

Supplementary Materials

Additional file 1: DPM search method

We adopt a nonparametric Bayesian Dirichlet process mixture (DPM) model to clustering the m sequence reads. Denote by $\{\mathbf{X}_i, i = 1, \dots, m\}$ the m independent sequence reads, with $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{in})^T$ a binary vector at n sites. The DPM model can be written as:

$$\begin{aligned} X_{ij} &\sim \text{Bernoulli}(p_{ij}), \quad j = 1, \dots, n, \quad i = 1, \dots, m \\ \mathbf{p}_i &= (p_{i1}, \dots, p_{in})^T \\ \mathbf{p}_i &\sim G, \quad G \sim \text{DP}(\alpha, G_0), \\ G_0 &= \prod_{j=1}^n G_{0j}, \quad G_{0j} \sim \text{Beta}(\alpha_j, \beta_j), \end{aligned}$$

where G is a random distribution of \mathbf{p}_i which is given a Dirichlet process (DP) prior with concentration parameter α and base measure G_0 . The base measure G_0 is formed by independent Beta distributions with pre-specified prior parameters $\{(\alpha_j, \beta_j), j = 1, \dots, n\}$. The DP prior has the effect of grouping similar \mathbf{p}_i 's into clusters hence is widely used for clustering.

While the posterior inference of the above model can be achieved through Markov chain Monte Carlo sampling, these sampling methods are often slow to implement. In this work, we employ a fast search algorithm to find the maximum a posteriori (MAP) solution, following the method of Daumé III (2007)²⁴. The main idea is to set a score function, and sequentially assign the ordered data points $\{\mathbf{X}_i, i = 1, \dots, m\}$ either to one of the existing clusters or to a new cluster based on the score function. We choose a fast in-admissible score function which has demonstrated extremely good performance on MAP search. In our algorithm, we need to determine the marginal joint likelihood: $H(\mathbf{X}_S) = \int f(\mathbf{X}_S|\mathbf{p})dG_0(\mathbf{p})$ and a conditional probability of a new \mathbf{X} conditional on \mathbf{X}_S : $H(\mathbf{X}|\mathbf{X}_S) = H(\mathbf{X}, \mathbf{X}_S)/H(\mathbf{X}_S)$, for $\mathbf{X}_S = \{\mathbf{X}_i : i \in S\}$ and S is an index set. Here f is the joint likelihood of \mathbf{X}_S assuming that the elements in \mathbf{X}_S are independent Bernoulli with a common parameter \mathbf{p} . Due to the fact that the Beta prior is conjugate to Bernoulli likelihood, we can calculate $H(\mathbf{X}_S)$ and $H(\mathbf{X}|\mathbf{X}_S)$ analytically.