Bayesian hierarchical modelling of single-cell methylation profiles

{ C.A.Kapourani and G.Sanguinetti }@ed.ac.uk

Motivation

- Bulk BS-seq technology ignores epigenetic heterogeneity among individual cells.
- Single cell BS-seq[1] measures methylation at single-cell level (binary states).
- Only 20% to 40% CpG coverage: limits statistical analysis to a semi-quantitative level.
- Bayesian hierarchical model: jointly learn methylation profiles (i.e. predict uncovered methylation states) and cluster cells based on genome-wide methylation patterns.

scBPR model

Bayesian generalised linear model (GLM) of basis function regression coupled with a Bernoulli observation model[2], where y_i is the methylation state, $\boldsymbol{h}_i$ the genomic location and $\boldsymbol{w}$ the model parameters.

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$z_i \sim \mathcal{N}(\boldsymbol{h}_i \boldsymbol{w}, 1)$$

$$\boldsymbol{w} = p(\boldsymbol{w})$$

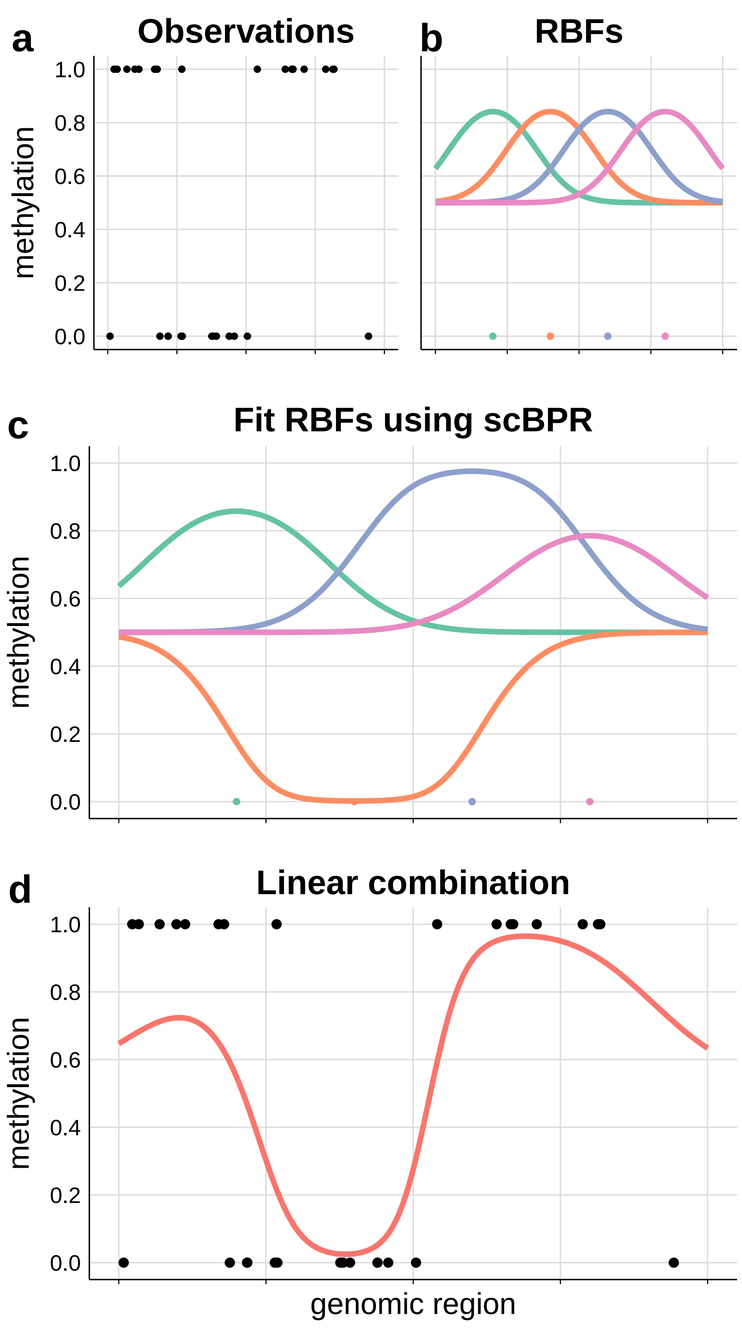


Figure 1: Process of learning methylation profiles using the single-cell Bernoulli Probit Regression (scBPR) model for a specific genomic region using 4 Radial Basis Functions (RBFs).

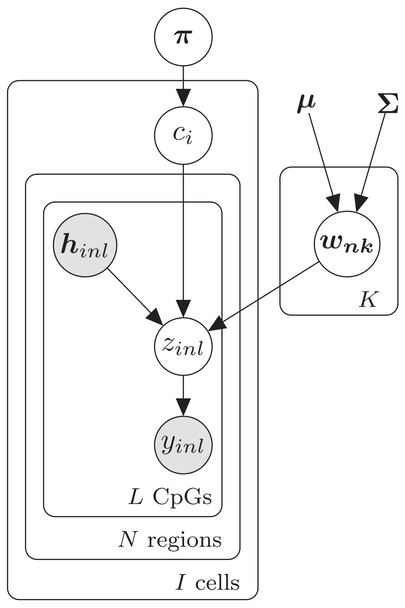
References

1. Smallwood, S.A. *et.al*, 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, **11**(8), pp.817-820.
2. CAK & GS, 2016. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, **32**(17), pp.i405-i412.
3. Angermueller, C. *et.al*, 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods.*, **13**(3), pp.229-232.

scBPR Finite Dirichlet Mixture Model

Joint posterior distribution

$$p(\mathbf{W}, \mathbf{Z}, \mathbf{c}, \boldsymbol{\pi} | \mathbf{H}, \mathbf{Y}) \propto \Big\{ \prod_i \prod_n p(\mathbf{y}_{in} | \mathbf{z}_{in}) p(\mathbf{z}_{in} | \mathbf{w}_{n,c_i}, \mathbf{H}_{in}, c_i) p(c_i | \boldsymbol{\pi}) \Big\} p(\boldsymbol{\pi}) \prod_n \prod_k p(\mathbf{w}_{nk})$$



Algorithm 1 Gibbs sampling for scBPR FDMM model

initialize Set $t \leftarrow 1$; set clusters K; set parameters $\boldsymbol{\pi}^{(0)}, \mathbf{W}^{(0)}$
while $t \leq T$ **do**
  Compute $\gamma(c_{ik})$, probability that cell i belongs to cluster k
  Generate $c_i^{(t)} \sim \mathcal{D}iscrete\big(\gamma(c_{i\cdot})\big)$
  Generate $\boldsymbol{\pi}^{(t)} \sim \mathcal{D}ir\Big(\big\{\alpha_k + \sum_i \mathbf{1}(c_i^{(t)} = k)\big\}_{k=1}^{K}\Big)$
  Generate $\mathbf{z}_{in}^{(t)} \sim \begin{cases} \mathcal{TN}(\mathbf{h}_{inl}\mathbf{w}_{nk}^{(t-1)}, 1, 0, \infty) & \text{if } y_{inl} = 1 \\ \mathcal{TN}(\mathbf{h}_{inl}\mathbf{w}_{nk}^{(t-1)}, 1, -\infty, 0) & \text{if } y_{inl} = 0 \end{cases}$
  Generate $\mathbf{w}_{nk}^{(t)} \sim \mathcal{N}\Big(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\Big)$
end while

Identifying cell sub-populations

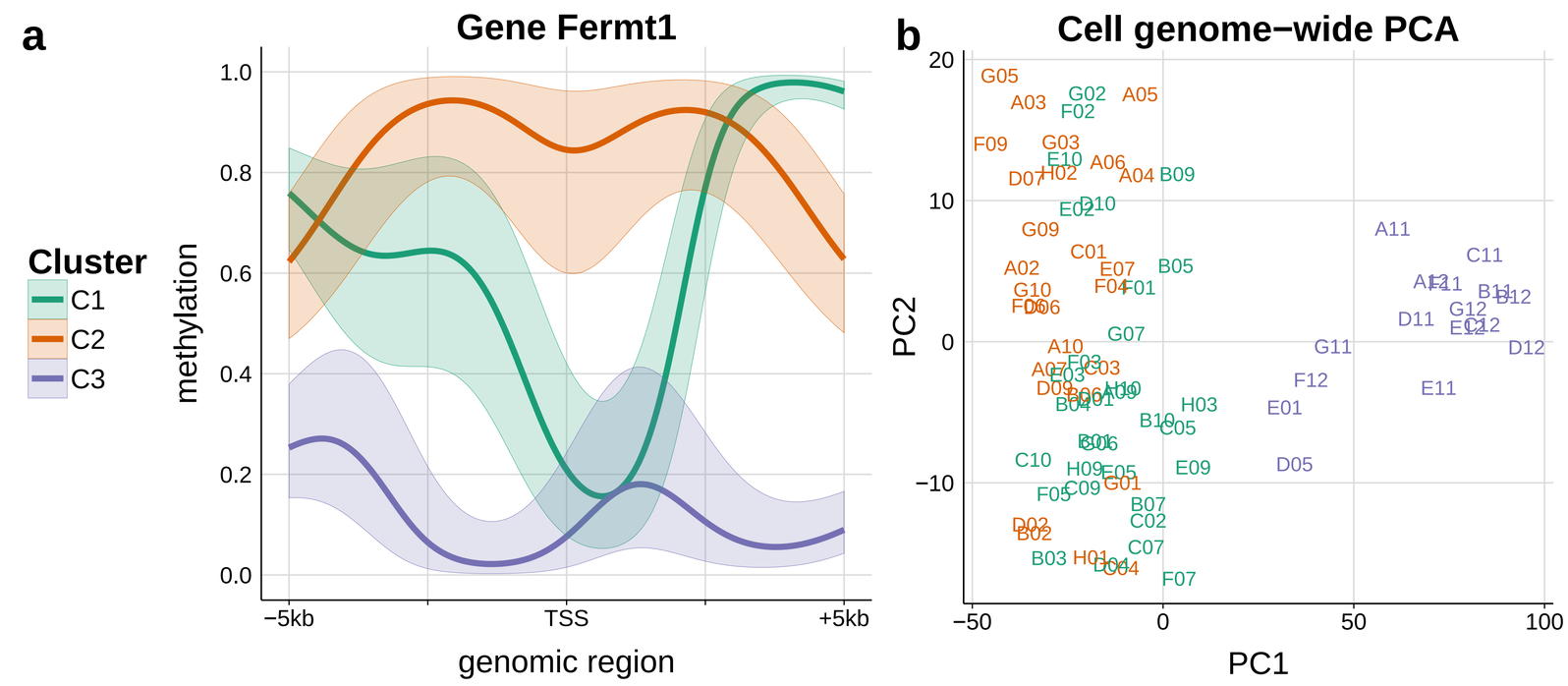


Figure 2: Clustering cells based on promoter methylation profiles on a real dataset[3]. **(a)** Methylation profiles for each sub-population of cells for gene Fermt1. **(b)** PCA of cells based on mean genome-wide methylation patterns; blue coloured cells are $2i$ cells which distinguish well from serum-cultured cells even based on mean methylation.

Imputing methylation states

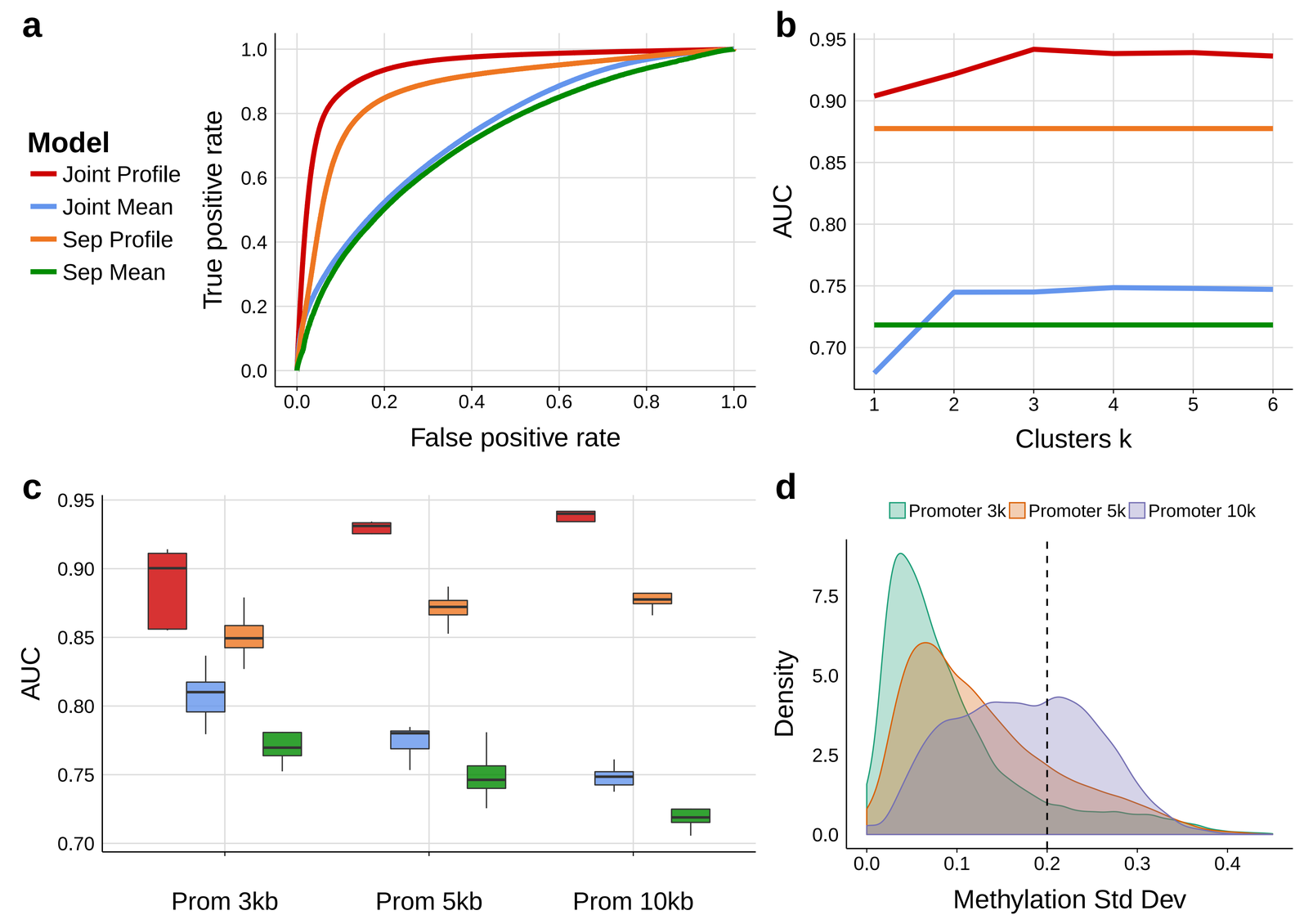


Figure 3: Imputation performance on real data[3]. **(a)** ROC curves for $10kb$ promoter windows and $K = 3$ clusters. **(b)** AUC while increasing the clusters for $10kb$ window. **(c)** Boxplot of AUCs for varying promoter windows, each dot represents a different experiment. **(d)** Mean methylation variability across cells on different promoter windows.