>>> 基于测序数据的肿瘤亚克隆重组

Name: 初砚硕 (计算机学院，生物信息)
Date: March 31, 2017

>>> 三个“?”

为什么肿瘤/癌症难以治疗?

肿瘤是由多种细胞组成的复杂混合体，大致包括发生变异的肿瘤细胞、未发生变异的肿瘤细胞、正常细胞（这种现象称为**异质性**）。一种治疗方法，只能针对个别成分发生作用，未被杀死的肿瘤细胞会迅速占据被杀死细胞的空间和营养，导致病情越治越坏。

什么是亚克隆?

目前没有一个标准的定义[1]。肿瘤中的细胞按照基因组相似度聚类（类别数未知），每一个类就是亚克隆。

亚克隆重组有什么意义?

获得肿瘤的组成成分有利于制定更清晰的肿瘤治疗方案。获得肿瘤组成成分的过程称为**亚克隆重组**。

---

[1] Charles Gawad, Winston Koh, and Stephen R Quake. “Single-cell genome sequencing: current state of the science”. In: *Nature Reviews Genetics* 17.3 (2016), pp. 175–188.

>>> 肿瘤的异质性是进化过程

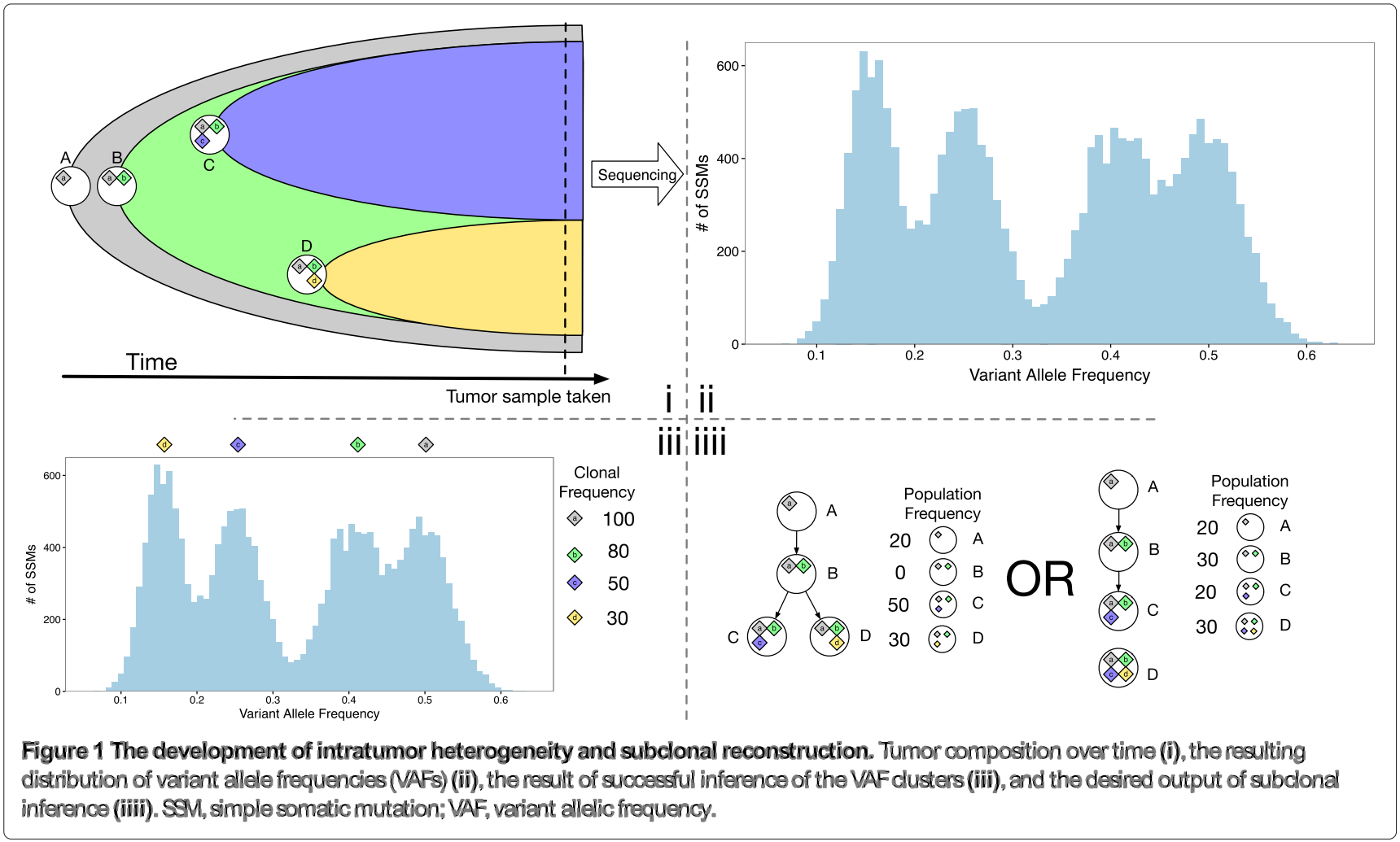


Figure 1 The development of intratumor heterogeneity and subclonal reconstruction. Tumor composition over time **(i)**, the resulting distribution of variant allele frequencies (VAFs) **(ii)**, the result of successful inference of the VAF clusters **(iii)**, and the desired output of subclonal inference **(iiii)**. SSM, simple somatic mutation; VAF, variant allelic frequency.

>>> 目录

>>> 目录

>>> 肿瘤/正常样本的测序数据

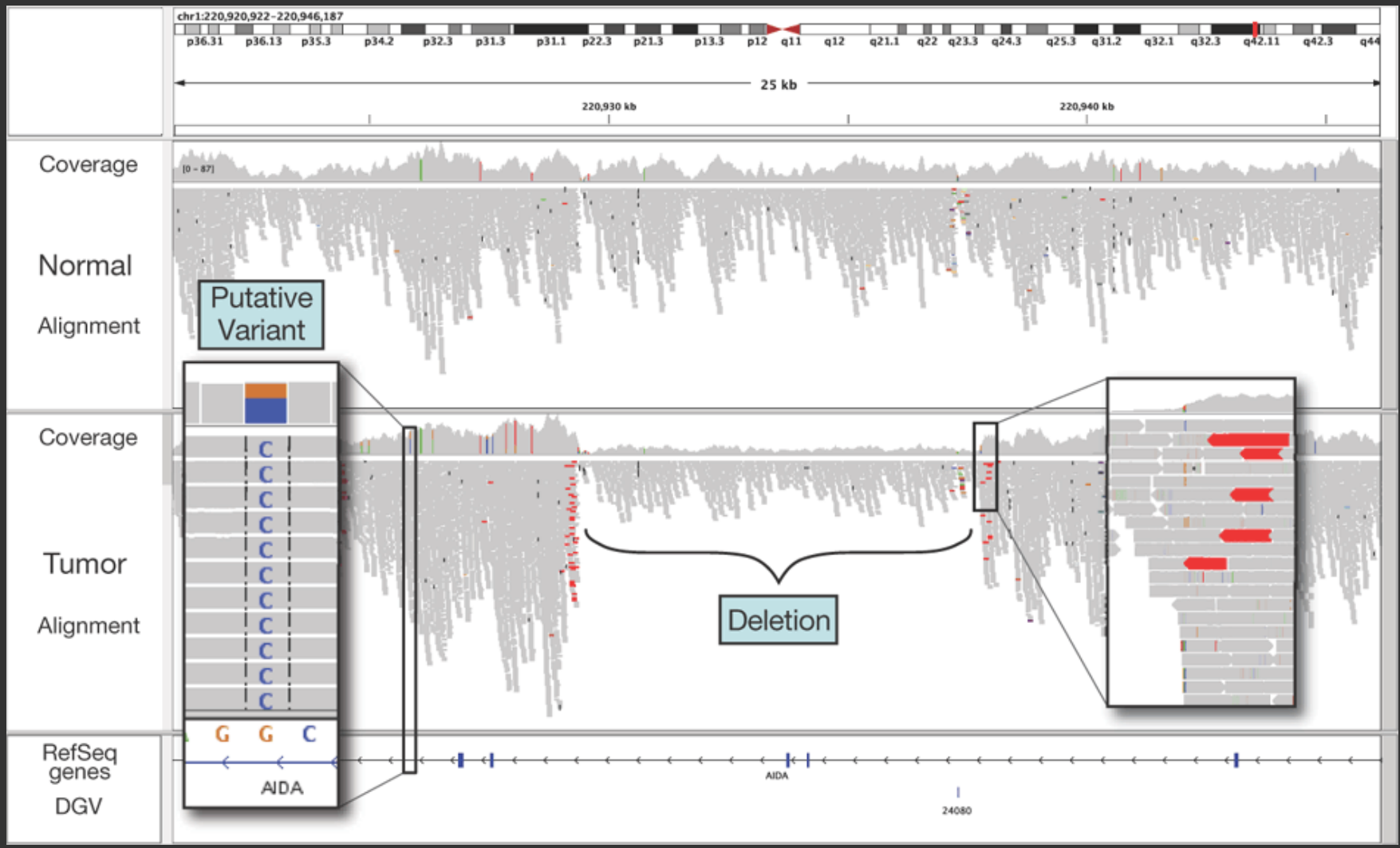
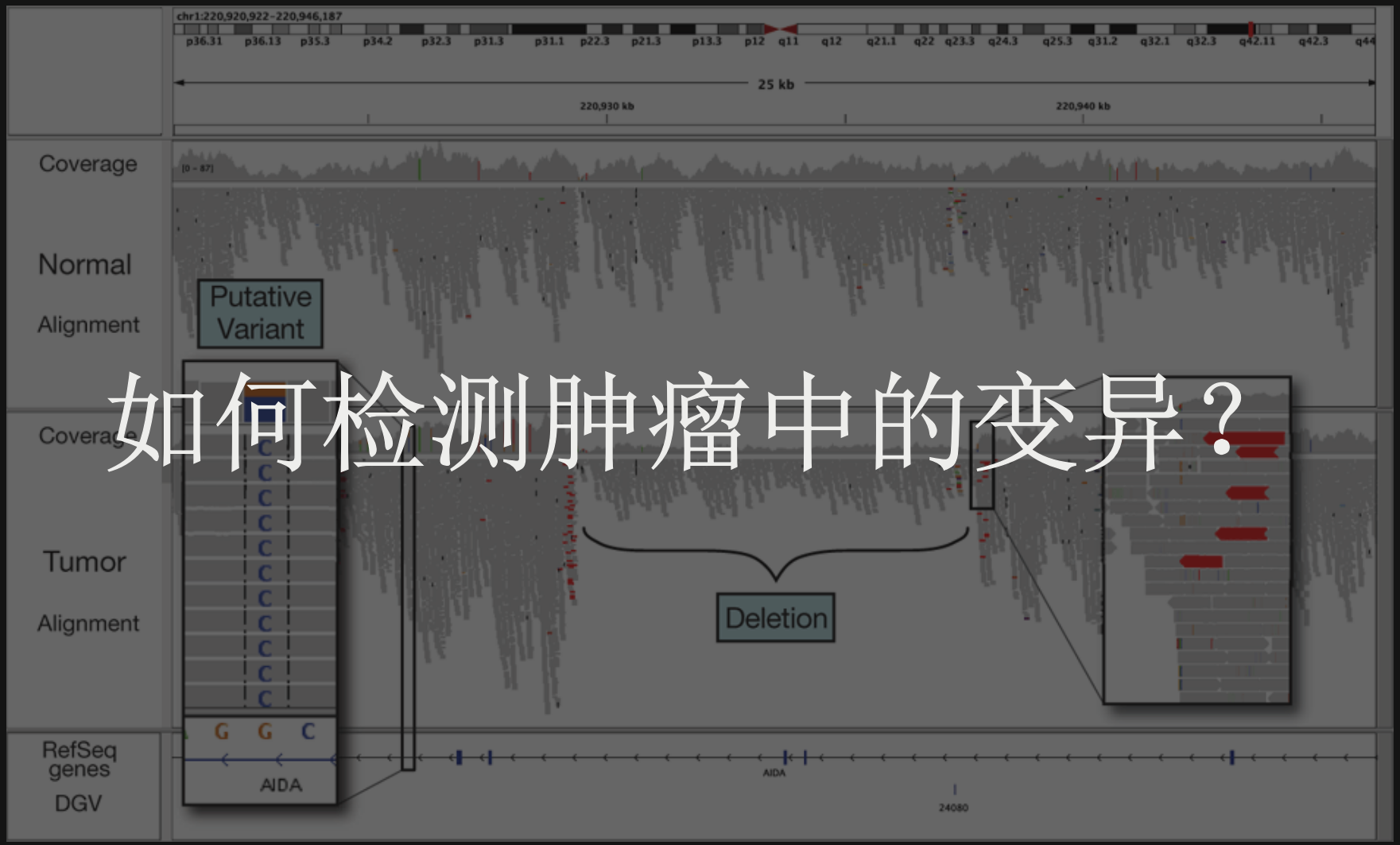


图 1： 肿瘤和正常样本配对测序（Tumor nomral paired sequencing）



如何检测肿瘤中的变异？

图 1： 肿瘤和正常样本配对测序（Tumor nomral paired sequencing）

1.1 SCNA 检测流程

>>> SCNA 检测流程

* 根据覆盖度片段 Segmentation (BIC-seq)
* 根据每一个片段（segment）之内的 read 性质，确定 SCNA 类型
  * 覆盖度 read depth/count
  * 等位型 allele type

>>> SCNA 检测流程

* 根据覆盖度片段 Segmentation (BIC-seq)
* 根据每一个片段（segment）之内的 read 性质，确定 SCNA 类型
  * 覆盖度 read depth/count
  * 等位型 allele type

影响覆盖度的因素

* 平均拷贝数 Average copy number
* Mappability (GC bias)

现有的对影响覆盖度的因素建模的方法如下[2]，使用 θ_j 表示第 j 个片段内的由片段长度和映射性（mappability）造成的不均匀性，$\bar{C}_j$ 表示该片段的平均拷贝数，D_j^N 表示在正常样本中的这个片段内的 read 数, D_j^T 表示在肿瘤样本中的这个片段内的 read 数, 那么对于片段 i 和片段 j，有 $D_i^T/D_i^T = \bar{C}_i\theta_i/\bar{C}_j\theta_j$，其中 $\theta_i/\theta_j = D_i^N/D_j^N$. 如果

$$\frac{D_i^T}{D_j^T} = \frac{\bar{C}_i\theta_i}{\bar{C}_j\theta_j} = \frac{\bar{C}_i}{\bar{C}_j} * \frac{D_i^N}{D_j^N}, \tag{1}$$

那么，

$$\log\frac{D_i^T}{D_i^N} - \log\frac{D_j^T}{D_j^N} = \log\frac{\bar{C}_i}{\bar{C}_j}. \tag{2}$$

[2] Yi Li and Xiaohui Xie. "MixClone: a mixture model for inferring tumor subclonal populations". In: *BMC genomics* 16.Suppl 2 (2015), S1.

现有的对影响覆盖度的因素建模的方法如下[2]，使用 θ_j 表示第 j 个片段内的由片段长度和映射性（mappability）造成的不均匀性，$\bar{C}_j$ 表示该片段的平均拷贝数，D_j^N 表示在正常样本中的这个片段内的 read 数，D_j^T 表示在肿瘤样本中的这个片段内的 read 数，那么对于片段 i 和片段 j，有 $D_i^T/D_i^T = \bar{C}_i\theta_i/\bar{C}_j\theta_j$，其中 $\theta_i/\theta_j = D_i^N/D_j^N$. 如果

$$\frac{D_i^T}{D_j^T} \approx \frac{\bar{C}_i\theta_i}{\bar{C}_j\theta_j} = \frac{\bar{C}_i}{\bar{C}_j} * \frac{D_i^N}{D_j^N} \tag{1}$$

公式 (2) 说明，对于所有的片段 $i = 1, \ldots, m$，$\log \frac{D_i^T}{D_i^N}$ 可以进行近邻聚类。

那么，

$$\log \frac{D_i^T}{D_i^N} - \log \frac{D_j^T}{D_j^N} = \log \frac{\bar{C}_i}{\bar{C}_j}. \tag{2}$$

[2] Yi Li and Xiaohui Xie. "MixClone: a mixture model for inferring tumor subclonal populations". In: *BMC genomics* 16.Suppl 2 (2015), S1.

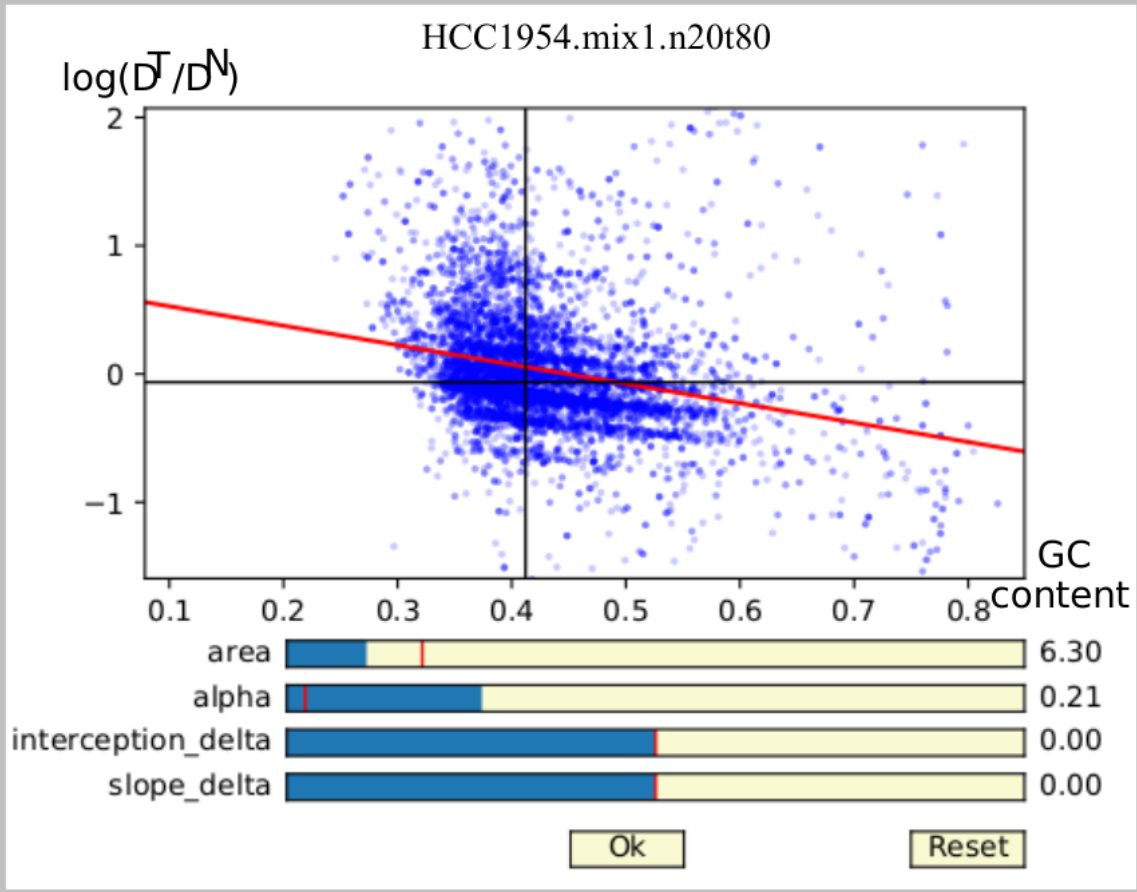


Fig. 1. Coupling GC bias of tumor sample 'HCC1954.mix1.n20t80'. Red line denotes the linear regression line, the black vertical and horizontal lines mark the median of GC content and $\log D_i^T/D_i^N$ respectively. In this figure, $\log D_i^T/D_i^N$ of the stripes vertically decreases along with the increase of GC content, while, the GC content of the stripes horizontally increases along with the decrease of $\log D_i^T/D_i^N$. The linear regression line is not parallel to the stripes, due to the horizontal GC bias.

图 2： Stripe

1.2 Pre-SCNAClonal — 肿瘤/正常样本全基因组测序数据的一种二维 GC-bias 校正和可视化工具

>>> 目录

2.1 VAF 信号校正

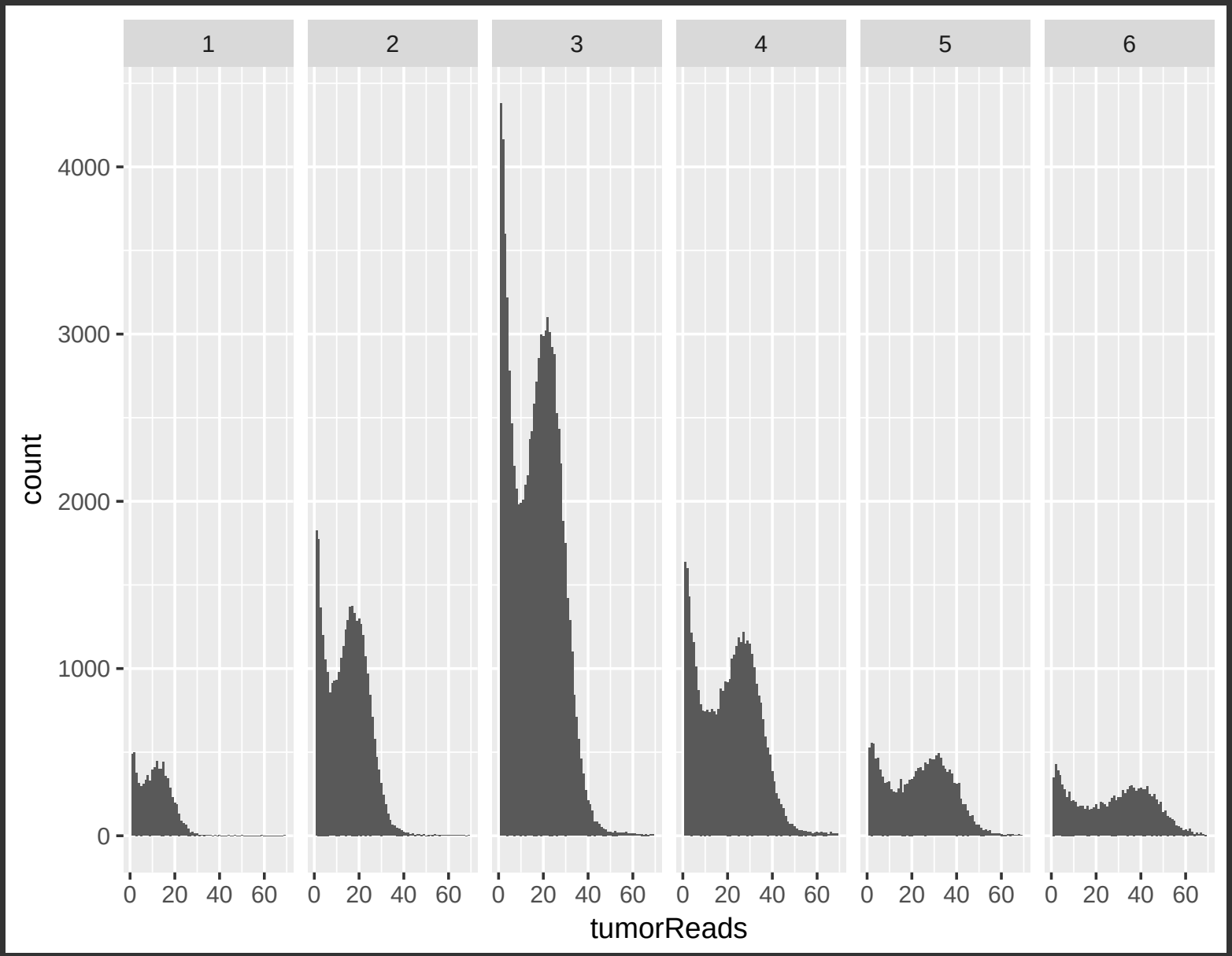


图 3： BAF read count distribution

>>> VAF 信号过滤（平滑）

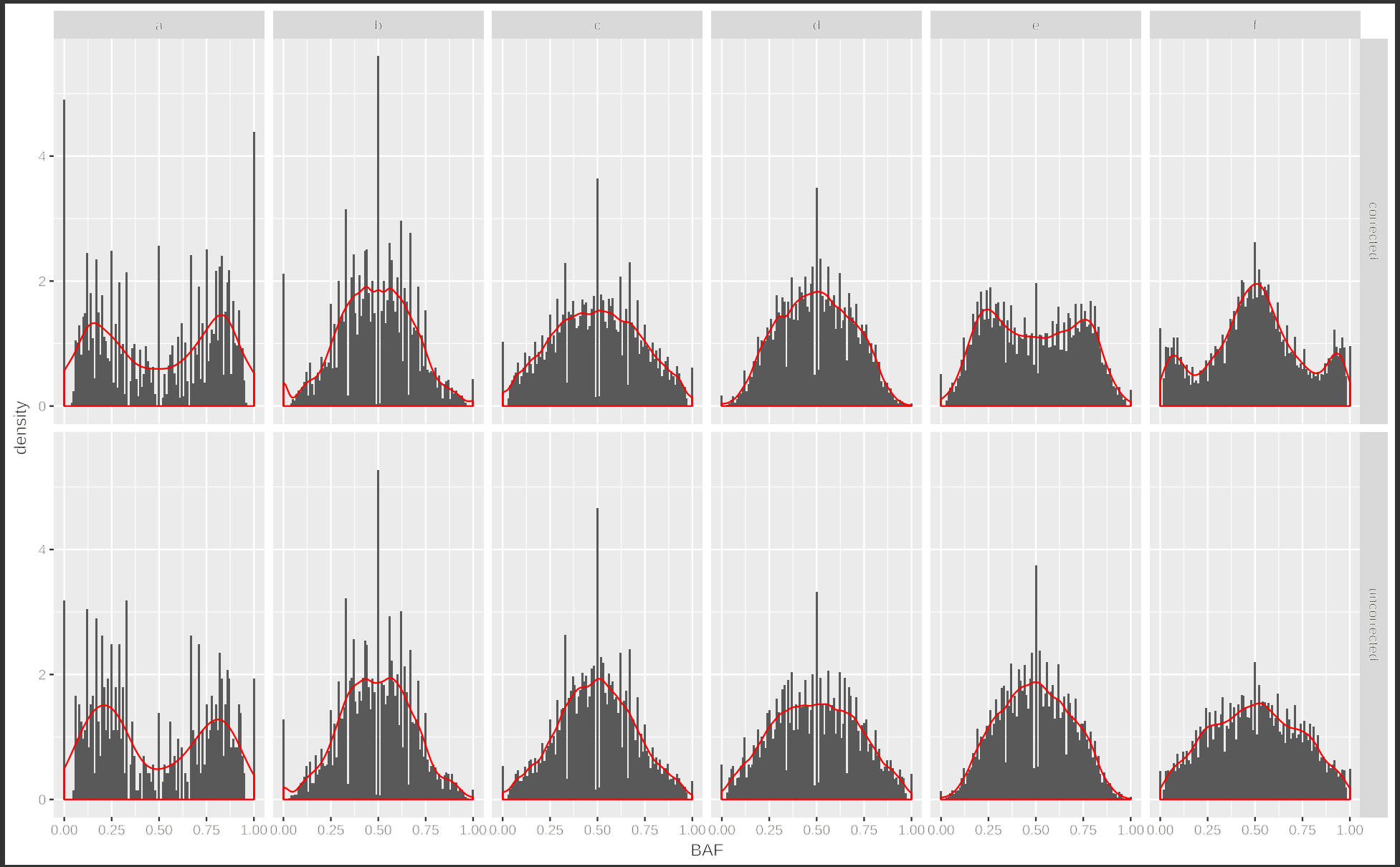


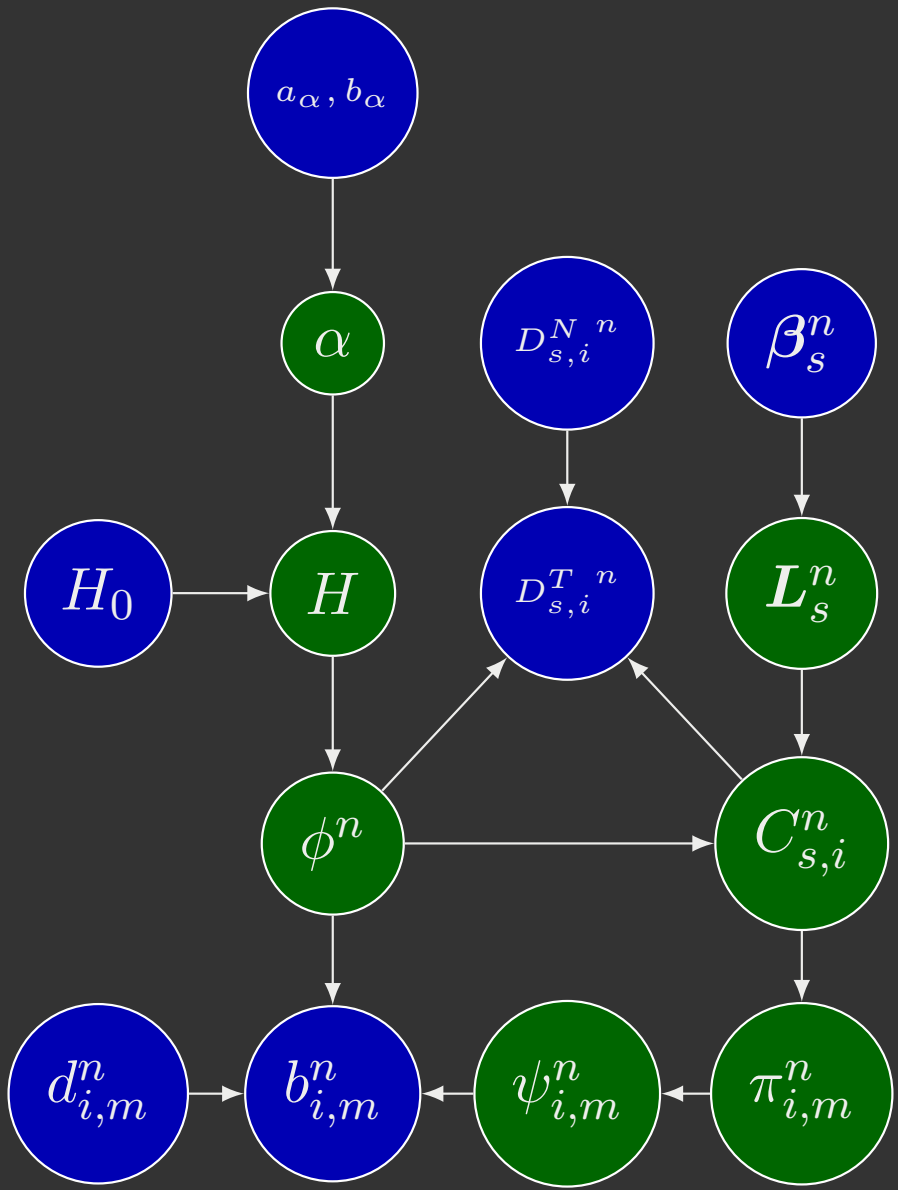
图 4： BAF distribution

2.2 贝叶斯概率模型

如何利用（条带的）先验信息？

* 平均拷贝数相等
* 基线拷贝数为 2
* 其余条带的平均拷贝数由下到上递增

>>> 概率模型



$$\alpha \sim \mathsf{Gamma}(a_\alpha, b_\alpha)$$

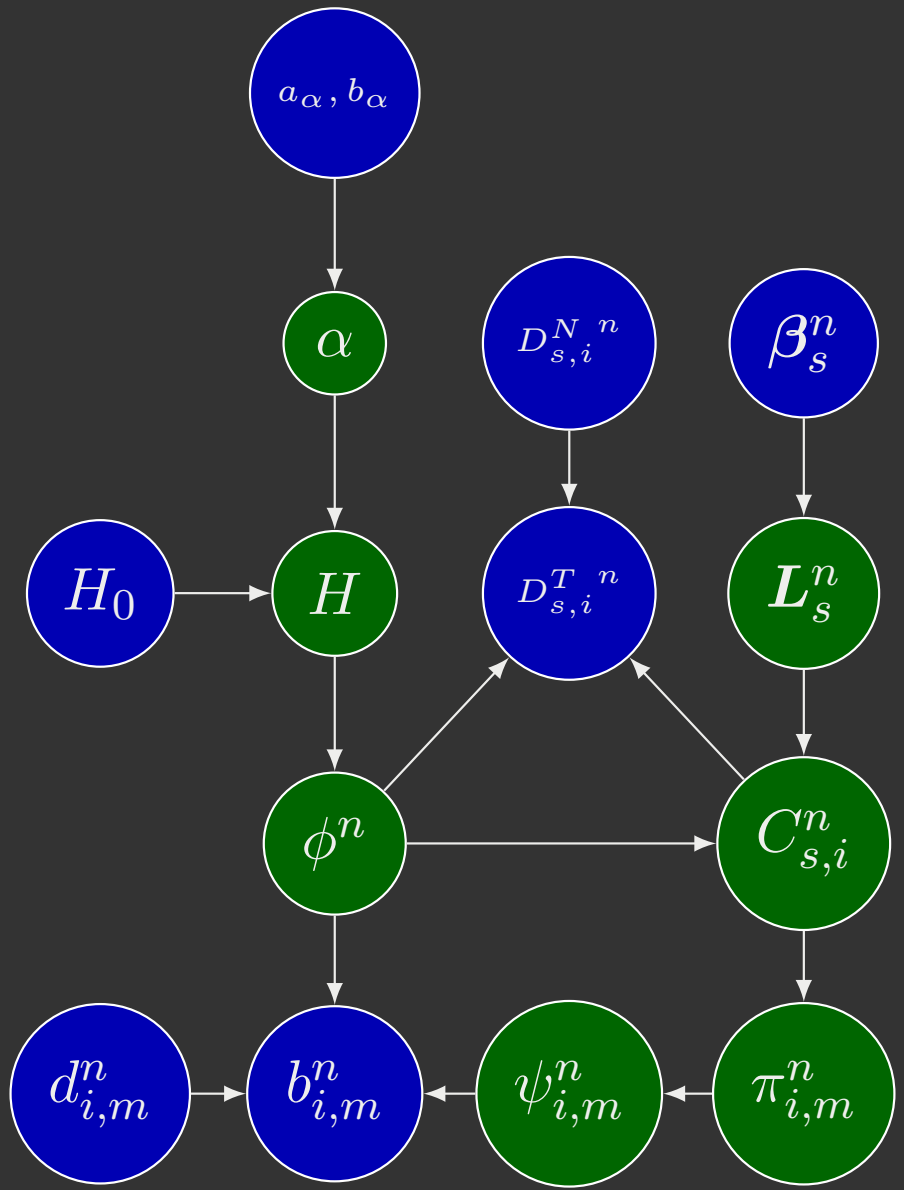
$$H_0 \sim \mathsf{Uniform}([0,1]^M)$$

$$H \sim \mathsf{DP}(H_0, \alpha)$$

$$\phi^n \sim H$$

$$C \sim \mathsf{Categorical}(\boldsymbol{L}_s^n)$$

$$\boldsymbol{L}_s^n \sim \mathsf{Dir}(\boldsymbol{\beta}_s^n)$$



$$\psi_{i,m}^n | \pi_{i,m}^n \sim \text{Categorical}(\pi_{i,m}^n)$$

$$\pi_{i,m}^n | C_{s,i}^n \sim \text{Categorical}(C_{s,i}^n)$$

$$b_{i,m}^n | d_{i,m}^n, \psi_{i,m}^n, \phi^n \sim \text{Binomial}(d_{i,m}^n, \xi(\phi^n, \psi_{i,m}^n))$$

$$\xi = \frac{\phi^n * C * \mu + (1 - \phi^n) * 2 * \frac{1}{2}}{\phi^n * C + (1 - \phi^n) * 2}$$

$$D_{s,i}^{T\ n} \sim \text{Possion}\left(\frac{\overline{C}_{s,i}^n}{2} * \sqrt[J]{\prod_{j=1}^{J} \frac{D_j^T}{D_j^N}} * D_{s,i}^{N\ n}\right)$$

>>> 概率模型参数

$$\pi_{i,m}^{n} \in \{\oslash, \mathsf{P}, \mathsf{M}, \mathsf{PP}, \mathsf{PM}, \mathsf{MM}, \mathsf{PPP}, \ldots, \mathsf{MMMMMMM}\}$$

$$C_{s,i}^{n} \in \{0, 1, 2, \ldots, 7\}$$

$$\boldsymbol{L}_{s}^{n} = \{l_0, l_1, \ldots, l_7\}$$

$$\boldsymbol{\beta}_{s}^{n} = \{\beta_{s0}^{n}, \beta_{s1}^{n}, \ldots, \beta_{s7}^{n}\}$$

$$j = 0, \ldots, 7$$

$$\beta_{sj}^{n} = \begin{cases} \frac{0.8}{m}, & \text{if } j = s \\ \frac{0.2}{8-m}, & \text{if } j \neq s \end{cases}$$

$$
\begin{aligned}
P(\phi, C|D, b) &\propto P(D, b|\phi, C) * P(\phi) * P(C|\phi) \\
&\propto P(D|\phi, C) * P(b|\phi, C) * P(\phi) * P(C|\phi) \\
&= P(\phi) * P(C|\phi) *
\end{aligned}
$$

$$
\prod_{n=1}^{N} \prod_{s=1}^{S} \prod_{i=1}^{I} \left[P\left(\left\{ D_{s,j}^{T} \right\}^{n} | \left\{ D_{s,j}^{N} \right\}^{n}, \phi^{n}, C_{s,i}^{n} \right) \prod_{m=1}^{M} P\left(b_{i,m}^{n} | d_{i,m}^{n}, \phi^{n}, \psi_{i,m}^{n} \right) \right]
$$