy-Weinberg equilibrium, with a germline mutation rate of μ. We set the value of μ relatively high at 0.1 to reduce false positive errors.

$$P(\sigma \mid \neg\text{sSNV}, \neg H) = \begin{cases} \mu^2 & \sigma = 2m \\ 2\mu(1-\mu) & \sigma = m \\ (1-\mu)^2 & \sigma = 0 \\ 0 & \text{else} \end{cases}$$

Continuing to assume HWE for the germline genotype, anneuploidy will only affect the alternate allele count for a heterozygous germline genotype. Here again we assume either allele may be dropped with equal probability.

$$P(\sigma \mid \neg\text{sSNV}, H) = \begin{cases} \mu^2 + \mu(1-\mu)(1 - P(H_T \mid H)) & \sigma = 2m \\ \mu(1-\mu)P(H_T \mid H)\frac{2m-1}{2(m-1)}T(m, m-\sigma) & 0 < \sigma < m \\ 0 & \sigma = m \\ \mu(1-\mu)P(H_T \mid H)\frac{2m-1}{2(m-1)}T(m, \sigma-m) & m < \sigma < 2m \\ (1-\mu)^2 + \mu(1-\mu)(1 - P(H_T \mid H)) & \sigma = 0 \end{cases}$$

2.5 Variant candidate site calling

To 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 [10]. We therefore must reject loci where the posterior probability of mutation is low. For a given locus i:

$$P(SNV_i \mid D_i) = 1 - P\left(\sum_{j=1}^{m} g_{ij} = 0 \mid D_i\right) = 1 - P(\sigma = 0 \mid D_i) \tag{12}$$

Using Bayes' formula:

$$P(\sigma \mid D_i) = \frac{P(D_i \mid \sigma)P(\sigma)}{\sum_{\sigma'=0}^{2m}[P(D_i \mid \sigma = \sigma')P(\sigma')]} \tag{13}$$

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 (3). 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

$$P(D_i \mid \sigma_i) = \sum_{\vec{G}} P(D_i \mid \vec{G})P(\vec{G} \mid \sigma_i) \tag{14}$$

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 (14) 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 (14) 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:

$$P(D_i \mid \sigma_i) = \binom{2m}{\sigma_i}^{-1} \sum_{g_1=0}^{2} \sum_{g_2=0}^{2} \cdots \sum_{g_m=0}^{2} \delta((g_1, \ldots g_m), \sigma_i) \left[\prod_{j=1}^{m} \binom{2}{g_j} P(D_{ij} \mid g_j) \right] \tag{15}$$

As has been done previously, we can employ a dynamic programming approach to compute these likelihoods for σ from cell genotype likelihoods [1, 2, 10]. If we let $F(k, l)$ be the subproblem objective given by

$$F(k,l) = \begin{cases} \sum_{g_1=0}^{2} \sum_{g_2=0}^{2} \cdots \sum_{g_k=0}^{2} \delta((g_1, \ldots g_k), l) \left[\prod_{j=1}^{k} \binom{2}{g_j} P(D_{ij} \mid g_j) \right] & 0 \leq l \leq 2k \\ 0 & \text{else} \end{cases} \tag{16}$$

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

$$F(k,l) = F(k-1,l)P(D_{ik} \mid g_k = 0) + 2F(k-1,l-1)P(D_{ik} \mid g_k = 1) + F(k-1,l-2)P(D_{ik} \mid g_k = 2) \tag{17}$$

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

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

The values for $F(k, l)$ are memoized in an array and the likelihood given in Equation 15 can be given by

$$P(D_i \mid \sigma_i) = \frac{F(m, \sigma_i)}{\binom{2m}{\sigma_i}} \tag{18}$$

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 (13).

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

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 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 [2]. This more efficient approach results in an overall asymptotic complexity of $O(nm^3 \log(m))$ [2]. The Monovar algorithm has an overall asymptotic complexity of $O(nm^3)$ [1].

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 13. 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:

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

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:

$$d = -\frac{3}{4} \log \left(1 - \frac{4}{3}\overline{p}\right) \tag{19}$$

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 [11].

2.7 Genotyping single cells

Assuming the tree created above is accurate, we now seek to infer genotypes from this phylogeny so as to overcome errors and noise associated with low coverage SCS data. We first determine weights for attaching point mutation and different types of LOH events to different edges of the tree, and then use these weights to determine genotype probabilities for each cell.

2.7.1 SNV weights

To begin with, we compute the probabilities of all descendents of each node having the same genotype: $\pi_0(e), \pi_1(e)$ and $\pi_2(e)$ are the probabilities that all descendents of e are homozygous reference, heterozygous and homozygous alternate respectively. These values are taken to be

$$\pi_g(e) = \prod_{\{j : c_j \succ e\}} P(g_j = g)$$

where $c_j \succ e$ indicates that the j^{th} cell is below e in T, and $P(g_j = g)$ are the posterior probabilities calculated in Equation (13). We also compute one more values, $\pi_\mu(e)$, defined as the probability that all descendents of e have genotype 1 or 2. These four values can be computed recursively in $O(m)$ time by multiplying the corresponding values from the two branches directy beneath each branch. In the case of a point mutation but no loss of heterozygosity, the weight given to attaching a point mutation at edge e is given by:

$$W(S_e) = \frac{\pi_\mu(e)\,[\pi_0(\rho)/\pi_0(e)]\,P(S_e)}{\sum_{e' \in E} \pi_\mu(e')\,[\pi_0(\rho)/\pi_0(e)]\,P(S'_e)} = \frac{d_e \pi_\mu(e)/\pi_0(e)}{\sum_{e' \in E} d_{e'} \pi_\mu(e')/\pi_0(e')} \tag{20}$$

where ρ represents the root edge and hence $\pi_0(\rho)/\pi_0(e)$ is a product of probabilities over only those cells that do not descend from e. The prior probability of an edge containing a mutation is taken to be the normalized edge length:

$$P(S_e) = \frac{d_e}{\sum_{e'} d_{e'}}$$

Weights For Loss of Heterozygosity

Idea: only cases 1 and 2 are complex. Case 3 can be modeled as a haploid full tree and case 4 is ignored. For cases 1 and 2 we can then work out conditional weights assuming mutation above the attachment points.

For a loss of heterozygosity events, we calculate weights in a similar way. Referring to Figure **??** above these weights are calculated for cases 1, 2 and 3 in the following way:

$$W^{(1)}(S_{e_1}, L_{e_2}) = \frac{\pi_0(e_2)\left[\pi_1(e_1)/\pi_1(e_2)\right]\left[\pi_0(\rho)/\pi_0(e_1)\right]P(S_{e_1})P(L_{e_2})}{\sum_{e'_1 \in E-E_l}\sum_{e'_2 \succeq e'_1}\pi_0(e'_2)\left[\pi_1(e'_1)/\pi_1(e'_2)\right]\left[\pi_0(\rho)/\pi_0(e'_1)\right]P(S_{e'_1})P(L_{e'_2})}$$

$$W^{(2)}(S_{e_1}, L_{e_2}) = \frac{\pi_2(e_2)\left[\pi_1(e_1)/\pi_1(e_2)\right]\left[\pi_0(\rho)/\pi_0(e_1)\right]P(S_{e_1})P(L_{e_2})}{\sum_{e'_1 \in E-E_l}\sum_{e'_2 \succeq e'_1}\pi_2(e'_2)\left[\pi_1(e'_1)/\pi_1(e'_2)\right]\left[\pi_0(\rho)/\pi_0(e'_1)\right]P(S_{e'_1})P(L_{e'_2})}$$

and

$$W^{(3)}(S_e) = \frac{\pi_2(e)\left[\pi_0(\rho)/\pi_2(e)\right]P(S_e)}{\sum_{e' \in E-E_l}\pi_2(e')\left[\pi_0(\rho)/\pi_2(e')\right]P(S'_e)}$$

Genotyping Cells

With all of these values computed for each of the edges in the tree, we can then begin a depth first traversal of the tree keeping track of genotype probabilities at every node. For the starting point we say the root node has genotype probabilities $P(g = 0) = 1$ and $P(g = 1) = P(g = 2)$. If node n_1 is the direct ancestor of n_2 separated by edge e we define the relations:

$$P(g_{n_2} = 0) = P(g_{n_1} = 0) * ((1 - W(S_e)) + W(S_e) * P(L_e) * \dots$$

$$P(g_{n_2} = 1) = P(g_{n_1} = 0) * W(S_e)(1 - P(L_e)) + P(g_{n_1} = 1)(1 - P(L_e) * (W^{(1)} + W \dots))$$

$$P(g_{n_2} = 2) = P(g_{n_1} = 0) * W(S_e) * P(L_e) + \dots$$

Other stuff: normalize probs if need, either call max prob or filter.

2.8 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? Overall question: is $O(nm^3)$ good enough at all stages? (No!) asymptotically only beats sciϕ by $\log(m)$... Ofc NJ is $O(m^3)$ but not nm^3 since you build a single tree from info on all sites! So my alg could be bounded by $O(nm^2)$?? Computing distances in $O(nm^2)$... Could have an $O(nm^2)$ algorithm!

References

[1] Hamim Zafar, Yong Wang, Luay Nakhleh, Nicholas Navin, and Ken Chen. Monovar: single-nucleotide variant detection in single cells. *Nature methods*, 13(6):505, 2016.

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

[3] Ruli Gao, Alexander Davis, Thomas O McDonald, Emi Sei, Xiuqing Shi, Yong Wang, Pei-Ching Tsai, Anna Casasent, Jill Waters, Hong Zhang, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature genetics*, 48(10):1119, 2016.

[4] Serena Nik-Zainal, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

[5] Tal Nawy. Single-cell sequencing. *Nature methods*, 11(1):18, 2013.

[6] Marc J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature genetics*, 48(3):238, 2016.

[7] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B Alexandrov, Jose MC Tubio, Elli Papaemmanuil, Daniel S Brewer, Heini ML Kallio, Gunilla Högnäs, Matti Annala, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353, 2015.

[8] Ugo Del Monte. Does the cell number 10^9 still really fit one gram of tumor tissue? *Cell Cycle*, 8(3):505–506, 2009.

[9] Shinichi Yachida, Siân Jones, Ivana Bozic, Tibor Antal, Rebecca Leary, Baojin Fu, Mihoko Kamiyama, Ralph H Hruban, James R Eshleman, Martin A Nowak, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319):1114, 2010.

[10] Si Quang Le and Richard Durbin. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome research*, 21(6):952–960, 2011.

[11] 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 Initial Parameter tuning

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