1 Model

Indices:

- Variant $i \in \{1, ..., I\}$
- Cluster $k \in \{1, ..., K\}$
- Sample $s \in \{1, ..., S\}$

Variables:

- $y[i, s]$ = variant read counts ∼ Binomial(n[i,s], θ[i,s])
- $n[i, s]$ = total read count (depth)
- $m[i, s]$ = multiplicity (# of variant alleles)
- $c[i, s]$ = total copy number
- $\omega[k, s] \in (0, 1]$; mutant cell fraction (MCF)
  prior: Beta(1,1)
- $z[i] \in \{1, ..., K\}$; cluster membership of variant
  prior: Categorical(π)
- $\pi[k] \in (0, 1)$; proportion of variants in each cluster
  prior: Dirichlet(1, ..., 1)
- $\theta[i, s] \in (0, 1]$; variant allele frequency (VAF)
  deterministic function of ω, m, c, n, z

$$\theta[i, s] = \frac{m[i, s] \times \omega[z[i], s]}{c[i, s] \times \omega[z[i], s] + 2 \times (1 - \omega[z[i], s])}$$

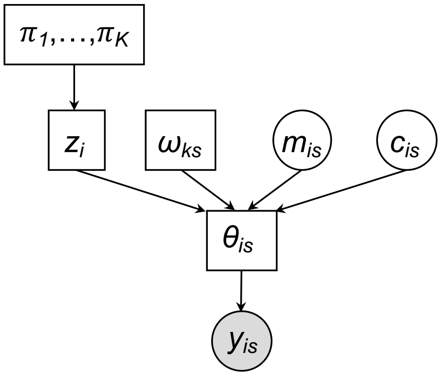

Figure 1: Bayesian hierarchical model for variant clustering and MCF estimation.

2 Current scheme

1. Split variants into sets based on presence in samples. Each set makes up a ”box” in crude tree structure. Ordering of variants is limited by this structure – can only make vertical connections.
2. Within each box, cluster variants and estimate MCFs. Use BIC to determine number of clusters, k.
3. Order variant clusters (i.e. connect cluster nodes to form tree).

3 Problems/Issues

- P(tree | data) ?
- Clustering and CCF estimation is done within a box, but tree spans all boxes