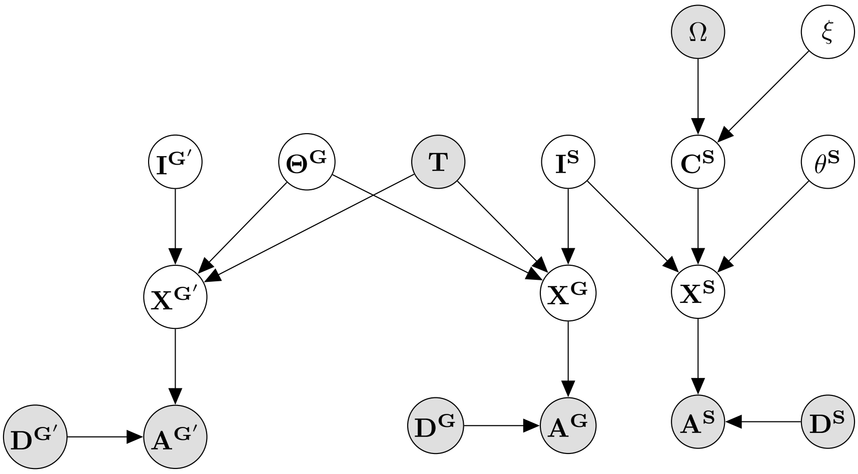
XClone

June 8, 2020

Contents

Proposed modification includes two binomial mixture models: for CNV and SNV modalities respectively. Two models are coupled by the clonal label assignment.



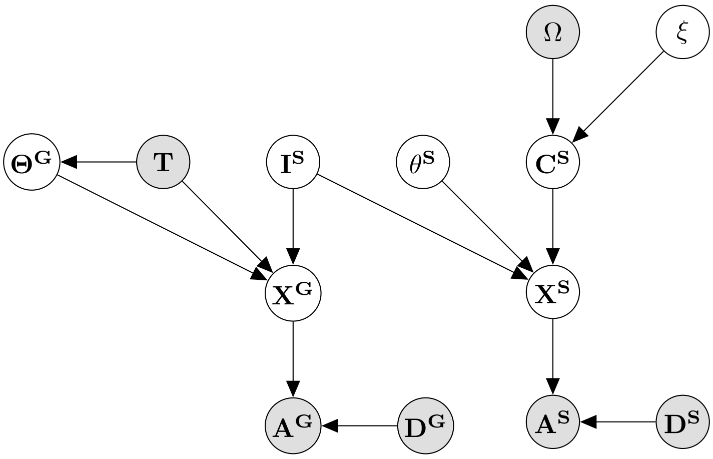
The most general model should have a capacity to integrate four modalities:

- scCNV inferred from scDNA-seq (let's call this **CNV-sample**).
- scCNV and scSNV inferred from scRNA-seq (from the same cells; let's call this **SNV-sample**).
- Clustering of variants inferred from bulk (Ω in the *Cardelino* paper).

The minimal implementation should be capable of integrating CNV and SNV information from the same sample. The SNV part doesn't change, it is exactly the same as in the original Cardelino paper.

1 CNV + SNV from the same sample

Here we suppose, that CNV information is only available for germline variants and SNV information — only for somatic variants.



The problem is that we usually can't robustly infer CNV. We can use ASE ratios of haplotype blocks instead to define an informative prior. This can be achieved in a slightly changed model.

1.1 Basic definitions

1. **Constants:**

- M — number of cells in the sample.
- K — estimated number of clones in the sample
- N_G, N_S — number of germline variant blocks and somatic variants in the sample.
- τ — set of possible CNV states ($\{1, \ldots, 8\}$ for now, as the case of zero should be treated with care if we can't say for sure whether the part of the chromosome (both arms) is deleted).

will be estimated via Dirichlet processes

2. **Other known quantities:**

- $\mathbf{D^S} \in \mathbb{N}^{N_S \times M}$, $\mathbf{D^G}$ — total read counts. To get a count of the block, one simply adds up the counts of the variants within the block. Here we assume that variants are far enough from each other, so that almost no reads overlap two variants at the same time. Otherwise, adding things up wouldn't make sense.
- $\mathbf{A^S} \in \mathbb{N}^{N_S \times M}$, $\mathbf{A^G}$ — same for allele-specific counts.
- $\mathbf{\Omega} \in \{0,1\}^{N_S \times K}$ — variant-to-clone assignment derived from bulk
- $\mathbf{f} = (f_1, \ldots, f_K)$ — K-dimensional vector of estimated clonal fractions ($\sum_{i=1}^{K} f_i = 1$)
- $\mathbf{T}$ — CNV states of blocks ($\mathbf{T_{i,k}}$ is a CNV state of a block i in clone k). **This information is only used to define a prior over the ASE ratio and can be ignored if absent. We can use ASE ratios of haplotype blocks as an informative prior instead.**

but how and what for?

3. **Inferred quantities:**

- $\mathbf{C} \in \{0,1\}^{N_S \times K}$ — clonal tree configuration matrix where $\mathbf{C_{i,k}} = 1$ if variant i is present in clone k and $\mathbf{C_{i,k}} = 0$ otherwise.
- $\mathbf{\Theta^G}$ — allelic rates of variants located within blocks of fixed CNV status. $\mathbf{\Theta^G_{i,t}}$ is an allelic rate of a block i with a CNV status t. Blocks are not the same across clones. Blocks are ordered as tuples of the form (*block start, block length*).
- $\theta^{\mathbf{S}} \in [0,1]^{K+1}$ — vector of allelic rates of somatic variants ($\theta_{\mathbf{0}}$ is a rate of an absent variant)
- $\mathbf{I^S} \in [K]^M$ — cell-to-clone assignment
- $\xi \in [0,1]$ — error rate of $\mathbf{C}$ configuration assignment (per element).

4. **Some notational conventions:**

- Capitalized letter without superscript denotes the information for somatic and germline variants combined. For instance, $\mathbf{C}$ stands for a configuration matrix that contains SNV-to-clone assignment information for somatic variants as well as variant-to-CNV assignment for germline variations. Same for $\mathbf{A}, \mathbf{D}, \mathbf{\Theta}$ etc.
- $\mathbf{H^G}$ — CNV status of the blocks in accordance with the current label assignment:

$$\mathbf{H^G_{i,j}} := \mathbf{T_{i,I^S_j}}$$

- $\mathbf{H^S}$ — presence/absence of variants in accordance with the current label assignment:

$$\mathbf{H^S_{i,j}} := \mathbf{C_{i,I^S_j}}$$

- $\mathbf{X^S} \in [0,1]^{N_S \times M}$ — a shortcut to simplify the notation:

$$\mathbf{X^S_{i,j}} := \begin{cases} \theta^{\mathbf{S}}_{\mathbf{i}} & \text{if } \mathbf{H^S_{i,j}} = 1, \\ \theta^{\mathbf{S}}_{\mathbf{0}} & \text{otherwise} \end{cases}$$

- $\mathbf{X^G}$ — same but for CNV modality: $\mathbf{X^G_{i,j}} := \mathbf{\Theta^G_{i,H^G_{i,j}}}$, meaning that $\mathbf{X^G_{i,j}}$ is an allelic rate of a block i in cell j based of the current cell-to-clone label assignment.

1.2 Generative model formulation

- **Cell-to-clone assignment posterior:**

$$\mathrm{P}(\mathbf{I_j^S} = k_0 \mid \mathbf{A_j}, \mathbf{D_j}, \mathbf{f}, \mathbf{C}, \boldsymbol{\Theta}) = \frac{\mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I_j^S} = k_0, \mathbf{C}, \boldsymbol{\Theta})\mathrm{P}(\mathbf{I_j^S} = k_0 \mid \mathbf{f})}{\sum\limits_{k=1}^{K} \mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I_j^S} = k, \mathbf{C}, \boldsymbol{\Theta})\mathrm{P}(\mathbf{I_j^S} = k \mid \mathbf{f})} \tag{1.1}$$

- **Posterior configuration:**

$$\mathrm{P}(\mathbf{C}, \boldsymbol{\Theta}, \xi \mid \mathbf{A}, \mathbf{D}, \boldsymbol{\Omega}, \mathbf{f}) = \mathrm{P}(\mathbf{C}, \theta^{\mathbf{S}}, \xi \mid \mathbf{A^S}, \mathbf{D^S}, \boldsymbol{\Omega}, \mathbf{f})\mathrm{P}(\boldsymbol{\Theta}^{\mathbf{G}} \mid \mathbf{A^G}, \mathbf{D^G}, \mathbf{f}) \tag{1.2}$$

- **ASE model:**

$$\begin{aligned} \mathrm{P}(\mathbf{A_{i,j}^S} \mid \mathbf{D_{i,j}^S}, \theta^{\mathbf{S}}) &= \mathrm{Binom}(\mathbf{A_{i,j}^S} \mid \mathbf{D_{i,j}^S}, \mathbf{X_{i,j}^S}) \\ \mathrm{P}(\mathbf{A_{i,j}^G} \mid \mathbf{D_{i,j}^G}, \boldsymbol{\Theta}^{\mathbf{G}}) &= \mathrm{Binom}(\mathbf{A_{i,j}^G} \mid \mathbf{D_{i,j}^G}, \mathbf{X_{i,j}^G}) \end{aligned} \tag{1.3}$$

- **ASE likelihood (both terms factorize over variants):**

$$\begin{aligned} &\mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I_j^S} = k, \mathbf{C}, \boldsymbol{\Theta}) = \mathrm{P}(\mathbf{A_j^S} \mid \mathbf{D_j^S}, \mathbf{I_j^S} = k, \mathbf{C}, \theta^{\mathbf{S}})\mathrm{P}(\mathbf{A_j^G} \mid \mathbf{D_j^G}, \mathbf{I_j^S} = k, \boldsymbol{\Theta}^{\mathbf{G}}) \\ &\mathrm{P}(\mathbf{A_j^S} \mid \mathbf{D_j^S}, \mathbf{I_j^S} = k, \mathbf{C^S}, \theta^{\mathbf{S}}) = \prod_{i=1}^{N_S} \mathrm{Binom}(\mathbf{A_{i,j}^S} \mid \mathbf{D_{i,j}^S}, \theta_{\mathbf{i}}^{\mathbf{S}})^{\mathbf{C_{i,k}}} \times \mathrm{Binom}(\mathbf{A_{i,j}^S} \mid \mathbf{D_{i,j}^S}, \theta_{\mathbf{0}}^{\mathbf{S}})^{\mathbf{1-C_{i,k}}} \\ &\mathrm{P}(\mathbf{A_j^G} \mid \mathbf{D_j^G}, \mathbf{I_j^S} = k, \boldsymbol{\Theta}^{\mathbf{G}}) = \prod_{i=1}^{N_G} \mathrm{Binom}(\mathbf{A_j^G} \mid \mathbf{D_j^G}, \boldsymbol{\Theta}_{\mathbf{i,k}}^{\mathbf{G}}) \end{aligned} \tag{1.4}$$

- **Allelic rate likelihood:**

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{j=1}^{M} \sum_{k=1}^{K} \mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I_j^S} = k, \boldsymbol{\Theta})\mathrm{P}(\mathbf{I_j^S} = k \mid \mathbf{f}) \tag{1.5}$$

  To view the clonal assignment in a Bayesian way, we introduce informative prior ν for $\boldsymbol{\Theta}$. Using that «posterior $\propto$ prior $\times$ likelihood», we obtain:

$$\begin{aligned} &\mathrm{P}(\boldsymbol{\Theta} \mid \mathbf{A}, \mathbf{D}, \mathbf{C}, \mathbf{f}, \nu) \propto \\ &\propto \mathrm{P}(\boldsymbol{\Theta} \mid \nu) \times \mathcal{L}(\boldsymbol{\Theta}) = \\ &= \left\{\mathrm{P}(\boldsymbol{\Theta}^{\mathbf{G}} \mid \nu^{\mathbf{G}}) \times \mathrm{P}(\theta^{\mathbf{S}} \mid \nu^{\mathbf{S}})\right\} \times \mathcal{L}(\boldsymbol{\Theta}) = \\ &= \left\{\mathrm{Beta}(\theta_{\mathbf{0}}^{\mathbf{S}} \mid \alpha_0^S, \beta_0^S) \prod_{i=1}^{N_S} \mathrm{Beta}(\theta_{\mathbf{i}}^{\mathbf{S}} \mid \alpha_1^S, \beta_1^S) \times \prod_{l=1}^{N_G} \prod_{t \in \tau} \mathrm{Beta}(\alpha_{l,\tau}^G, \beta_{l,\tau}^G)\right\} \times \mathcal{L}(\boldsymbol{\Theta}) \end{aligned} \tag{1.6}$$

  Parameters α_t^G, β_t^G are selected in such a way that the mode of $\mathrm{Beta}(\alpha_t^G, \beta_t^G)$ equals to $1/t$ [1]

- **Cell-to-clone assignment likelihood**:
  The unknown parameters Θ can be marginalized out:

$$\mathrm{P}(I_j = k \mid \mathbf{A_j}, \mathbf{D_j}, \mathbf{C}, \mathbf{f}) = \int_{\boldsymbol{\Theta}} \mathrm{P}(I_j = k \mid \mathbf{A_j}, \mathbf{D_j}, \mathbf{C}, \mathbf{f}, \boldsymbol{\Theta})\mathrm{P}(\boldsymbol{\Theta} \mid \mathbf{A}, \mathbf{D}, \mathbf{C}, \mathbf{f}, \nu) d\boldsymbol{\Theta} \tag{1.7}$$

  This integral can be evaluated via Monte-Carlo estimation or via analytic computation (maybe).

[1] because if we assume that allelic rates only depend on the CNV status t then those rates could be computed as $1/t$. Given that $\mathbb{E}\ \mathrm{Beta}(\alpha, \beta) = \frac{\alpha}{\alpha+\beta}$, $\alpha = 1,\ \beta = t - 1$ is a feasible solution.

1.3 Inference (Gibbs sampler)

To use a Gibbs sampler, we define conditional probability distribution for each scalar random variable:

1. **Cell-to-clone label assignment:**

$$\begin{aligned}\mathrm{P}(\mathbf{I^S_j} = k \mid \mathbf{I^S_{-j}}, \mathbf{A}, \mathbf{D}, \mathbf{C}, \mathbf{f}, \boldsymbol{\Theta}) = \mathrm{P}(\mathbf{I^S_j} = k \mid \mathbf{A_j}, \mathbf{D_j}, \mathbf{C}, \mathbf{f}, \boldsymbol{\Theta}) \propto \\ \propto \mathrm{P}(\mathbf{I^S_j} = k \mid \mathbf{f})\mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I^S_j} = k, \mathbf{C}, \boldsymbol{\Theta}) =\end{aligned} \tag{1.8}$$

2. **Allelic rates:** Assuming fixed assignment, let's expand the joint likelihood equation:

$$\begin{aligned}
&\mathrm{P}(\boldsymbol{\Theta} \mid \mathbf{A}, \mathbf{D}, \mathbf{C}, \mathbf{I^S}, \mathbf{f}, \nu) \propto \\
&\propto \left\{ \mathrm{Beta}(\theta^{\mathbf{S}}_{\mathbf{0}} \mid \alpha^S_0, \beta^S_0) \prod_{i=1}^{N_S} \mathrm{Beta}(\theta^{\mathbf{S}}_{\mathbf{i}} \mid \alpha^S_1, \beta^S_1) \times \prod_{l=1}^{N_G} \prod_{t\in\tau} \mathrm{Beta}(\alpha^G_{T_{l,t}}, \beta^G_{T_{l,t}}) \right\} \times \\
&\times \prod_{j=1}^{M} \left[\mathrm{P}(\mathbf{A^S_j} \mid \mathbf{D^S_j}, \mathbf{I^S_j}, \mathbf{C}, \theta^{\mathbf{S}}) \times \mathrm{P}(\mathbf{A^G_j} \mid \mathbf{D^G_j}, \mathbf{I^S_j}, \boldsymbol{\Theta}^{\mathbf{G}}) \right] = \\
&= \left\{ \mathrm{Beta}(\theta^{\mathbf{S}}_{\mathbf{0}} \mid \alpha^S_0, \beta^S_0) \prod_{i=1}^{N_S} \mathrm{Beta}(\theta^{\mathbf{S}}_{\mathbf{i}} \mid \alpha^S_1, \beta^S_1) \times \prod_{l=1}^{N_G} \prod_{t\in\tau} \mathrm{Beta}(\alpha^G_{l,t}, \beta^G_{l,t}) \right\} \times \\
&\times \prod_{j=1}^{M} \left[\left(\prod_{i=1}^{N_S} \mathrm{Binom}(\mathbf{A^S_{i,j}} \mid \mathbf{D^S_{i,j}}, \theta^{\mathbf{S}}_{\mathbf{i}})^{\mathbf{H^S_{i,j}}} \times \mathrm{Binom}(\mathbf{A^S_{i,j}} \mid \mathbf{D^S_{i,j}}, \theta^{\mathbf{S}}_{\mathbf{0}})^{\mathbf{1-H^S_{i,j}}} \right) \times \prod_{l=1}^{N_G} \mathrm{Binom}(\mathbf{A^G_{l,j}} \mid \mathbf{D^G_{l,j}}, \mathbf{X^G_{l,j}}) \right] = \\
&= \left[\mathrm{Beta}(\theta^{\mathbf{S}}_{\mathbf{0}} \mid \alpha^S_0, \beta^S_0) \prod_{j=1}^{M} \prod_{i=1}^{N_S} \mathrm{Binom}(\mathbf{A^S_{i,j}} \mid \mathbf{D^S_{i,j}}, \theta^{\mathbf{S}}_{\mathbf{0}})^{\mathbf{1-H^S_{i,j}}} \right] \times \\
&\times \prod_{i=1}^{N_S} \left[\mathrm{Beta}(\theta^S_i \mid \alpha^S_1, \beta^S_1) \prod_{j=1}^{M} \mathrm{Binom}(\mathbf{A^S_{i,j}} \mid \mathbf{D^S_{i,j}}, \theta^{\mathbf{S}}_{\mathbf{i}})^{\mathbf{H^S_{i,j}}} \right] \times \\
&\times \prod_{l=1}^{N_G} \prod_{t\in\tau} \left[\mathrm{Beta}(\alpha^G_{l,t}, \beta^G_{l,t}) \prod_{j=1}^{M} \mathrm{Binom}(\mathbf{A^G_{l,j}} \mid \mathbf{D^G_{l,j}}, \mathbf{X^G_{l,j}})^{\mathbb{I}\{\mathbf{H^G_{l,j}}=t\}} \right]
\end{aligned} \tag{1.9}$$

From here we derive update rules for individual allelic rates:

$$\begin{gathered}\theta^{\mathbf{S}}_{\mathbf{0}} \mid \mathbf{I^S} \sim \mathrm{Beta}(\alpha^S_0 + u^S_i, \beta^S_0 + v^S_i), \quad \theta^{\mathbf{S}}_{\mathbf{i}} \mid \mathbf{I^S} \sim \mathrm{Beta}(\alpha^S_1 + u^S_i, \beta^S_1 + v^S_i), \\ \theta^{\mathbf{G}}_{\mathbf{l,t}} \mid \mathbf{I^S} \sim \mathrm{Beta}(\alpha^G_{l,t} + u^G_{l,t}, \beta^G_{l,t} + v^G_{l,t})\end{gathered} \tag{1.10}$$

where

$$\begin{aligned}
u^S_0 &= \sum_{i=1}^{N_S} \sum_{j=1}^{M} \mathbf{A^S_{i,j}}(1 - \mathbf{H^S_{i,j}}), & v^S_0 &= \sum_{i=1}^{N_S} \sum_{j=1}^{M} (\mathbf{D^S_{i,j}} - \mathbf{A^S_{i,j}})(1 - \mathbf{H^S_{i,j}}), \\
u^S_i &= \sum_{j=1}^{M} \mathbf{A^S_{i,j}} \mathbf{H^S_{i,j}}, \ i > 0, & v^S_i &= \sum_{j=1}^{M} (\mathbf{D^S_{i,j}} - \mathbf{A^S_{i,j}}) \mathbf{H^S_{i,j}}, \ i > 0 \\
u^G_{l,t} &= \sum_{j=1}^{M} \mathbf{A^G_{l,j}} \cdot \mathbb{I}\left\{\mathbf{H^G_{l,j}} = t\right\} & v^G_{l,t} &= \sum_{j=1}^{M} (\mathbf{D^G_{l,j}} - \mathbf{A^G_{l,j}}) \cdot \mathbb{I}\left\{\mathbf{H^G_{l,j}} = t\right\}
\end{aligned} \tag{1.11}$$

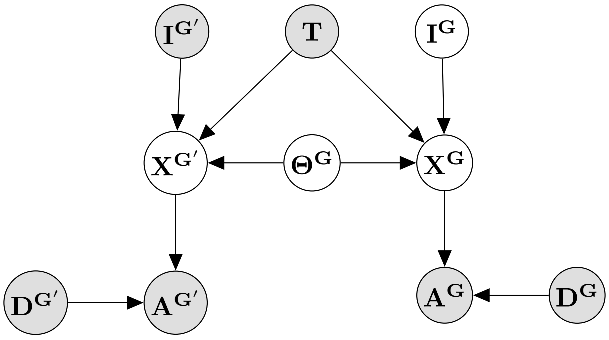
3. **Configuration matrix entries:**

$$\begin{aligned}
&\mathrm{P}(\mathbf{C_{i,k}} = 1 \mid \mathbf{C_{-i,k}}, \mathbf{A}, \mathbf{D}, \mathbf{I^S}, \mathbf{f}, \boldsymbol{\Theta}, \xi) = \\
&= \frac{|\Omega_{i,k} - \xi| \prod_{j=1}^{M} \mathbb{I}(\mathbf{I^S_j} = k) \cdot \mathrm{Binom}(\mathbf{A^S_{i,j}} \mid \mathbf{D^S_{i,j}}, \theta^{\mathbf{S}}_{\mathbf{i}})}{|\Omega_{i,k} - \xi| \prod_{j=1}^{M} \mathbb{I}(\mathbf{I^S_j} = k) \cdot \mathrm{Binom}(\mathbf{A^S_{i,j}} \mid \mathbf{D^S_{i,j}}, \theta^{\mathbf{S}}_{\mathbf{i}}) + |\Omega_{i,k} - (1 - \xi)| \prod_{j=1}^{M} \mathbb{I}(\mathbf{I^S_j} = k) \cdot \mathrm{Binom}(\mathbf{A^S_{i,j}} \mid \mathbf{D^S_{i,j}}, \theta^{\mathbf{S}}_{\mathbf{0}})}
\end{aligned} \tag{1.12}$$

4. **Configuration assignment error rate:** We introduce a prior Beta(κ_0, κ_1) for ξ, then:

$$P(\xi \mid \mathbf{C}, \mathbf{\Omega}, \kappa) = \text{Beta}\left(\kappa_0 + \sum_{i,k} \mathbb{I}\{\mathbf{\Omega_{i,k}} \neq \mathbf{C_{i,k}}\}, \kappa_1 + \sum_{i,k} \mathbb{I}\{\mathbf{\Omega_{i,k}} = \mathbf{C_{i,k}}\}\right) \tag{1.13}$$

2 scDNA + CNV and scRNA, different samples



In this model, clonal assignment $\mathbf{I^{G'}}$ of cells in scDNA sample is assumed fixed. For each clone k, for each same-CNV block i the clonal CNV state $\mathbf{T}_{i,k}$ is defined to be the most frequent CNV state in that position across all cells assigned to clone k. Clonal assignment $\mathbf{I^{G}}$ in scRNA sample is learned.

This model learns the clonal structure in scRNA sample using ASE profiles from scDNA.

2.1 Basic definitions

1. **Constants:**

   - M', M — number of cells in scDNA and scRNA samples respectively.
   - K — estimated number of clones in the sample.
   - N_G — number of same-CNV blocks in scDNA sample (also used in scRNA sample).
   - $T_{\max}$ — maximal possible CNV number. User-defined with the default of 5.
   - τ — set of possible CNV configurations:

     $$\{\{(1,0),(0,1)\},\{(2,0),(1,1),(1,2)\},\ldots,\{(T_{\max},0),\ldots,(0,T_{\max})\}\}$$

     The case of zero CNV number should be treated with care if we can't say for sure whether the part of the chromosome (both arms) is deleted).

2. **Other known quantities:**

   - $\mathbf{D^{G'}}, \mathbf{D^{G}}$ — total read counts. To get a count of the block, one simply adds up the counts of the variants within the block. Here we assume that variants are far enough from each other, so that almost no reads overlap two variants at the same time. Otherwise, adding things up wouldn't make sense.
   - $\mathbf{A^{G'}}$, $\mathbf{A^{G}}$ — same for allele-specific counts.
   - $\mathbf{f} = (f_1, \ldots, f_K)$ — K-dimensional vector of estimated clonal fractions ($\sum_{i=1}^{K} f_i = 1$)
   - $\mathbf{T}$ — CNV states of blocks ($\mathbf{T_{i,k}}$ is a CNV state of a block i in clone k).

3. **Inferred quantities:**

- $\mathbf{\Theta^G}$ — allelic rates of variants located within blocks of fixed CNV status. $\mathbf{\Theta^G_{i,t}}$ is an allelic rate of a block i with a CNV status t. Set of blocks for each clone is unique. Blocks are ordered as tuples of the form (*block start, block length*).
- $\mathbf{I^G} \in [K]^M$ — cell-to-clone assignment in scRNA sample.

4. **Some notational conventions:**
   - Capitalized letter without superscript (like $\mathbf{\Theta}$) denotes the information for both samples.
   - $\mathbf{H^{G'}}, \mathbf{H^G}$ — CNV status of the blocks in accordance with the current label assignment:

$$\mathbf{H^{G'}_{i,j}} := \mathbf{T_{i,I^{G'}_j}}, \quad \mathbf{H^{G}_{i,j}} := \mathbf{T_{i,I^{G}_j}}$$

   - $\mathbf{X^{G'}}$, $\mathbf{X^G}$ — a shortcut to simplify the notation: $\mathbf{X^{G'|G}_{i,j}}$ is an allelic rate of a block i in cell j based of the current cell-to-clone label assignment.

$$\mathbf{X^{G'}_{i,j}} := \mathbf{\Theta^G_{i,H^{G'}_{i,j}}}, \quad \mathbf{X^{G}_{i,j}} := \mathbf{\Theta^G_{i,H^{G}_{i,j}}}$$

2.2 Generative model formulation

- **Cell-to-clone assignment posterior:**

$$\mathrm{P}(\mathbf{I^G_j} = k_0 \mid \mathbf{A_j}, \mathbf{D_j}, \mathbf{f}, \mathbf{\Theta}) = \frac{\mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I^G_j} = k_0, \mathbf{\Theta})\mathrm{P}(\mathbf{I^G_j} = k_0 \mid \mathbf{f})}{\sum\limits_{k=1}^{K} \mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I^G_j} = k, \mathbf{\Theta})\mathrm{P}(\mathbf{I^G_j} = k \mid \mathbf{f})} \tag{2.1}$$

- **ASE model:**

$$\begin{gathered} \mathrm{P}(\mathbf{A^{G'}_{i,j}} \mid \mathbf{D^{G'}_{i,j}}, \mathbf{\Theta^G}) = \mathrm{Binom}(\mathbf{A^{G'}_{i,j}} \mid \mathbf{D^{G'}_{i,j}}, \mathbf{X^{G'}_{i,j}}) \\ \mathrm{P}(\mathbf{A^{G}_{i,j}} \mid \mathbf{D^{G}_{i,j}}, \mathbf{\Theta^G}) = \mathrm{Binom}(\mathbf{A^{G}_{i,j}} \mid \mathbf{D^{G}_{i,j}}, \mathbf{X^{G}_{i,j}}) \end{gathered} \tag{2.2}$$

- **ASE likelihood (both terms factorize over variants):**

$$\begin{gathered} \mathrm{P}(\mathbf{A^{G'}_j} \mid \mathbf{D^{G'}_j}, \mathbf{I^{G'}_j} = k', \mathbf{\Theta^G}) = \prod_{i=1}^{N_G} \mathrm{Binom}(\mathbf{A^{G'}_j} \mid \mathbf{D^{G'}_j}, \mathbf{\Theta^G_{i,k}}) \\ \mathrm{P}(\mathbf{A^{G}_j} \mid \mathbf{D^{G}_j}, \mathbf{I^{G}_j} = k, \mathbf{\Theta^G}) = \prod_{i=1}^{N_G} \mathrm{Binom}(\mathbf{A^{G}_j} \mid \mathbf{D^{G}_j}, \mathbf{\Theta^G_{i,k}}) \end{gathered} \tag{2.3}$$

- **Allelic rate likelihood:**

$$\begin{aligned} \mathcal{L}(\mathbf{\Theta}) = &\left(\prod_{j=1}^{M'} \sum_{k'=1}^{K} \mathrm{P}(\mathbf{A^{G'}_j} \mid \mathbf{D^{G'}_j}, \mathbf{I^{G'}_j} = k', \mathbf{\Theta^G}) \right) \times \\ &\times \left(\prod_{j=1}^{M} \sum_{k=1}^{K} \mathrm{P}(\mathbf{A_j} \mid \mathbf{D_j}, \mathbf{I^G_j} = k, \mathbf{\Theta^G}) \cdot \mathrm{P}(\mathbf{I^G_j} = k \mid \mathbf{f}) \right) \end{aligned} \tag{2.4}$$

  To view the clonal assignment in a Bayesian way, we introduce informative prior ν for $\mathbf{\Theta}$. Using that «posterior $\propto$ prior × likelihood», we obtain:

$$\begin{gathered} \mathrm{P}(\mathbf{\Theta^G} \mid \mathbf{A}, \mathbf{D}, \mathbf{f}, \nu) \propto \mathrm{P}(\mathbf{\Theta} \mid \nu) \times \mathcal{L}(\mathbf{\Theta}) = \\ = \prod_{l=1}^{N_G} \prod_{t \in \tau} \mathrm{Beta}(\alpha^G_{l,\tau}, \beta^G_{l,\tau}) \times \mathcal{L}(\mathbf{\Theta}) \end{gathered} \tag{2.5}$$

  Parameters α^G_t, β^G_t are selected in such a way that the mode of $\mathrm{Beta}(\alpha^G_t, \beta^G_t)$ equals to $1/t$ [2].

[2] because if we assume that allelic rates only depend on the CNV status t then those rates could be computed as $1/t$

2.3 Selecting a prior for $\Theta^{\mathbf{G}}$

CNV state of t hides a plethora of possible configurations: it can mean "k copies of maternal chromosome and $t-k$ copies of paternal" for any $k \in \{0, \dots, t\}$, all the variants are possible. But they are not equally possible: some are more supported by evidence than the rest. During initialization, for each block in each clone we should find the (k,t)-configuration (k_0, t), such that k_0/t is as close to the observed ASE ratio as possible. Then we choose values (α, β) such that the mode of $\mathrm{Beta}(\alpha, \beta)$, given by $(\alpha-1)/(\alpha+\beta-2)$, equals to k_0/t. That means, we must solve the following problem:

$$\frac{\alpha-1}{\alpha+\beta-2} = \frac{k_0}{t}, \ \alpha \geqslant 1, \ \beta \geqslant 1$$

Let's derive the solution. If $k_0 = 0$, it is clear that $\alpha = 1$, while any $\beta > 1$ works[3]. Otherwise:

$$\begin{aligned} (\alpha-1)t &= (\alpha+\beta-2)k_0 \\ k_0\beta &= (t-k_0)\alpha - t + 2k_0 \\ \beta &= \left(\frac{t}{k_0} - 1\right)\alpha - \left(\frac{t}{k_0} - 2\right) \\ \implies \alpha &= 1 + \frac{t-2k_0}{t-k_0}, \ \beta = 1 \end{aligned} \tag{2.6}$$

As β is linearly dependent from α, any increase in α will pull β up, "sharpening" the shape of the distribution and making it more biased, thereby we decided to choose the minimal feasible α.

2.4 Inference (Gibbs sampler)

To use a Gibbs sampler, we define conditional probability distribution for each scalar random variable:

1. **Cell-to-clone label assignment:**

$$\mathrm{P}(\mathbf{I_j^G} = k \mid \mathbf{I_{-j}^G}, \mathbf{A^G}, \mathbf{D^G}, \mathbf{f}, \mathbf{\Theta^G}) \propto \mathrm{P}(\mathbf{A_j^G} \mid \mathbf{D_j^G}, \mathbf{I_j^G} = k, \mathbf{\Theta}^G) \cdot \mathrm{P}(\mathbf{I_j^G} = k \mid \mathbf{f}) \tag{2.7}$$

2. **Allelic rates:** Assuming fixed assignment, let's expand the joint likelihood equation:

$$\begin{aligned} &\mathrm{P}(\mathbf{\Theta} \mid \mathbf{A}, \mathbf{D}, \mathbf{I^{G'}}, \mathbf{I^G}, \mathbf{f}, \nu) \propto \\ &\propto \left\{\prod_{l=1}^{N_G}\prod_{t\in\tau}\mathrm{Beta}(\alpha^G_{T_{l,t}}, \beta^G_{T_{l,t}})\right\} \times \left[\prod_{j=1}^{M'}\mathrm{P}(\mathbf{A_j^{G'}} \mid \mathbf{D_j^{G'}}, \mathbf{I_j^{G'}}, \mathbf{\Theta^G})\right] \times \left[\prod_{j=1}^{M}\mathrm{P}(\mathbf{A_j^G} \mid \mathbf{D_j^G}, \mathbf{I_j^G}, \mathbf{\Theta^G})\right] = \\ &= \left\{\prod_{l=1}^{N_G}\prod_{t\in\tau}\mathrm{Beta}(\alpha^G_{T_{l,t}}, \beta^G_{T_{l,t}})\right\} \times \left[\prod_{j=1}^{M'}\prod_{i=1}^{N_G}\mathrm{Binom}(\mathbf{A_j^{G'}} \mid \mathbf{D_j^{G'}}, \mathbf{\Theta^G_{i,k}})\right] \times \left[\prod_{j=1}^{M}\prod_{i=1}^{N_G}\mathrm{Binom}(\mathbf{A_j^G} \mid \mathbf{D_j^G}, \mathbf{\Theta^G_{i,k}})\right] = \\ &= \prod_{l=1}^{N_G}\prod_{t\in\tau}\left[\mathrm{Beta}(\alpha^G_{l,t}, \beta^G_{l,t})\prod_{j'=1}^{M'}\prod_{j=1}^{M}\mathrm{Binom}(\mathbf{A^{G'}_{l,j'}} \mid \mathbf{D^{G'}_{l,j'}}, \mathbf{X^{G'}_{l,j'}})^{\mathbb{I}\left\{\mathbf{H^{G'}_{l,j'}}=t\right\}}\mathrm{Binom}(\mathbf{A^G_{l,j}} \mid \mathbf{D^G_{l,j}}, \mathbf{X^G_{l,j}})^{\mathbb{I}\{\mathbf{H^G_{l,j}}=t\}}\right] \end{aligned} \tag{2.8}$$

From here we derive update rules for individual allelic rates:

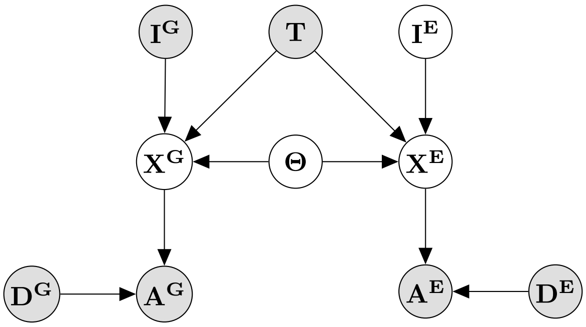
$$\mathbf{\Theta^G_{l,t}} \mid \mathbf{I^{G'}}, \mathbf{I^G} \sim \mathrm{Beta}(\alpha^G_{l,t} + u^G_{l,t}, \beta^G_{l,t} + v^G_{l,t}) \tag{2.9}$$

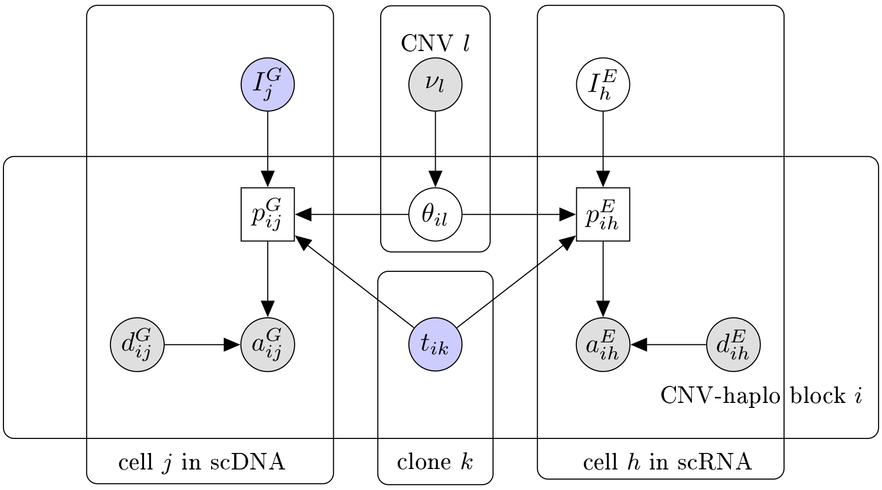
where

$$\begin{aligned} u^G_{l,t} &= \sum_{j'=1}^{M'}\mathbf{A^{G'}_{l,j'}} \cdot \mathbb{I}\left\{\mathbf{H^{G'}_{l,j'}} = t\right\} + \sum_{j=1}^{M}\mathbf{A^G_{l,j}} \cdot \mathbb{I}\left\{\mathbf{H^G_{l,j}} = t\right\} \\ v^G_{l,t} &= \sum_{j'=1}^{M'}(\mathbf{D^{G'}_{l,j'}} - \mathbf{A^{G'}_{l,j'}}) \cdot \mathbb{I}\left\{\mathbf{H^{G'}_{l,j'}} = t\right\} + \sum_{j=1}^{M}(\mathbf{D^G_{l,j}} - \mathbf{A^G_{l,j}}) \cdot \mathbb{I}\left\{\mathbf{H^G_{l,j}} = t\right\} \end{aligned} \tag{2.10}$$

[3] Nevertheless, it is not clear which one to choose. As we try to reduce prior bias, let's set it to be equal $1 + \varepsilon$ for some reasonable $\varepsilon > 0$

3 Notes





$$p_{i,j} = \theta_{i,l_j}; l_j = t_{i,I_j}$$

$$P(a_{i,j}|d_{i,j}, p_{i,j}) = \text{Binom}(a_{i,j}; d_{i,j}, p_{i,j})$$