Module 2: Inferring clonal population structure from SNV data
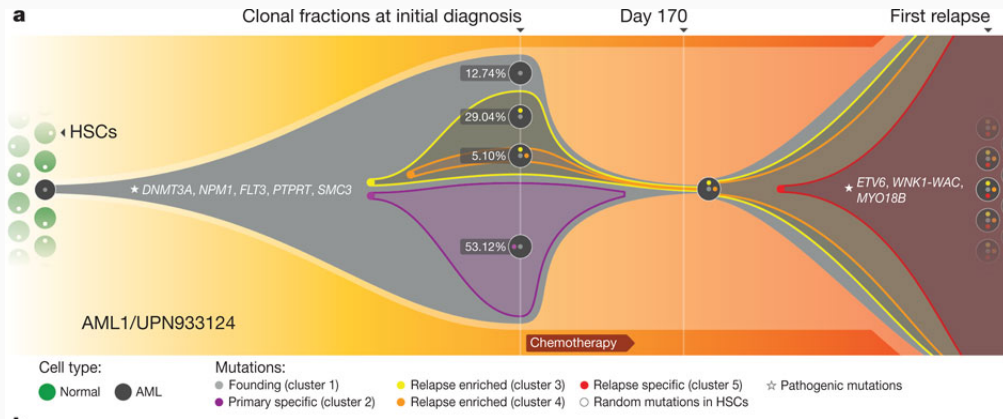
5th June 2019

Motivation
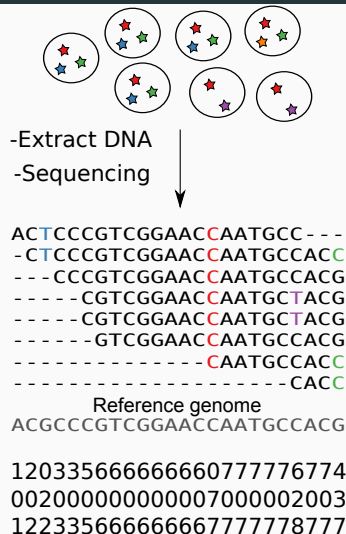
Clonal population structure

Ding et al. Nature 2012
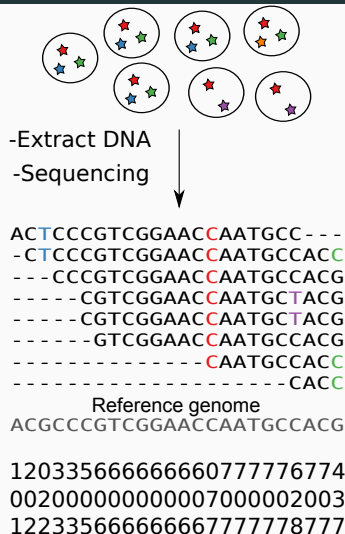
- Tumours are a mixture of clonal populations
- The number of reads with a mutation is proportional to how common the mutation is
- Mutations in the most common clones should have the most reads with variants

-Extract DNA
-Sequencing

```
ACTCCCGTCGGAACCAATGCC---
-CTCCCGTCGGAACCAATGCCACC
---CCCGTCGGAACCAATGCCACG
-----CGTCGGAACCAATGCTACG
-----CGTCGGAACCAATGCTACG
------GTCGGAACCAATGCCACG
-------------CAATGCCACC
--------------------CACC
```
Reference genome
ACGCCCGTCGGAACCAATGCCACG

a 120335666666660777776774
b 002000000000007000002003
d 122335666666667777778777

- *a* number of reads with reference allele
- *b* number of reads with variant allele
- *d* total number of reads
- Variant allele frequency (VAF) - proportion of reads with mutation $\frac{b}{d}$
- Cellular prevalence (cancer cell fraction) - proportion of cancer cells with a mutation


```
-Extract DNA
-Sequencing
```

```
ACTCCCGTCGGAACCAATGCC---
-CTCCCGTCGGAACCAATGCCACC
---CCCGTCGGAACCAATGCCACG
-----CGTCGGAACCAATGCTACG
-----CGTCGGAACCAATGCTACG
------GTCGGAACCAATGCCACG
-------------CAATGCCACC
--------------------CACC
```
Reference genome
ACGCCCGTCGGAACCAATGCCACG

a 12033566666666660777776774
b 00200000000000007000002003
d 12233566666666667777778777

4

High level questions

- Given high throughput bulk sequencing can we estimate how common a mutation is?
- If so can we infer what mutations occur together within a clonal population?

Simple solution
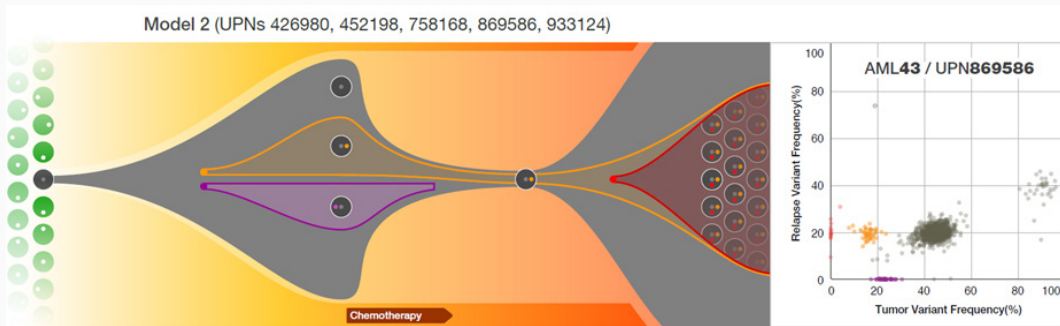
Ding et al. Nature 2012

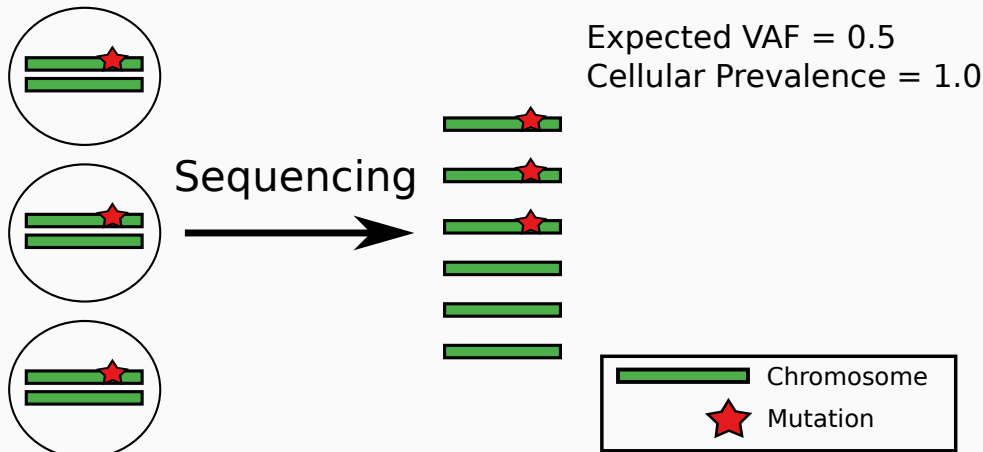- Cluster VAF to infer groups of mutations are similar prevalence

Should we use the VAF

- We measure two values *a* and *b* from HTS
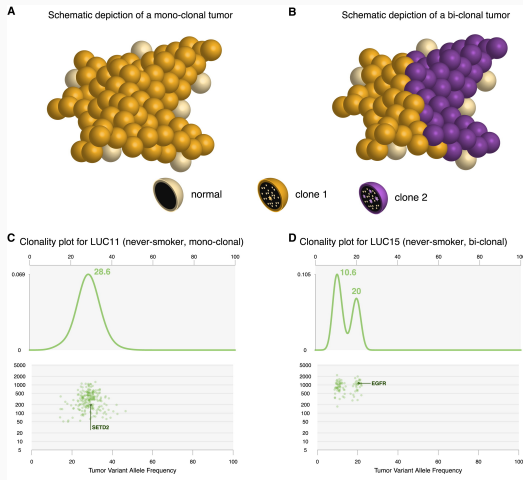- The VAF summarises these two value into one $f = \frac{b}{a+b}$

Questions

- Is it better to use VAF or allele counts as input for a model?
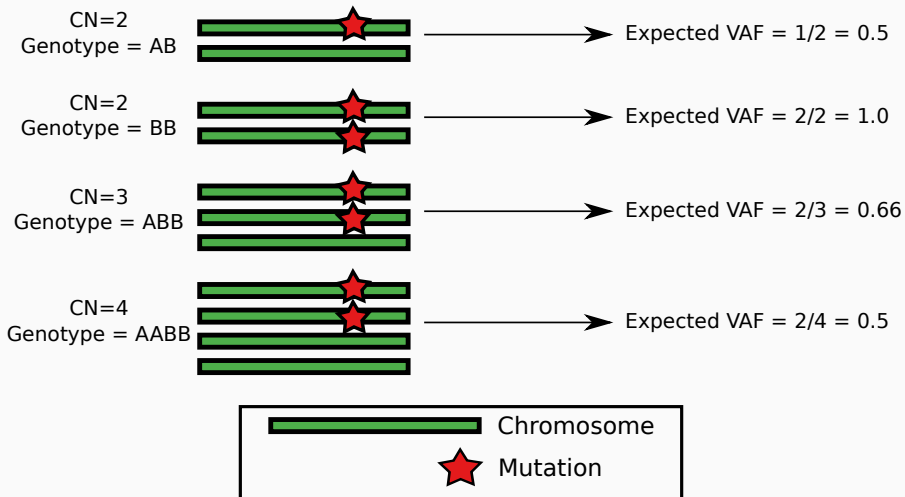- What distributions would be appropriate to model each type of data?

Questions

- Is it reasonable to interpret the the two cluster example as bi-clonal?
- What are some other explanations for the observed data?

Govidan et al. Cell 2012

Mutational genotype

CN=2
Genotype = AB → Expected VAF = 1/2 = 0.5

CN=2
Genotype = BB → Expected VAF = 2/2 = 1.0

CN=3
Genotype = ABB → Expected VAF = 2/3 = 0.66

CN=4
Genotype = AABB → Expected VAF = 2/4 = 0.5

Chromosome

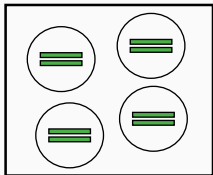Mutation

Mutational genotype model

Objective

The goal of this section will be to define a model which relates the observed allelic count data b, d to the latent (hidden) cellular prevalence of a mutation ϕ. To do so we will construct a simple model of how the data is generated. We follow a common pattern in probabilistic modelling of relating a latent parameter of interest. We will think about a few key factors in this process.
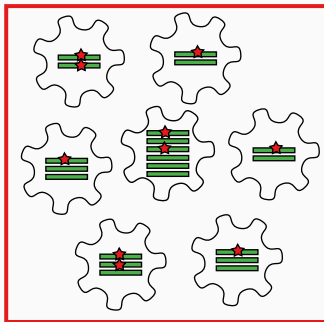
- How does the data relate to latent parameter?
- What is the model that generates the data given the cellular prevalence?
- What assumptions do we need to be able to evaluate the probability defined by this model?
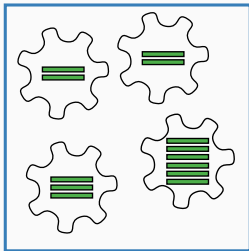
Populations structure

Normal Population

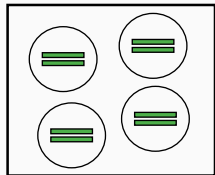Variant Population

Reference Population

12

We need to simplify

- The mutational genotypes of the cancer cells is a nuisance parameter we need to know to evaluate the likelihood.
- In the model suggested thus far, every cancer cell can have a different mutational genotype.
- This introduces a potentially large number of parameters, and would quickly lead to an unidentifiable model.
- We need to make a simplifying assumption to get to a tractable model.
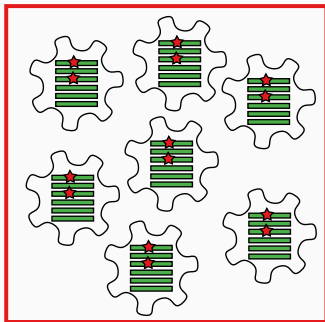
Simple population structure

Normal Population

Variant Population
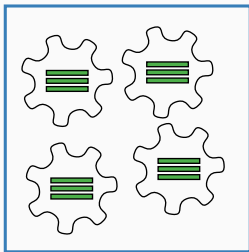
Reference Population

Modelling the data

We know pose the following question: What is the probability of sampling a read with the mutation? To answer this question we will imagine the following.

- Cells are lysed and DNA released into a pool.
- There are an infinite number of fragments in the pool.
- We select a fragment uniformly at random.

Questions

- Why do we assume the pool of DNA fragments is infinite?
- How do these assumptions help us compute the probability of sampling a read with the mutation?

We will break the computation of the probability of sampling a read with the mutation into two parts.

1. We assign the fragment to a cell population.
2. Given that the fragment comes from a population we then ask what the probability of sampling a mutant read is.

Notation

- $\psi = (g_N, g_R, g_V)$ is the mutational genotypes of the normal, reference and variant population.
- $c(g)$ is the number of chromosomes for genotype g
- $\mu(g)$ is the probability of sampling a mutant read given the genotype g
- t is the proportion of cancer cells

Probability of sampling from a population

We assume the probability of sampling a fragment from a population is proportional to the prevalence of the population and the number of copies of the locus that population has.

- Normal - $(1-t)c(g_N)$
- Reference - $t(1-\phi)c(g_R)$
- Variant - $t\phi c(g_V)$

Question
Why is it important to include the copy number in these probabilities?

Probability of sampling from a population

We can now define the probability of sampling from a population. Let $E \in \{N, R, V\}$ be the random variable indicating which population a fragment was sample from. Then

$$
P(E = e) = \begin{cases} \frac{(1-t)c(g_N)}{Z} & e = N \\ \frac{t(1-\phi)c(g_R)}{Z} & e = R \\ \frac{t\phi c(g_V)}{Z} & e = V \end{cases}
$$

$$
Z = (1-t)c(g_N) + t(1-\phi)c(g_R) + t\phi c(g_V)
$$

Probability of sampling a mutant read

Now we will condition on the population and define the probability of sampling a mutant read given the cell came from population e. The obvious way to do this is to assume we choose a chromosome at random from the cell. The probability of the chromosome having the mutation is then $\frac{\#\text{mutant chromosomes}}{\#\text{total chromosomes}}$. Let F be a random variable indicating if the read has a mutation.

$$
\begin{aligned}
P(F = 1 | E = e) &= \mu(g_e) \\
&= \begin{cases} \varepsilon & b(g_e) = 0 \\ \frac{b(g_e)}{c(g_e)} & b(g_e) \in \{1, \ldots, c(g_e)\} \\ 1 - \varepsilon & b(g_e) = c(g_e) \end{cases}
\end{aligned}
$$

Question
What is ε used for?

Joint probability

Putting everything together. We can now compute compute the probability of observing a read with a mutation.

$$
\begin{aligned}
P(F = 1 | \psi, \phi, t) &= \sum_{e \in \{N,R,V\}} P(F = 1 | E = e) P(E = e) \\
&= \frac{(1-t)c(g_N)}{Z} \mu(g_N) + \frac{t(1-\phi)c(g_R)}{Z} \mu(g_R) + \frac{t\phi c(g_V)}{Z} \mu(g_V) \\
&:= \xi(\psi, \phi, t)
\end{aligned}
$$

Note that this equation assumes the mutational genotype, cellular prevalence and tumour content are known which we show in the conditional probability.

Multiple reads

- In practice we observe d reads.
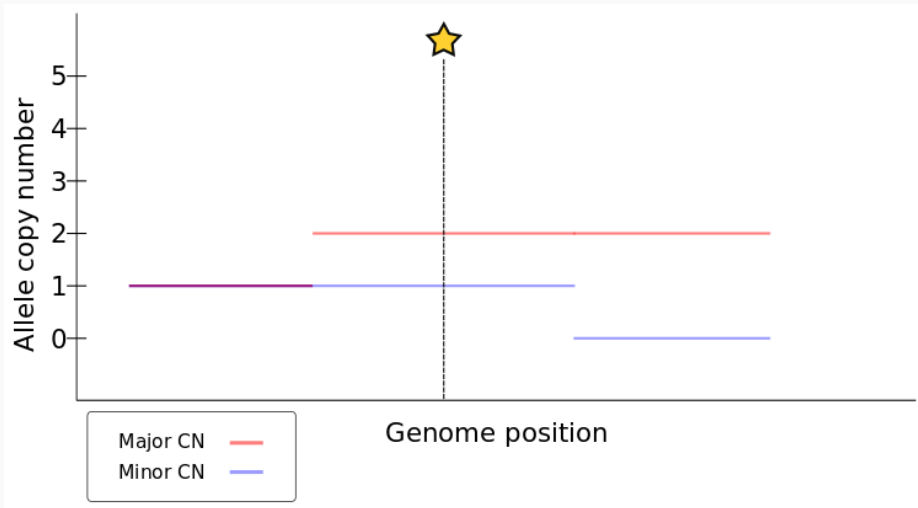- The probability b of these are mutants is then

$$p(b|d, \psi, \phi, t) \;=\; \text{Binomial}(b|d, \xi(\psi, \phi, t))$$

Genotype prior

- Thus far we have treated the population genotypes as known
- In practice we do not know the genotypes
- The genotypes are latent variables in our model and we need to specify a prior

Question
What prior should we use for the genotypes?

Copy number and mutation

Putting it together

- To finish off the model we need a prior for ϕ. In the absence of any knowledge we use a continuous prior.
- We assume t is known and fixed. It could come from CNV analysis
- We can now compute the joint distribution and posterior

$$
\begin{aligned}
p(b, d, \phi, \pi, t) &= p(\phi) \sum_{\psi} \pi_{\psi} \text{Binomial}(b|d, \xi(\psi, \phi, t)) \\
p(\phi|b, d, \pi, t) &= \frac{p(b, d, \phi, \pi, t)}{\int p(b, d, \phi, \pi, t) \mathrm{d}\phi}
\end{aligned}
$$

- The normalisation constant, $\int p(b, d, \phi, \pi, t)\mathrm{d}\phi$, does not have a closed form. Since it is a 1-D integral we use numerical quadrature.

26

Posteriors

a=100, b=100, CN=(1,1)

a=200, b=100, CN=(2,1)

Cellular prevalence

Typo: top is CN=(2,0)

27
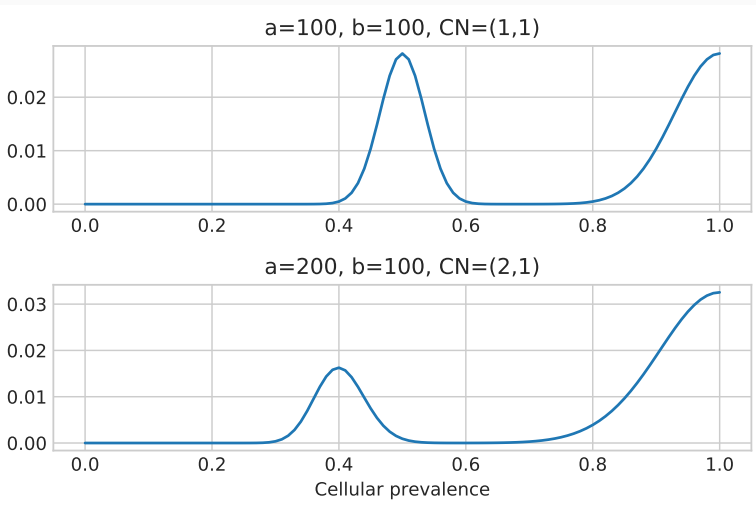
Mutational clustering

- We have a model to relate cellular prevalence to allele counts.
- So far each mutation is treated independently.
- The posteriors we obtain are multi-modal reflecting uncertainty in genotypes.

Question
Can anyone think of a way to deal with the mult-modality?
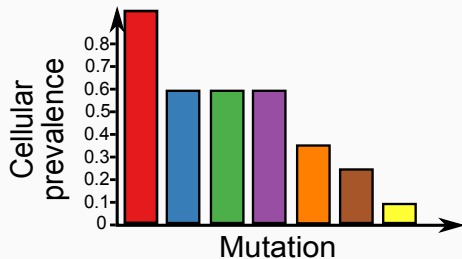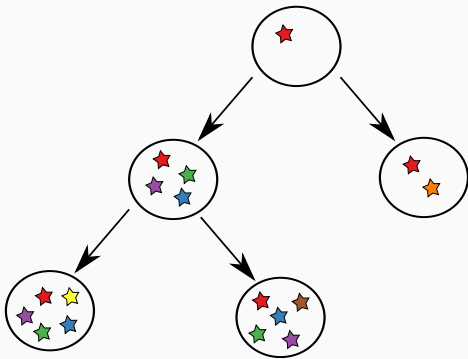
Mixture models

- Mixture models provide a probabilistic way to cluster data.
- The key idea is to associate data point x_n with a (latent) cluster indicator variable z_n.
- The cluster indicator serves to choose which parameter generated the data.

Generic mixture model

A generic mixture model with *K* components (clusters) is as follows

$$
\begin{aligned}
\kappa &\in \mathbb{R}_+^K \\
\rho | \kappa &\sim \text{Dirichlet}(\cdot | \kappa) \\
z_n | \rho &\sim \text{Categorical}(\cdot | \rho) \\
\theta_k &\sim G(\cdot) \\
x_n | z_n = k, \theta &\sim F(\cdot | \theta_k)
\end{aligned}
$$

where *G* and *F* are arbitrary distributions.

Improving the model

The main idea is to share the cellular prevalences by making them the parameters of the mixture distribution. Each cluster then represents a set of mutations which share the same evolutionary history.

$$
\begin{aligned}
\rho | \kappa &\sim \text{Dirichlet}(\cdot | \kappa) \\
z_n | \rho &\sim \text{Categorical}(\cdot | \rho) \\
\phi_k &\sim \text{Uniform}(\cdot | [0, 1]) \\
b_n | d_n, \phi, \pi_n, t, z_n &\sim \sum_{\psi} \pi_{n\psi} \text{Binomial}(\cdot | d_n, \xi(\psi, \phi_{z_n}, t))
\end{aligned}
$$

MCMC inference

We can no longer use quadrature to integrate all the variables to compute the posterior. We could use EM to compute MAP estimates. However, it is useful to quantify uncertainty in this problem so we will do full MCMC inference. To update the parameters we use the following strategy.

- Update ϕ_k with an MH move
- Update z_n with a Gibbs sampler
- Update ρ with a Gibbs sampler

The most complicated update is for the cluster indicators. To derive the Gibbs update we need to compute the conditional distribution of z_n given all the other parameters. This has the form

$$p(z_n = k|-) = \frac{\rho_k p(b_n|d_n, \phi, \pi_n, t, z_n = k)}{\sum_\ell \rho_\ell p(b_n|d_n, \phi, \pi_n, t, z_n = \ell)}$$

Question
How do we set the number of clusters?

A different view of mixture models

- Note that in a mixture model if $z_i = z_j$, then we have $\theta_{z_i} = \theta_{z_j}$ that is data point i and j share the same parameter.
- Instead of explicitly tracking the indicators we could instead assign each data point it own θ_n.
    - If $\theta_i = \theta_j$ we say i and j are in the same cluster.
- Now if $\theta_i \sim G$ and G is a continuous distribution, there is zero probability of sharing parameters.
- Implicitly a mixture model solves this by constructing a discrete distribution $G(\cdot) = \sum_{k=1}^{K} \rho_k \delta_{\theta_k}(\cdot)$

Dirichlet process

- The Dirichlet process (DP) is a non-parametric Bayesian prior.
- Formally it is a distribution over distributions, or a stochastic process.
- There are two parameters to a DP: α the concentration and G_0 the base measure.
- α roughly controls the number of clusters
- G_0 controls where the cluster parameters are located.
- If we have $G \sim \text{DP}(\cdot | \alpha, G_0)$ then G is almost surely discrete.

Dirichlet process mixture models

- The last two slides suggest a way to use the DP to estimate the number of clusters.
- In the finite case we set G to have K components, and update the mix-weights ρ_k and cluster parameters θ_k.
- If we instead sample G from a DP then the number of components K will be random.
- Thus we can learn the value of K.

Non-parametric model

$$
\begin{aligned}
G_0 &= \text{Uniform}(\cdot|[0,1]) \\
G|\alpha, G_0 &\sim \text{DP}(\cdot|\alpha, G_0) \\
\phi_n|G &\sim G \\
b_n|d_n, \phi_n, \pi_n, t &\sim \sum_{\psi} \pi_{n\psi}\text{Binomial}(\cdot|d_n, \xi(\psi, \phi_n, t))
\end{aligned}
$$

Summarising the posterior

- MCMC sampling will produce a set of parameter samples called the trace.
- This gives us an approximate posterior over all possible clusterings.

Question
How should we report a single "best" clustering?

Summarising the posterior

- We compute a consensus clustering, $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_N)$, using the MPEAR technique to approximately optimise the Adjusted Rand Index loss function.
  - This requires computing the posterior similarity matrix S where S_{ij} is the proportion of MCMC samples where mutation i and j appear in the same cluster.
  - S is invariant to the number of clusters and reordering the cluster labels.
- We can also compute the posterior cellular prevalence of a cluster as

$$p(\phi_k | X, \hat{\mathbf{z}}) = \frac{p(\phi_k) \prod_{\{n: \hat{z}_n = k\}} p(x_n | \phi_k)}{\int p(\phi_k) \prod_{\{n: \hat{z}_n = k\}} p(x_n | \phi_k) \mathrm{d}\phi_k}$$

Multiple samples

The extension to multiple samples is straightforward. The only major modification is to change ϕ_k to a vector ϕ_k and the prior from Uniform on the interval [0,1] to the Uniform on the hyper-cube $[0,1]^M$. Let M be the number of samples and m be the index for samples. Then the multi-sample model is

$$
\begin{aligned}
G_0 &= \text{Uniform}(\cdot|[0,1]^M) \\
G|\alpha, G_0 &\sim \text{DP}(\cdot|\alpha, G_0) \\
\phi_n|G &\sim G \\
b_{nm}|d_{nm}, \phi_{nm}, \pi_{nm}, t &\sim \sum_{\psi} \pi_{nm\psi} \text{Binomial}(b|d, \xi(\psi, \phi_{nm}, t_m))
\end{aligned}
$$

Overdispersion

- HTS allele counts can have more variance than a Binomial can account for.
- For high depth sequencing this problem becomes important.
- To overcome the issue we use an overdipsersed distribution, the Beta-Binomial.
- We set the mean to ξ and introduce a global variance parameter.