Supplementary Methods

Bayesian Dirichlet process for clustering VAFs across multiple samples

We develop a model for clustering variant allele fractions (VAFs) of mutations called in a single sample to mutation data across multiple samples from the same individual. In normal somatic cells, the vast majority of the genome retains its normal, diploid copy number, which means that we can cluster the VAFs directly (excluding mutations on the X and Y chromosomes in males) – this has the considerable advantage that the Dirichlet Process model we build can rely directly on conjugate prior distributions. The model could be extended to include a correction for different copy number states in given samples for a particular mutation through, for example, a Metropolis-Hastings update, but at considerable computational cost. The full mathematical development of the model is detailed in the Supplementary Information.

For every somatic mutation, we have a vector of the number of sequencing reads reporting the variant allele across each sample, together with the total sequencing depth for each mutation in each sample. We assume that each mutation can be assigned to one of an unknown number of clones – each clone has an expected VAF across the set of sequenced samples. We want to estimate:

1. The number of clusters (clones);
2. The location of each cluster in the n-dimensional VAF hypercube;
3. The allocation of mutations to each cluster.

We model these data using a hierarchical Bayesian model, where the distribution of clone sizes and numbers follows a Dirichlet process.

We define N as the number of somatic mutations across all M samples in a given sample; $n_{i,j}, i = 1, \dots, N, j = 1, \dots, M$ as the total read depth for mutation i in sample j; of which $y_{i,j}$ report the reference allele. Then $y_{i,j} \sim \text{Bin}(n_{i,j}, \pi_{i,j})$, where $\pi_{i,j}$ is the expected proportion of reads reporting the reference allele. Here, $\pi_{i,j}$ follows a Dirichlet process: $\boldsymbol{\pi_i} \sim DP(\alpha P_0) \in [0,1]^M$. We use the stick-breaking representation of the Dirichlet process:

$$P = \sum_{h=1}^{\infty} \omega_h \delta_{\boldsymbol{\pi_h}}, \text{ with } \boldsymbol{\pi_h} \sim P_0,$$

where δ_π is a point mass at π and ω_h is the weight of the hth mutation cluster (that is, effectively the proportion of mutations found in cluster h). To capture the stick-breaking formulation, we let $\omega_h = V_h \prod_{l<h}(1 - V_l)$, with $V_h \sim \text{Beta}(1, \alpha)$. We set a practical maximum number of clusters, C, as 100. As priors, we set $P_0 \sim U(0,1)^M$ and $\alpha \sim \Gamma(0.01, 0.01)$.

To model the posterior distribution of the Dirichlet process, we use Gibbs sampling, as follows:

Step 1: Allocating each mutation to one of the clusters

We set indicator variables, $S_i \epsilon \{1,2,\dots,C\}$, to denote allocation of mutation i to a cluster. The posterior distribution of these variables is therefore:

$$\Pr(S_i = h|-) = \frac{(V_h \prod_{l<h}(1 - V_l)) \left(\prod_j \binom{n_{i,j}}{y_{i,j}} p_{h,i,j}^{y_{i,j}} (1 - p_{h,i,j})^{n_{i,j} - y_{i,j}} \right)}{\sum_{r=1}^{C} (V_h \prod_{l<h}(1 - V_l)) \left(\prod_j \binom{n_{r,j}}{y_{r,j}} p_{h,r,j}^{y_{r,j}} (1 - p_{h,r,j})^{n_{r,j} - y_{r,j}} \right)},$$

where $h = 1,2,\dots,C$.

Step 2: Up-dating the stick-breaking weights

These are conditionally conjugate beta posterior distributions:

$$(V_h|-) \sim \text{Beta}\left(1 + \sum_{i=1}^{N} 1(S_i = h)\,,\, \alpha + \sum_{i=1}^{N} 1(S_i > h)\right),$$

where $h = 1,\dots,C-1$ and $V_C = 1$.

Step 3: Up-dating the cluster positions in the M-dimensional VAF hypercube

We want to generate draws from the posterior distribution of $(\boldsymbol{\pi_h}|-)$. Since the prior is $U(0,1)^M$, we have:

$$\pi_{h,j} \sim \text{Beta}\left(1 + \sum\nolimits_{i:S_i=h} y_{i,j}\,,\, 1 + \sum\nolimits_{i:S_i=h} (n_{i,j} - y_{i,j})\right)$$

Step 4: Testing a split-merge step on the current clustering

With large numbers of samples, many of which are not clonally related to other samples from the same individual, the clusters tend to fall on the edges of the VAF hypercube. The posterior distribution can be rather peaky, with local maxima separated by large regions of very low probability. This can make the Gibbs sampler prone to becoming imprisoned, without fully exploring the complete posterior distribution. To alleviate this problem, we include a potential split-merge step at each cycle of the Gibbs sampler, following a previously described Metropolis-Hastings proposal for conjugate distributions (see D. Dahl: *An improved merge-split sampler for conjugate Dirichlet process mixture models*. University of Wisconsin-Madison Technical Report 1086; 2003).

Two mutations, denoted i and j, are drawn at random from the current state of the Gibbs sampler. If the two mutations are currently in the same cluster, a potential split step is considered:

- Form two new sets seeded with each mutation: $S_i = \{i\}$ and $S_j = \{j\}$.
- Generate a random permutation of the remaining mutations in the given cluster.
- Taking each mutation, k, in turn, allocate to cluster S_i with probability given by the beta-binomial posterior distribution based on mutations already allocated (*Equation 1*):

$$\Pr(k \in S_i | \boldsymbol{y}, \boldsymbol{n}) \propto |S_i| \prod_r \binom{n_{k,r}}{y_{k,r}} \frac{\mathrm{B}\left(y_{k,r} + \sum_{l<k} y_{l,r} + 1,\ \ n_{k,r} - y_{k,r} + \sum_{l<k}\left(n_{l,r} - y_{l,r}\right) + 1\right)}{\mathrm{B}\left(\sum_{l<k} y_{l,r} + 1,\ \ \sum_{l<k}\left(n_{l,r} - y_{l,r}\right) + 1\right)},$$

  where this is normalised by the sum of this and the equivalent expression for $k \in S_j$. Otherwise, allocate to S_j.
- Compute the Metropolis-Hastings ratio as below.

If the two mutations are in different clusters, a potential merge step is considered:

- Form one merged set with all mutations currently allocated to the two clusters: $S = S_i \cup S_j$.
- Compute the Metropolis-Hastings ratio as below.

The Metropolis-Hastings ratio is computed as follows: Given the proposed new allocation $\boldsymbol{\eta}^*$, and the current existing allocation $\boldsymbol{\eta}$, the ratio is calculated as:

$$\alpha(\boldsymbol{\eta}^*|\boldsymbol{\eta}) = \min\left(1, \frac{p(\boldsymbol{\eta}^*|\boldsymbol{y})\,\Pr(\boldsymbol{\eta}|\boldsymbol{\eta}^*)}{p(\boldsymbol{\eta}|\boldsymbol{y})\,\Pr(\boldsymbol{\eta}^*|\boldsymbol{\eta})}\right).$$

If the proposal is a merge step, $\Pr(\boldsymbol{\eta}^*|\boldsymbol{\eta}) = 1$, since there is only one way to merge two clusters into one. $\Pr(\boldsymbol{\eta}|\boldsymbol{\eta}^*)$ is calculated as the product of the sequential beta-binomial posterior probabilities as if the merged cluster were being split into the two original clusters (*Equation 1* above), noting that the mutations should be in a random order for the calculation. The partition probabilities are given by:

$$p(\boldsymbol{\eta}^*|\boldsymbol{y})$$
$$\propto \alpha\Gamma(|S|)\prod_{k\in S}\prod_{r}\binom{n_{k,r}}{y_{k,r}}\frac{\mathrm{B}\left(y_{k,r}+\sum_{l<k}y_{l,r}+1,\;\; n_{k,r}-y_{k,r}+\sum_{l<k}\left(n_{l,r}-y_{l,r}\right)+1\right)}{\mathrm{B}\left(\sum_{l<k}y_{l,r}+1,\;\; \sum_{l<k}\left(n_{l,r}-y_{l,r}\right)+1\right)}$$

$$p(\boldsymbol{\eta}|\boldsymbol{y})$$
$$\propto \alpha^2\Gamma(|S_i|)\Gamma\left(|S_j|\right)\prod_{k\in S_i}\prod_{r}\binom{n_{k,r}}{y_{k,r}}\frac{\mathrm{B}\left(y_{k,r}+\sum_{l<k}y_{l,r}+1,\;\; n_{k,r}-y_{k,r}+\sum_{l<k}\left(n_{l,r}-y_{l,r}\right)+1\right)}{\mathrm{B}\left(\sum_{l<k}y_{l,r}+1,\;\; \sum_{l<k}\left(n_{l,r}-y_{l,r}\right)+1\right)}$$
$$\times\prod_{k\in S_j}\prod_{r}\binom{n_{k,r}}{y_{k,r}}\frac{\mathrm{B}\left(y_{k,r}+\sum_{l<k}y_{l,r}+1,\;\; n_{k,r}-y_{k,r}+\sum_{l<k}\left(n_{l,r}-y_{l,r}\right)+1\right)}{\mathrm{B}\left(\sum_{l<k}y_{l,r}+1,\;\; \sum_{l<k}\left(n_{l,r}-y_{l,r}\right)+1\right)}$$

If the proposal is a split step, $\Pr(\boldsymbol{\eta}|\boldsymbol{\eta}^*) = 1$ and $\Pr(\boldsymbol{\eta}^*|\boldsymbol{\eta})$ is the product of the probabilities in *Equation 1* above. The partition probabilities are the reverse of the two equations immediately above.

The proposed new split or merge step is accepted with probability given by the Metropolis-Hastings ratio.

Step 5: Up-dating the hyperparameter

The posterior distribution for α is:

$$(\alpha|-)\sim\Gamma\left(C+A-1, B-\sum_{l=1}^{C-1}\log\left(1-V_l\right)\right)$$

where the prior is $\alpha\sim\Gamma(A,B)$.