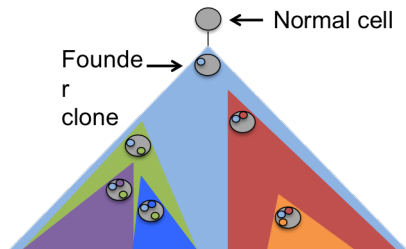


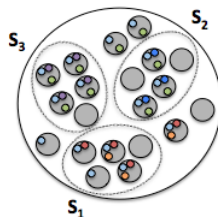
Clustering with the Multivariate binomial model/DP

January 17, 2017

Cancer is an evolutionary disease.



Clonal tree



Multiple samples

This mixture of clones is called intratumor heterogeneity.

Ancestry.

The purpose of Ancestry 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.

$$\mathbf{F} = \frac{1}{2} \mathbf{U} \mathbf{B}$$

Frequency Matrix \mathbf{F} Usage Matrix \mathbf{U} Clonal Matrix \mathbf{B}

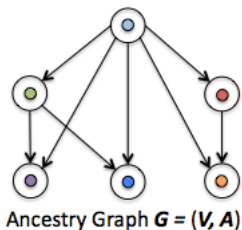
The diagram illustrates the relationship between three matrices used in the Ancestry tool:

- Frequency Matrix \mathbf{F}** : A 3x6 matrix of variant allele frequencies (VAFs) indexed by samples (rows) and mutations (columns). The values are: $\begin{bmatrix} 0.4 & 0.4 & 0.4 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.3 & 0.0 & 0.3 & 0.0 & 0.0 \\ 0.4 & 0.0 & 0.0 & 0.0 & 0.3 & 0.2 \end{bmatrix}$.
- Usage Matrix \mathbf{U}** : A 3x6 matrix representing the proportion of each clone in each sample. The values are: $\begin{bmatrix} 0.0 & 0.0 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.0 & 0.2 & 0.4 \end{bmatrix}$.
- Clonal Matrix \mathbf{B}** : A 6x6 matrix representing the relationship between mutations and clones. The values are: $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$.

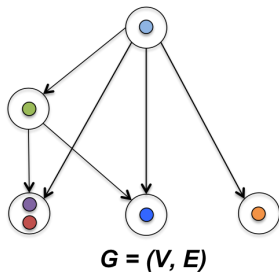
The equation shows that the Frequency Matrix \mathbf{F} is equal to half the product of the Usage Matrix \mathbf{U} and the Clonal Matrix \mathbf{B} .

Clustering mutations.

Why cluster mutations? What does that mean?



Unclassified

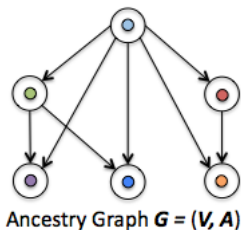


Clustered

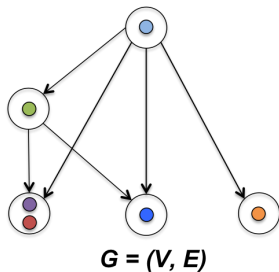
Clustering mutations.

Why cluster mutations? What does that mean?

- ▶ Passenger mutations.



Unclassified

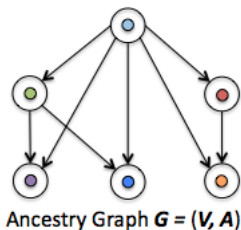


Clustered

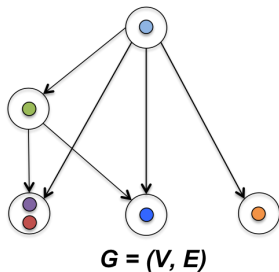
Clustering mutations.

Why cluster mutations? What does that mean?

- ▶ Passenger mutations.
- ▶ Low read coverage.



Unclassified



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

$$\frac{1}{m} \sum_{p=1}^m \frac{1}{T} \sum_{t=1}^T \min_j |\phi_t - \hat{\phi}_j|$$

- ▶ Number of clusters
- ▶ Number of mutations placed by Ancestry

Plots

- ▶ Violin plots
- ▶ Posterior plots