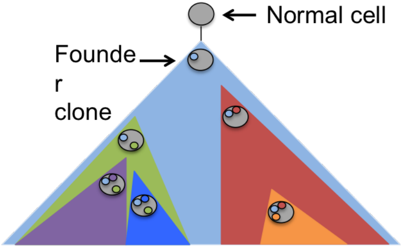
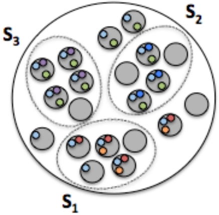
Clustering with the Multivariate binomial model/DP

January 16, 2017

Cancer is an evolutionary disease.



Clonal tree



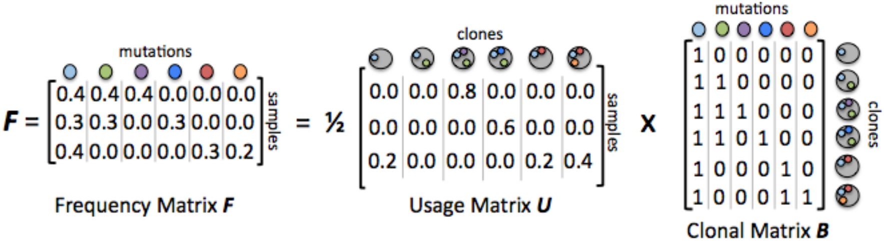
Multiple samples

This mixture of clones is called intratumor heterogeneity.

Thanks to Mohammed for letting me use his figures.

Ancestree.

The purpose of Ancestree is to infer the clonal tree that relates the clones and the proportions of each clone in each sample, given a matrix of variant allele frequencies (VAFs) indexed by samples and SNVs.



$$F = \begin{bmatrix}0.4&0.4&0.4&0.0&0.0&0.0\\0.3&0.3&0.0&0.3&0.0&0.0\\0.4&0.0&0.0&0.0&0.3&0.2\end{bmatrix} = ½\begin{bmatrix}0.0&0.0&0.8&0.0&0.0&0.0\\0.0&0.0&0.0&0.6&0.0&0.0\\0.2&0.0&0.0&0.0&0.2&0.4\end{bmatrix} \text{ X } \begin{bmatrix}1&0&0&0&0&0\\1&1&0&0&0&0\\1&1&1&0&0&0\\1&1&0&1&0&0\\1&0&0&0&1&0\\1&0&0&0&1&1\end{bmatrix}$$

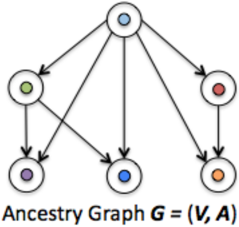
Frequency Matrix ***F***

Usage Matrix ***U***

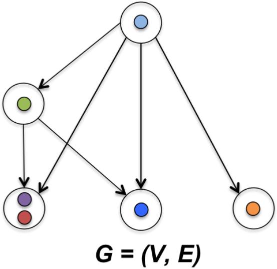
Clonal Matrix ***B***

Clustering mutations.

Why cluster mutations? What does that mean?



Unclustered

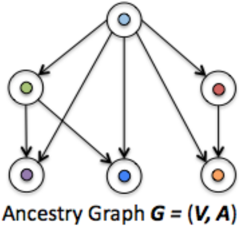


Clustered

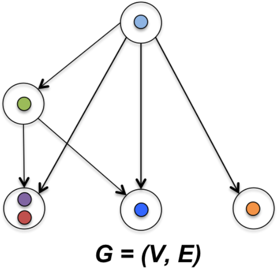
Clustering mutations.

Why cluster mutations? What does that mean?

- Passenger mutations.



Unclustered

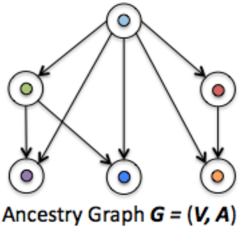


Clustered

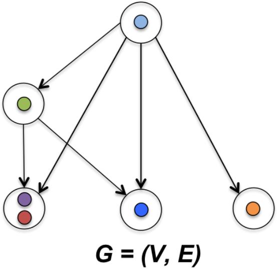
Clustering mutations.

Why cluster mutations? What does that mean?

- Passenger mutations.
- Low read coverage.



Unclustered



Clustered

Model, Inference, Equations

I will defer to the pdf here.

Results

Simulated datasets:

Number of clusters	10
Number of SNVs	100
Number of samples	4, 5, 6
Coverage	50, 100, 1000

Results

Simulated datasets:

Number of clusters	10
Number of SNVs	100
Number of samples	4, 5, 6
Coverage	50, 100, 1000

Evaluating cluster assignments

- Adjusted Rand Index
- Cluster frequency error

$$\frac{1}{m}\sum_{p=1}^{m}\frac{1}{T}\sum_{t=1}^{T}\min_{j}|\phi_t - \widehat{\phi}_j|$$

- Number of clusters
- Number of mutations placed by Ancestree

Plots

- Violin plots
- Posterior plots