PyClone : Software For Inferring Cellular Frequencies From Allelic Count Data

by Andrew Roth

[1]Graduate Program in Bioinformatics, University Of British Columbia
[2]Shah Lab, Department of Molecular Oncology, British Columbia Cancer Agency
e-mail: andrewjlroth@gmail.com

Introduction

The problem to be addressed here is that of inferring the fraction of cells in the population containing a mutation, which I will refer to as the *cellular frequency*. I will assume we are given information about the frequency of variant of B-alleles in the bulk population, and also some possibly incorrect information about the genotype of cells with the mutation. The key challenge is to go from the observed allelic frequency to the cellular frequency. Note that cellular frequency and genotype of a mutation both contribute to the observed allelic frequency.

The PyClone software implements a probabilistic model to estimate the frequency at which mutations occur in a population of cells. PyClone works by first converting next generation sequence data allelic abundance data. Next information about copy number (CN) and loss of heterozygosity (LOH) at the mutation are used to elicit prior beliefs about the genotype of cells with the mutation. Using the allelic abundance data and prior information a probabilistic model is applied to infer the *cellular frequency* of the mutation, that is the fraction of cells in the population containing the mutation.

When multiple mutations are present in a single sample, PyClone analyses them jointly under the assumption mutations tend to co-occur in cells. Given this assumption there should be groups of mutations at the same cellular frequency. Thus another output given by the PyClone software is the probability a pair of genes co-occur in the same mutated cell (more accurately co-occur at the same cellular frequency, however the most parsimonious way for this to occur is if the mutations are in the same cell).

Notation

In what follows we assume that we have aligned sequence data. For each position of interest this data is summarised by counting how many reads contain a given nucleotide from the set $\Sigma = \{A, C, T, G\}$. For simplicity we further assume that there is a reference genome used and that there are only two alleles at a given position, the reference allele A$\in\Sigma$ and the variant allele B$\in\Sigma$. For each position, i, we reduce read data to count data, a^i, the number of reads with nucleotides matching this reference allele and b^i the number of reads with nucleotides matching the variant allele.

With this formalism a diploid genome can have one of three possible genotypes, AA, AB, BB, at each position. Cancer genomes are not strictly diploid, so we need to consider an extended set of possible genotypes. For each position the genotype can be fully specified by two value. The number of copies of that position, c^i, and the number reference alleles in the genotype r^i. For example the genotype AAAB would have $c^i = 4$ and $r^i = 3$.

In the absence of any other confounding factors, the frequency of reference alleles we observe at a given position should be $f^i = \frac{r^i}{c^i}$. Due to sampling and sequencing error the observed reference allelic frequency $\hat{f}^i$ will generally not be exactly equal to f^i. However, as depth of coverage increases $\hat{f}^i$ should converge to f^i.

Sequencing Error

Sequencing error means that in general $\hat{f}^i$ will differ slightly from f^i. If the errors are random, then for genotypes where $r^i \neq 0$ or $r^i \neq c^i$ the errors should cancel out and this effect should not be noticeable. However, in the two exceptional cases where $r^i = 0$ or $r^i = c^i$ we assume there will be a small random deviation, $\varepsilon \ll 1$, in observed frequency associated with the error rate of the sequencing technology. That is when $r^i = 0$ we should expect $f^i = \varepsilon$ and when $r^i = c^i$ then $f^i = 1 - \varepsilon$.

Heterogeneity

We define heterogeneity to mean there are two or more populations of cells with different genotypes at position i. Heterogeneity at a locus means that f^i is no longer well defined, the $\hat{f}^i$ will in general not directly inform us about the underlying genotypes.

For example we can imagine a simple example where 50% of the cells have the genotype AA and 50% have the genotype BB. Then at high depth, we expect $\hat{f}^i$ to converge to $0.5 \cdot f_{\mathrm{AA}} + 0.5 \cdot f_{\mathrm{AB}}$, where with some abuse of notation we use f_g to be the allelic frequency associated with genotype g.

If we assume there are J sub-populations at site i we can label the genotypes as using c_j^i, r_j^i, f_j^i to indicate the copy number, number of reference alleles and associated reference allele frequency in the j^{th} sub-population. Note the if we have $r_j^i = 0$ or $r_j^i = c_j^i$, for all j, the heterogeneity becomes unobservable since f_j^i will be ε or $1-\varepsilon$ for all populations.

Again an example may clarify this. Consider a position with two sub-populations each at 50% with genotypes AA and AAA. Then $f_{\mathrm{AA}} = f_{\mathrm{AAA}} = 1-\varepsilon$, so that $\hat{f}^i$ will converge to $0.5 \cdot f_{\mathrm{AA}} + 0.5 \cdot f_{\mathrm{AAA}} = 0.5 \cdot (1-\varepsilon) + 0.5 \cdot (1-\varepsilon) = 1-\varepsilon$.

This observation is important because it means that allelic frequencies are not useful for distinguishing copy number variants.

Model

Assumptions

To simplify the subsequent analysis we make a simplifying assumption that there are two populations of cells at a given position i. One population consists of cells for which $r_j^i = c_j^i$, which we refer to as the *reference* population. This includes both normal cells, and tumour cells with copy number variations. The common factor in this group is that the genotypes contain no variant alleles. As discussed above, using allele frequencies we cannot deconvolve the fraction of normal cells, and tumour cells with no mutations. We assign the variable $f_r^i = 1-\varepsilon$ to the reference allele frequency for this population.

The second population which we refer to as the *variant* population will have $r_j^i < c_j^i$, that is at least one variant allele is present in the genotype. We will further assume that only one such population exists, so that we have the parameters c_v^i, r_v^i, f_v^i for this population.

We let ϕ^i be the fraction of cells from the variant population and $1-\phi^i$ the fraction of cells from the reference population. It then follows that $\hat{f}^i$ converges to $(1-\phi^i) f_r^i + \phi^i f_v^i$. This implies that if we knew f_v^i, we could estimate ϕ^i via

$$\phi^i = \frac{\hat{f}^i - f_r^i}{f_v^i - f_r^i}$$

The key difficulty is that knowledge of f_v^i, would imply knowledge of the variant populations genotype. This information is generally not available, however it is possible that some prior beliefs over possible genotypes exists.

Accomodating Genotype Uncertainty

The discussion above implies that we can model the number of reference allele counts observed at position i, as a binomial distribution with parameters $n = d^i = a^i + b^i$, and $p = (1-\phi^i) f_r^i + \phi^i f_v^i$. However, the uncertainty in genotype means in f_v^i is unknown.

At this point it is useful to switch notation slightly. We will let $\mu_r = 1-\varepsilon$ be the probability of sampling a reference allele from the reference population. To reiterate, $\varepsilon \ll 1$ is a value associated with the error rate of the sequencing technology.

We will also introduce an infinite vector $\boldsymbol{\mu}_v = (\mu_{\mathrm{A}}, \mu_{\mathrm{B}}, \mu_{\mathrm{AA}}, ...)$. The entries, $\mu_{v:g}$, in this vector are the probability of sampling a reference allele from the variant population with genotype g.

In addition for each position i we will introduce a vectors $\boldsymbol{\pi}^i = (\pi^i_{\mathrm{A}}, \pi^i_{\mathrm{B}}, \pi^i_{\mathrm{AA}}, ...)$, in which the entries, π^i_g, are the probabilities the variant population at site i has genotype g. Only a finite number of entries in $\boldsymbol{\pi}^i$ will be non-zero.

Finally we introduce a variable G^i, which indicates which genotype the variant population has.

We can the create a hierarchical Bayesian model of the data as follows.

$$\begin{aligned} \phi^i &\sim \text{Uniform}(0,1) \\ G^i &\sim \text{Discrete}(\boldsymbol{\pi}^i) \\ a^i | d^i, G^i = g, \phi^i, \mu_r, \boldsymbol{\mu}_v &\sim \text{Binomial}(d^i, (1-\phi^i)\mu_r + \phi^i \, \mu_{v:g}) \end{aligned}$$

We can then compute a posterior distribution on ϕ^i

$$\mathbb{P}(\phi^i, G^i = g | a^i, d^i, \phi^i, \mu_r, \boldsymbol{\mu}_v, \boldsymbol{\pi}^i) \;\propto\; \mathbb{P}(a^i | d^i, G^i = g, \phi^i, \mu_r, \boldsymbol{\mu}_v) \mathbb{P}(G^i | \boldsymbol{\pi}^i) \mathbb{P}(\phi^i)$$

We can then marginalise out G^i to obtain a posterior distribution over ϕ^i.

Hierarchical Modeling Of Genotype Uncertainty

Eliciting prior information about $\boldsymbol{\pi}^i$ will be difficult. It is beneficial to add another level to the model by placing a Dirichlet prior over $\boldsymbol{\pi}^i$. This frees us from having to directly input prior probabilities about genotypes, and instead allows us to work with the simpler pseudo-count parameters of the Dirichlet distribution. In addition, deepening the model hierarchy will provide some measure of protection against inaccurate prior specification for $\boldsymbol{\pi}$.

Formally we have

$$\boldsymbol{\pi}^i | \boldsymbol{\delta}^i \;\sim\; \text{Dirichlet}(\boldsymbol{\delta}^i)$$

Calculation of the posterior then becomes

$$\mathbb{P}(\phi^i, G^i = g, \boldsymbol{\pi}^i | a^i, d^i, \phi^i, \mu_r, \boldsymbol{\mu}_v, \boldsymbol{\delta}^i) \;\propto\; \mathbb{P}(a^i | d^i, G^i = g, \phi^i, \mu_r, \boldsymbol{\mu}_v) \mathbb{P}(G^i | \boldsymbol{\pi}^i) \mathbb{P}(\phi^i) \mathbb{P}(\boldsymbol{\pi}^i | \boldsymbol{\delta}^i)$$

Again we can marginalise the nuisance parameter G^i, and now $\boldsymbol{\pi}^i$. The final form of the posterior after marginalization is

$$\mathbb{P}(\phi^i | a^i, d^i, \phi^i, \mu_r, \boldsymbol{\mu}_v, \boldsymbol{\delta}^i) \;\propto\; \sum_g \left[\frac{\prod_{g' \neq g} \Gamma(\delta_{g'}) \times \Gamma(\delta_g + 1)}{\Gamma\left(\sum_{g'} \delta_{g'} + 1\right)} \right] \text{Binomial}(d^i, (1-\phi^i)\mu_r + \phi^i \, \mu_{v:g})$$

Since ϕ^i is a one dimensional parameter with support in the interval [0, 1] it is trivial to compute the normalization constant using numerical integration techniques.

Sharing Statistical Strength Across Samples

In the above formulation prior uncertainty about variant population genotypes, will be translated into posterior uncertainty about ϕ^i. In particular, if we believe several genotypes are equally probable, then there will be equal number of modes of the same height in the posterior for ϕ^i corresponding to each genotype. Without further assumptions we have no reason to believe any of these modes are more probable than the others.

If we consider the set of all mutations in a sample together, this no longer need be true. By utilizing a shared prior for the cellular frequencies ϕ^i across all positions in the sample, we can impose extra constraints to resolve ambiguities in genotypes. The choice of such a prior distribution will have a dramatic effect on inference, and should be considered carefully. In particular, the biological implications of the prior must be fully understood.

Dirichlet Process Prior

If we are willing to assume that modes the posterior of ϕ^i which are common between positions are more likely, we can resolve the multimodal posteriors. Biologically these means we are assuming there are groups of mutations, which tend to occur at the same frequency.

We can explicitly incorporate such an assumption into the model, if we change the prior distribution of ϕ^i to be shared across all positions in a sample. In particular we can use a shared discrete distribution ϕ^i to yield a traditional clustering model. However, using a discrete distribution would entail selecting the number of clusters of mutations in advance. Since this information is not generally available we use a semi-parametric Dirichlet process prior (DPP) over ϕ^i.