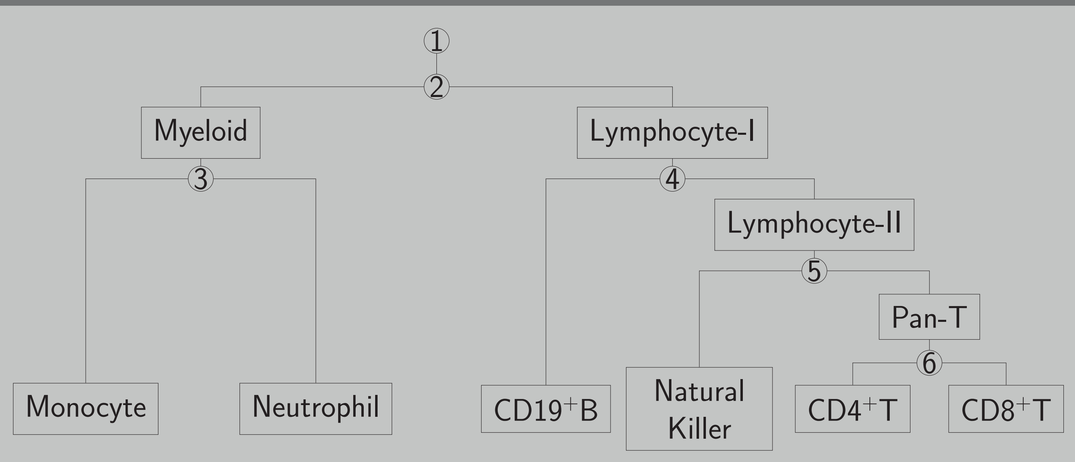
Cell-type specific analysis of heterogeneous methylation signal

Daniel W. Kennedy, Nicole M. White, Miles C. Benton, Lyn Griffiths, Rodney A. Lea, and Kerrie Mengersen

ARC Centre of Excellence for Mathematical and Statistical Frontiers

QUT Institute of Health and Biomedical Innovation

Figure: Haematopoeitic Lineage



Lineage Matrix:

$$\mathbf{A} = \begin{array}{c} \\ \text{Monocyte} \\ \text{Neutrophil} \\ \text{CD19}^{+}\text{B} \\ \text{Natural Killer} \\ \text{CD4}^{+}\text{T} \\ \text{CD8}^{+}\text{T} \end{array} \begin{array}{c} \begin{array}{cccccc} \textbf{Node:} & \textbf{1} & \textbf{2} & \textbf{3} & \textbf{4} & \textbf{5} \end{array} \\ \begin{pmatrix} 1 & -1/2 & -1/2 & 0 & 0 & 0 \\ 1 & -1/2 & +1/2 & 0 & 0 & 0 \\ 1 & +1/2 & 0 & -1/2 & 0 & 0 \\ 1 & +1/2 & 0 & +1/2 & -1/2 & 0 \\ 1 & +1/2 & 0 & +1/2 & +1/2 & -1/2 \\ 1 & +1/2 & 0 & +1/2 & +1/2 & +1/2 \end{pmatrix} \end{array}$$

Introduction

- Methylation arrays from blood are the most common type of epigenetic data collected, and are generally comprised of measurements from >100,000 genomic locations, called *loci*.
- When the methylation level at a locus is different between two sample groups, this is called *differential methylation*.
- Cell-type methylation levels vary, but are known to be related by the *haematopoeitic lineage*.
- Differential methylation can be cell-type-specific, where only a subset of blood cell-types in the sample are differentially methylated.
- **Objective**: Identify loci with differential methylation for *specific cell-types*

Model

Let y_i be the blood methylation level of sample i. Since y_i is constrained to the unit interval, a logit-Normal distribution was used.

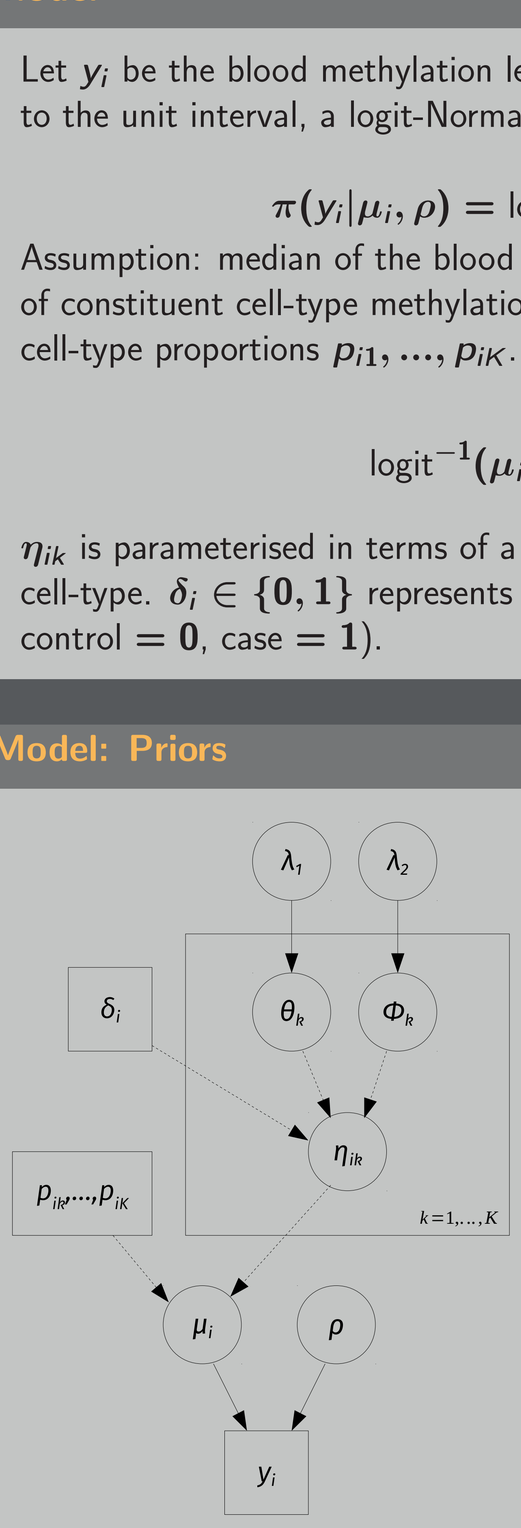
$$\pi(y_i|\mu_i, \rho) = \text{logitNormal}(y_i; \mu_i, \rho),$$

Assumption: median of the blood methylation level is a *linear combination* of constituent cell-type methylation levels $\eta_{i1}, ..., \eta_{iK}$, weighted by the cell-type proportions $p_{i1}, ..., p_{iK}$.

$$\text{logit}^{-1}(\mu_i) = \sum_{k=1}^{K} p_{ik}\eta_{ik}$$

η_{ik} is parameterised in terms of a baseline θ_k and a shift ϕ_k for each cell-type. $\delta_i \in \{0, 1\}$ represents the binary covariate of interest (e.g. control $= 0$, case $= 1$).

Model: Priors



Priors set on lineage-based contrasts:

$$\theta = \mathbf{A}\xi,$$
$$\phi = \mathbf{A}\zeta.$$

For $q \in \{2, ..., K\}$:

$$\pi(\xi_q|\lambda_1) = \text{Normal}\left(\xi_q; 0, \sqrt{\lambda_1}\right),$$
$$\pi(\zeta_q|\lambda_2) = \text{Normal}\left(\zeta_q; 0, \sqrt{\lambda_2}\right).$$

$$\pi(\lambda_1) = \text{Gamma}(\lambda_1; 1, \lambda_0),$$
$$\pi(\lambda_2) = \text{Gamma}(\lambda_2; 1, \lambda_0),$$
$$\pi(\rho) = \text{half-Cauchy}(\rho; 0, 5).$$

ξ_1 and ζ_1 are not cell-type related:

$$\pi(\xi_0) = \text{Cauchy}(\xi_0; 0, 10),$$
$$\pi(\zeta_0) = \text{Cauchy}(\zeta_0; 0, 10).$$

Method: Inference and Case Study

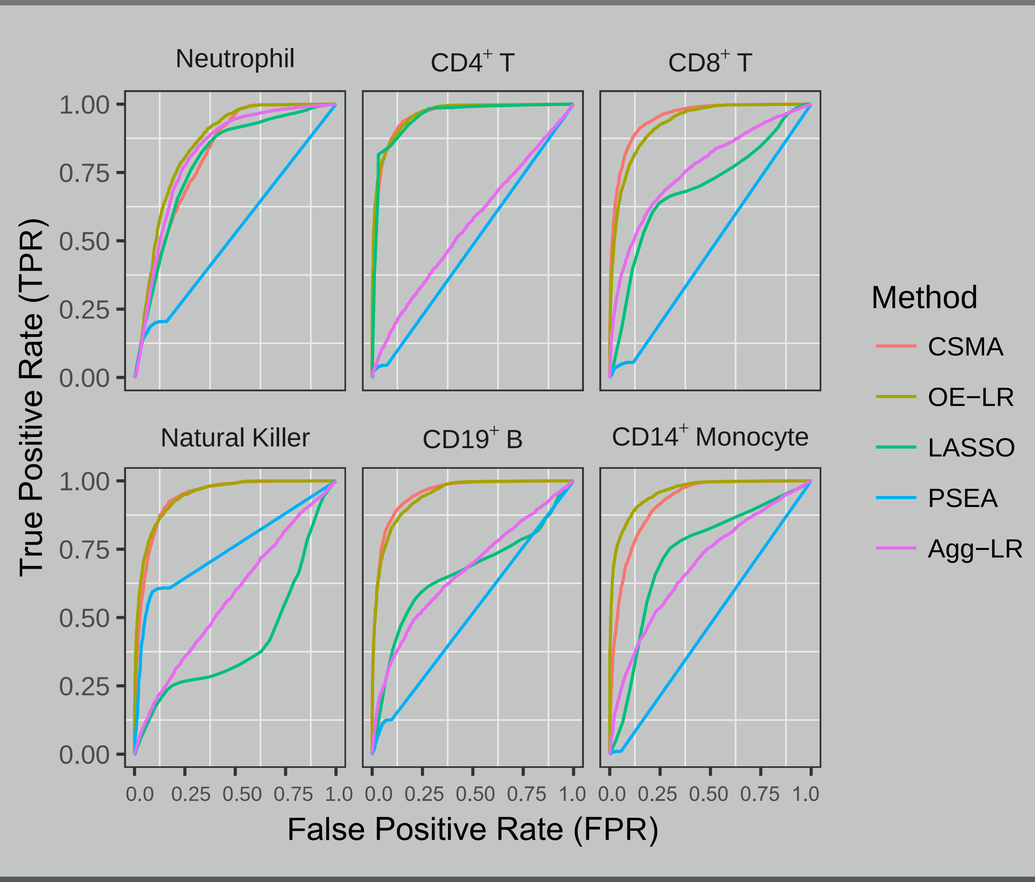
- Model fitted using numerical optimisation procedure in STAN [1].
- Obtained *Maximum A Posteriori* (MAP) estimates for ϕ parameters and a Hessian matrix estimate.
- Laplace approximations to the posterior calculated for each for ϕ_k.
- Predicted differential methylation for cell-type k if

$$\Pr\left(\left|\phi_k - \phi_k^{\text{MAP}}\right| > 0 | \text{Data}\right) < \alpha$$

Case study: find differentially methylated loci associated with sex.

- Data-set contained 5 females and 9 males.
- Cell-sorted data contained ground-truth for comparison with predictions.

Results



Results

- CSMA method outperformed other methods for CD8^{+}T and CD19^{+}B.
- OE-LR method outperformed CSMA for the Monocyte and Neutrophil cell-types.
- CSMA tended to detect more differentially methylated loci specific to the given cell-type.
- OE-LR tended to detect differentially methylated loci where all cell-types were differentially methylated.
- PSEA, Agg-LR, and LASSO methods were sub-optimal.

Conclusions

- CSMA and OE-LR are both useful for finding differentially methylated loci.
- Best method may be to use an *ensemble approach* for finding both cell-type specific and unspecific differentially methylated loci.

Current Work:

- Extending CSMA model to include multiple covariates and different data distributions.
- Reducing potential bias from inaccurate proportion estimates.
- Developing empirical Bayes approach for optimal value of λ_0.

References

[1] Stan Development Team.
RStan: the R interface to Stan, 2016.
R package version 2.14.1.

Acknowledgments

QUT ihbi Institute of Health and Biomedical Innovation

ACEMS AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

@danwkenn

dan.w.kennedy@outlook.com