Cell-type specific analysis of heterogeneous methylation signal using a Bayesian model-based approach

Daniel W Kennedy[1,2], Nicole M White[2], Miles C Benton[2,3], Rodney A Lea[2], and Kerrie Mengersen[1]✉

[1] School of Mathematical Sciences, Queensland University of Technology, Brisbane, 4000, Australia
[2] Institute for Health and Biomedical Innovation, Queensland University of Technology, Kelvin Grove, 4059, Australia
[3] Kenepuru Science Centre, Institute of Environmental Science and Research, Wellington 5240, New Zealand

Motivation: Epigenome-wide studies are often performed using heterogeneous methylation samples, especially when there is no prior information as to which cell-types are disease associated. While much work has been done in ascertaining cell-type fractions and removing cell-type heterogeneity variation, relatively little work has been done in identifying cell-type specific variation in heterogeneous samples.

Results: In this paper, we present a Bayesian model-based approach for making cell-type specific inferences in heterogeneous settings, by using a logit-Normal sampling distribution and incorporating *a priori* knowledge of cell-type lineage. The method is applied to the detection of cell-type specific sex effects in methylation, where cell-type information is present as an independent verification of the results. Panels derived from this method contained more loci where CD8$^+$T, CD19$^+$B and Natural Killer cell-types were differentially methylated. The analysis suggests that an ensemble approach with this method included could be used for discovering cell-type specific methylation changes.

Availability: https://github.com/danwkenn/Bayes_CDM

EWAS | DNA methylation | cell-type heterogeneity

Correspondence: *k.mengersen @qut.edu.au*

Introduction

Epigenome-wide Association Studies (EWASs) have been used extensively to find genomic loci where epigenetic changes are associated with some phenotype of interest. Epigenetic changes are commonly identified by differences in the DNA methylation levels of cells in accessible tissues such as blood. DNA methylation is a biological process whereby a methyl group (CH_3) attaches to a Cytosine base which is proceeded by a Guanine base, referred to as a CpG locus. Methylation is known to vary between cell-types (1), and to play a role in cell differentiation (2) and normal cell function (3). Reinius et al. (1) were able to recover the haematopoietic lineage of whole blood immune cell-types using the methylation profiles of cell-sorted samples, indicating there are distinct cell-type and cell-type grouping methylation profiles.

The methylome has been found to have numerous associations with normal biological variation attributable to age (4) and sex (5, 6), as well as autoimmune disorders including multiple sclerosis (see Webb and de Arellano (7), Zulet et al. (8) for recent reviews), diabetes (9), rheumatoid arthritis (10), and many types of cancer (11, 12). Additionally, there are known to be both phenotype- and disease-associated loci which only show association for specific cell-types. White et al. (13) used a Bayesian model selection approach to identify sets of panels with sex associations associated with cell-types as a function of cell lineage. It has been demonstrated that age has a measurably different effect on methylation for different cell-subtypes and tissue (4, 14), and Multiple Sclerosis is known to be specifically associated with T-cell differential methylation (15).

Many EWASs are conducted using samples from mixed cell-type tissues, such as Whole Blood or PBMC samples, because of (a) the prohibitive cost of obtaining cell-sorted samples, (b) the lack of *a priori* knowledge for which cell-types are associated with the phenotype, and (c) the proliferation of heterogeneous data in the public domain. Given mixed cell samples are comprised of multiple constituent cell-types, mixed cell methylation data exhibit a profile which is a convolution of the profiles from the constituent cell-types. This gives rise to several issues when conducting an EWAS on mixed cell tissue samples. Firstly, differences in methylation between constituent cell-types introduce a large amount of variation which is unrelated to the phenotype of interest (16). Secondly, phenotype-related changes to the cell-type composition are manifested as methylation changes in mixed cell profiles at cell-type associated loci, and could lead to false associations at these loci. Thirdly, phenotype associations with less prevalent cell-types are less likely to be detected compared to that of more common cell-types.

Methylation data are often in the form of beta-values, each of which can be viewed as a measure of the probability that a strand of DNA in the sample is methylated at a given locus. Beta-values from mixed cell samples have been modelled as a linear combination of the underlying cell-type specific methylation levels, weighted by the cell-type proportions of the sample (17–19), an assumption referred to here as the linear mixing process, a term used to describe the equivalent process for gene expression data (20). While many algorithms have been developed which account for and remove variation associated with cell-type heterogeneity (21–23), there has been comparatively little research into finding and estimating effects of a phenotype or disease state on constituent cell-type methylation levels. This paper presents a statistical model for detecting and estimating differential methylation on the cell-type level in mixed cell samples.

One method of predicting associations at the cell-type level uses *a priori* known cell-type-specific regulatory information to suggest which cell-types are likely to be differentially methylated, given a set of phenotypically associated loci (24). However, since the method does not make this determination based on the data itself, it can only predict differential methylation for a given cell-type at loci that are previously known to be associated with that cell-type via regulatory information. Therefore, there is merit in a method which uses only the mixed-cell-derived data, with the greater potential of inferring the direction and size of the phenotype effects in each cell-type.

In this paper we propose a novel statistical model, which combined with an optimisation fitting procedure we refer to as Bayesian Cell-type level Differential Methylation (Bayes-CDM). The model preserves the linear mixing process assumption, but also implicitly enforces the boundary restrictions on the data and the parameters using a logit link function. Furthermore, this model incorporates prior information concerning relatedness of cell-type methylation profiles, based on the haematopoeitic lineage. It is known that methylation plays a role in cell-type differentiation, so the parameters governing cell-type methylation can be made to reflect this lineage via contrasts. None of the previous cell-type inference methods leverage this information, but in our method we establish a prior covariance structure to incorporate it. We therefore propose an extension of EWAS to the cell-type level, when only mixed cell samples are available.

The goal of an EWAS is to identify loci where there is an association between the methylation data and an underlying phenotype. The focus of this paper is to identify an association between the methylation level of a given cell-type's methylation level and an underlying phenotype when only mixed cell data are available. Further, we assume that estimates of the cell-type proportions in each sample are available.

A simulation of mixed cell data is conducted to demonstrate the utility of the method in detecting cell-type level differential methylation. The method is then used for a combined data set of methylation samples from male and female subjects, with sex used as the phenotype to which related differential methylation is identified. Given that corresponding cell-sorted data was available, this provided a ground-truth for comparison with the output of Bayes-CDM and several competing methods.

Methods

Let I be the number of methylation samples. Methylation data for a locus can be represented of an I-length vector $\mathbf{y} = (y_1, ..., y_I)^{\mathrm{T}}$ where y_i corresponds to the beta-value of the i^{th} sample. Beta-values are constrained to the unit interval, that is $0 \le y_i \le 1$ for $i = 1, ..., I$.

In this paper we consider a binary phenotype, whereby the i^{th} is assigned one of two possible phenotype levels, $\delta_i = 0$, or $\delta_i = 1$. The goal is to infer whether there is a difference in the methylation level between the two phenotype levels (0 and 1) for each constituent cell-type.

Given that the data are restricted to the unit interval, y_i is assumed to follow a logit-Normal distribution,

$$\begin{aligned} p(y_i|\mu_i, \rho) &= \mathrm{logitNormal}(y_i; \mu_i, \rho) \\ &= \frac{\rho}{\sqrt{2\pi}x\,(1-x)} \exp\left(-\frac{\rho}{2}\,(y_i - \mu_i)^2\right), \end{aligned}$$

which is defined by the logit-median parameter $\mu_i \in \mathbb{R}$ and the precision parameter $\rho \in \mathbb{R}_+$. The model is applied to each locus individually, meaning that for I samples, the likelihood is

$$p(\mathbf{y}|\mu_1, ..., \mu_I, \rho) = \prod_{i=1}^{I} \mathrm{logitNormal}(y_i; \mu_i, \rho).$$

Several previous methods (18, 21, 25) used a Normal distribution for the beta-value. The mean value was specified as the linear combination of the underlying cell-type methylation values, weighted by their respective cell-type proportions. Since the mean and mode of the logit-Normal distribution are not available in closed form, the median of y_i was used.

Let π_{ik} be the cell-type proportion estimate for sample i and cell-type k, where $i = 1, ..., I$ and $k = 1, ..., K$:

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

The median of y_i is therefore assumed to be the linear combination of the underlying cell-type methylation levels $\eta_{i1}, ..., \eta_{iK}$, where each η_{ik} is constrained between 0 and 1. The cell-type proportion values are constrained to be positive, and values from each sample add to 1.

The linear predictor for each η_{ik} is parametrised in terms of a baseline θ_k and a shift ϕ_k for each cell-type.

$$\mathrm{logit}(\eta_{ik}) = \theta_k + \delta_i\phi_k$$

Therefore, η_{ik} can be thought of as the cell-type k methylation level in sample i. This parametrisation allows for the effect of phenotype on cell-type k methylation level to be estimated from ϕ_k. When $\phi_k = 0$, there is no difference between the methylation levels of cell-type k for the two phenotype levels. By having $\theta_k + \delta_i\phi_k$ equal the logit of the cell-type methylation level, θ_k and ϕ_k do not need to be constrained in order for η_k and thus μ_i to remain on the unit interval.

We adopted a Bayesian modelling approach to overcome the issue of constrained data and parameters, and to incorporate cell-type lineage relationships as prior information. The model has $2K$ free parameters specifying the mean methylation level, but typical methylation data sets tend to have small sample sizes. Therefore, regularisation in the form of an informative prior distribution was used for the θ and ϕ parameters. A set of Normal shrinkage priors was placed on special linear combinations which incorporate the haematopoeitic lineage (see Figure 1). For the remainder of the paper we fo-

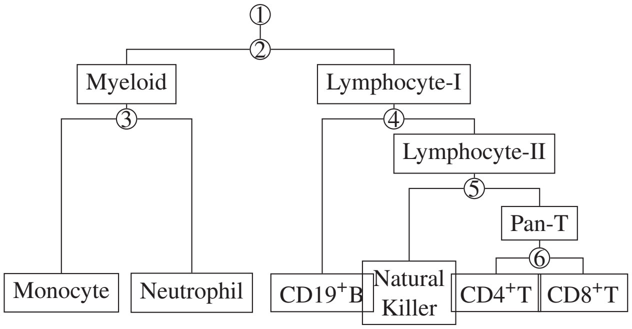


Fig. 1. Cell-types and cell-type groupings as clustered by the haematopoietic lineage, with *nodes* each given a number. The nodes relate to the columns of the lineage matrix $\mathbf{A}$ while each of the cell-types relate to a row.

Parameter	Mapping
θ_{Monocyte}	$\xi_1 - {}^1/_2\xi_2 - {}^1/_2\xi_3$
$\theta_{\text{Neutrophil}}$	$\xi_1 - {}^1/_2\xi_2 + {}^1/_2\xi_3$
$\theta_{\text{CD19}^+\text{B}}$	$\xi_1 + {}^1/_2\xi_2 - {}^1/_2\xi_4$
$\theta_{\text{Natural Killer}}$	$\xi_1 + {}^1/_2\xi_2 + {}^1/_2\xi_4 - {}^1/_2\xi_5$
$\theta_{\text{CD4}^+\text{T}}$	$\xi_1 + {}^1/_2\xi_2 + {}^1/_2\xi_4 + {}^1/_2\xi_5 - {}^1/_2\xi_6$
$\theta_{\text{CD8}^+\text{T}}$	$\xi_1 + {}^1/_2\xi_2 + {}^1/_2\xi_4 + {}^1/_2\xi_5 + {}^1/_2\xi_6$
ϕ_{Monocyte}	$\xi_1 - {}^1/_2\zeta_2 - {}^1/_2\zeta_3$
$\phi_{\text{Neutrophil}}$	$\zeta_1 - {}^1/_2\zeta_2 + {}^1/_2\zeta_3$
$\phi_{\text{CD19}^+\text{B}}$	$\zeta_1 + {}^1/_2\zeta_2 - {}^1/_2\zeta_4$
$\phi_{\text{Natural Killer}}$	$\zeta_1 + {}^1/_2\zeta_2 + {}^1/_2\zeta_4 - {}^1/_2\zeta_5$
$\phi_{\text{CD4}^+\text{T}}$	$\zeta_1 + {}^1/_2\zeta_2 + {}^1/_2\zeta_4 + {}^1/_2\zeta_5 - {}^1/_2\zeta_6$
$\phi_{\text{CD8}^+\text{T}}$	$\zeta_1 + {}^1/_2\zeta_2 + {}^1/_2\zeta_4 + {}^1/_2\zeta_5 + {}^1/_2\zeta_6$

Table 1. Mapping of contrast parameters ξ and ζ terms to the cell-type level parameters θ and ϕ.

cus on the example of whole blood methylation data, which is composed of 6 major cell-types; CD14⁺ Monocyte, CD56⁺ Natural Killer (henseforth abbreviated to Monocyte and Natural Killer respectively), Neutrophil, CD19⁺B, CD4⁺ and CD8⁺T, however this method can be extended to any heterogeneous tissue.

By associating a parameter with the node of the lineage rather than the cell-type, we can represent cell-type methylation as successive contrasts between the two cell-type groupings distinguished after each node. The phenotype level-0 cell-type methylation levels are parametrised by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^{\mathrm{T}}$. Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^{\mathrm{T}}$ be a set of contrast parameters associated with each node, such that $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ are related via a lineage matrix $\mathbf{A}$:

$$\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\xi}$$

$$\mathbf{A} = \begin{bmatrix} 1 & -{}^1/_2 & -{}^1/_2 & 0 & 0 & 0 \\ 1 & -{}^1/_2 & +{}^1/_2 & 0 & 0 & 0 \\ 1 & +{}^1/_2 & 0 & -{}^1/_2 & 0 & 0 \\ 1 & +{}^1/_2 & 0 & +{}^1/_2 & -{}^1/_2 & 0 \\ 1 & +{}^1/_2 & 0 & +{}^1/_2 & +{}^1/_2 & -{}^1/_2 \\ 1 & +{}^1/_2 & 0 & +{}^1/_2 & +{}^1/_2 & +{}^1/_2 \end{bmatrix}$$

Each column relates to a node in the cell-type lineage, and each row refers to a cell-type. The corresponding mapping of contrast parameters to cell-type level parameters is given in Table 1. The parameters ξ_q differentiate between the cell-type groupings formed by the associated node q. For example, ξ_2 acts as the contrast between Myeloid and Lymphocyte-I types, as each Myeloid cell-type methylation has a $-\frac{1}{2}\xi_2$ term, and each Lymphocyte-I cell-type baseline methylation has a $+\frac{1}{2}\xi_2$ term. The subsequent ξ terms describe later cell-type differentiations in the lineage such as ξ_6, which is the difference in baseline methylation between CD8⁺T and CD4⁺T. The exception is the first parameter ξ_1 acts as an intercept, since it appears as a component in all constituent cell-types, and does not differentiate between cell-type groupings.

The influence of cell-type lineage on the phenotype effect is also considered. In the same way as above, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)^{\mathrm{T}}$ can be described in terms of node contrasts $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_K)^{\mathrm{T}}$.

$$\boldsymbol{\phi} = \mathbf{A}\boldsymbol{\zeta}.$$

Here ζ_1 acts as an overall shift in methylation level from phenotype level 0 to 1. The second parameter ζ_2 is the difference in phenotype effect between Myeloid and Lymphocyte-I types, and so on.

We placed Normal priors with mean 0 on the linear contrast parameters, which have the effect of shrinking the posterior distributions and parameter estimates towards 0, and shrinking associated cell-type methylation level parameters (θ and ϕ) of related cell-types together. The degree of shrinkage is dependent on the precision of the prior distributions relative to the informativeness of the data as expressed by the likelihood. Therefore, as sample size and thus information from data increases, bias incurred from shrinkage decreases. Thus for $q \in \{2, \ldots, K\}$,

$$p(\xi_q|\lambda_1) = \text{Normal}(\xi_q; 0, \lambda_1),$$
$$p(\zeta_q|\lambda_2) = \text{Normal}(\zeta_q; 0, \lambda_2).$$

The prior distributions for the two precision parameters λ_1 and λ_2 were Gamma distributed with a shape parameter $\alpha = 1$ and the shrinkage parameter λ_0,

$$p(\lambda_1) = \text{Gamma}(\lambda_1; 1, \lambda_0),$$
$$p(\lambda_2) = \text{Gamma}(\lambda_2; 1, \lambda_0).$$

The mean of $p(\lambda_1)$ and $p(\lambda_2)$ is $1/\lambda_0$, so the effect of increasing λ_0 is to decrease the prior mean of λ_1 and λ_2 and thereby increase the amount of regularisation. The value $\lambda_0 = 1$ was chosen as a reasonable level of regularisation, given the high noise-low data context of methylation studies. It is straightforward to show that the marginal prior distribution of the ξ and ζ parameters is a standard Student's t distribution with 2 degrees of freedom. For a simple empirical justification and investigation of the prior distributions of θ and ϕ parameters as a consequence of the above prior specification of ξ and ζ refer the Supplementary Material.

A half-Cauchy prior was used for the precision parameter ρ:

$$p(\rho) = \text{halfCauchy}(\rho; 0, 5) = \frac{2}{5\,(1 + x^2/25)}$$

Since ξ_1 and ζ_1 are not cell-type related, we chose not to shrink them and instead use a weakly informative Cauchy distribution:

$$p(\xi_0) = \text{Cauchy}(\xi_0; 0, 10) = \frac{1}{10\,(1 + x^2/100)},$$
$$p(\zeta_0) = \text{Cauchy}(\zeta_0; 0, 10) = \frac{1}{10\,(1 + x^2/100)}.$$

A. Model Inference. The posterior densities for the model parameters were approximated using Laplace approximations. Maximum A Posteriori (MAP) estimates for the parameters $\left(\boldsymbol{\xi}^{\text{MAP}}, \boldsymbol{\zeta}^{\text{MAP}}, \lambda_1^{\text{MAP}}, \lambda_2^{\text{MAP}}, \rho^{\text{MAP}}\right)$ and an estimate of the Hessian matrix were calculated using numerical optimisation implemented in the STAN (26) software R package. To calculate posterior standard deviations of ϕ and θ, the standard deviation of the marginal distributions needed to be calculated from the corresponding Hessian estimate.

Since the parameter ϕ_k determined the effect size of phenotype on the methylation level of cell-type k, it is the basis for predicting if a given cell-type k is differentially methylated. Since we used a Laplacian approximation to the posterior, the probability

$$2\text{Pr}\left(Z > \left|\phi_k^{\text{MAP}}\right| / \hat{\text{SD}}(\phi_k)\right)$$

is a measure of the magnitude of the standardised effect size, where $Z \sim \text{Normal}(0, 1)$ and $\hat{\text{SD}}(\phi_k)$ is the estimate of the standard deviation of the posterior for ϕ_k. Consequently, we used this value with cut-off value α as a decision boundary. If this posterior probability for ϕ_k was less than α, then we predict that cell-type k is differentially methylated.

B. Simulation study. We conducted a simulation study of mixed cell data to investigate the ability of the method to detect differential methylation in cell-types, in comparison to a number of other potential methods. An underlying set of cell-type methylation levels was constructed for four different scenarios (see Suppl. Figure 1) to investigate behaviour of detection methods over different profiles of differential methylation. We presumed a reasonable number of samples for a methylation study is $I = 50$, and precision parameter $\rho = 30$.

The four simulation scenarios were designed to investigate how varying the number of differentially methylated cell-types and the effect sizes affected the probability of detection (see Suppl. Figure 1 for visual depiction). Scenario 1 simulates a locus where the differential methylation is restricted to cell-types of the myeloid lineage grouping (Monocyte and Neutrophil), playing to the strength of Bayes-CDM, which uses the prior covariance structure to shrink related cell-type methylation levels together. Scenario 2 showed a similar situation, except the effect is much smaller. Scenario 3 showed a situation where differential methylation is exhibited in three cell-types, not all related. Finally, Scenario 4 showed a situation where only a single cell-type, the Monocyte, is differentially methylated. The method and a set of alternatives, described in Table 2, were used to predict if each cell-type was differentially methylated in all four scenarios. This was repeated for 100 simulations of each scenario to investigate the methods in terms of specificity and sensitivity.

In all scenarios, there was a non-phenotypic difference in the methylation levels between the Myeloid and Lymphocyte, as well as a difference between CD19$^+$B and the other Lymphocyte types. In terms of the model, this translates to differing θ-values between cell-types. The Fluorescence-Activated Cell-Sorting (FACS) composition data from a mixed-sex data set (GSE88824) was used to fit Dirichlet parameters (via diri.est from the Compositional R package), and simulated composition data were drawn from this Dirichlet distribution.

Several alternative methods were used as a comparison for Bayes-CDM and are described in table 2. The Orthogonal Effects Linear Regression (OE-LR) method presented here models the phenotype effect as non-specific to cell-type, and as such should have a performance similar to methodologies which correct for cell-type heterogeneity without allowing for cell-types to have different phenotype effects. Some examples of this type of methodologies include (CITE,CITE, CITE), which all use the proportion estimation method by (27) to obtain proportion estimates. The Aggregated Linear Regression (Agg-LR) method represents a suggested extension of the method used by (18) for more than two cell-types. Here a separate model is fitted for each cell-type by aggregating other cell-type proportions together. Population-Specific Expression Analysis (PSEA) (25) and an implementation of the LASSO (28) represent the same, fully specified linear model, with a separate effect for each cell-type. They differ in that PSEA uses best-subset selection while LASSO uses l_1-regularisation to select the optimal parameter set. In the case of the LASSO, the variable selection was limited to the interaction effects between cell-type fraction and phenotype, and 10-fold cross-validation was used to find the optimal tuning parameter for the LASSO. To the best of our knowledge, the LASSO has not been applied for this specific problem in methylation data, however given the interest in regularisation techniques we wanted to investigate how such a method would perform.

Since there are multiple cell-types; and therefore multiple possible cell-types with differential methylation, an ideal method correctly detects all cell-types with differential methylation. Consequently, the detection rate of differential methylation in each cell-type was calculated, as well as the rate of choosing five multi-cell-type measures described in Table 3. These measures were designed to quantify the simultaneous predictive performance over multiple cell-types.

C. Case study: Finding Cell-type differential methylation associated with sex. Two publicly available data sets (GEO Accession Numbers: GSE35069 and GSE88824) containing both mixed cell and cell-sorted data, as well as Fluorescence Activated Cell Sorting (FACS) estimates of the cell-type proportions (see Supplementary Material of Reinius et al. (1)), were merged into a single combined data set. In total there were 5 female and 9 male subjects, which allowed us to identify differential methylation associated with sex using

Method	Abbreviation	Description
Orthogonal Effects Linear Regression (standard heterogeneity-corrected linear regression)	OE-LR	Linear regression with cell-type proportion estimates included to account for cell-type heterogeneity, and a single orthogonal effect of phenotype. Model: $y_i = \delta_i \beta + \sum_{k=1}^{K} \pi_{ik} \gamma_k + \epsilon_i$
Aggregated Linear Regression (extension of (18))	Agg-LR	Method is run once for each cell-type. For cell-type k, the proportions of the other cell-types are aggregated together, and the following model is fitted: $y_i = \beta_0 + \beta_1 \pi_{ik} + \beta_2 \delta_i \pi_{ik} + \beta_3 \delta_i (1 - \pi_{ik}) + \epsilon_i$ The value of β_2 determines if cell-type differential methylation occurs.
Population-Specific Expression Analysis (25)	PSEA	Method uses a complete model with a phenotype effect for each cell-type: $y_i = \beta_0 + \sum_{k=1}^{K} \pi_{ik} \beta_k + \sum_{k=1}^{K} \pi_{ik} \delta_i \gamma_k + \epsilon_i$ The method then uses best subset selection based on the Akaike Information Criterion to select a subset of the coefficients. We set the cut-off value as 1 if the effect was not selected, and the standard F-test p-value from the OLS fit of the model if selected.
Least Absolute Shrinkage and Selection Operator (28)	LASSO	Method is a complete model as with PSEA, but applies a regularising penalty function on the coefficients which acts to select variables by shrinking some to 0. We only applied the shrinkage on the phenotype effect coefficients, and obtained a cut-off by choosing the value of the shrinkage parameter where the coefficient was shrunk to 0.

Table 2. Alternative methods for finding differential methylation in cell-type heterogeneous samples. These methods represent four possible alternative approaches to making inferences; Orthogonal Effects Linear Regression (OE-LR); where only considering orthogonal effects to the cell-type proportion estimates; Aggregated Linear regression (Agg-LR), where each cell-type's differential methylation is inferred separately; Popoulation-Specific Expression Analysis (PSEA), which uses the full linear model and a subset selection method; and LASSO, which also uses the full model but uses the sparsity property of the regularising function in contrast with the PSEA method.

Measure	Description
Predicted DM	Differential methylation detected in at-least one cell-type.
At least one true	Differential methylation correctly predicted in at least one cell-type.
At least one true, no false	Differential methylation correctly detected in at least one cell-type, without any incorrect detections.
All true associations	Differential methylation correctly detected for all cell-types with differential methylation.
Correct Subset	Differential methylation correctly detected for all cell-types with differential methylation, with no incorrect detections.

Table 3. Measures for the evaluating predictive performance over multiple cell-types.

both the mixed cell data and independently using the cell-sorted data.

C.1. Obtaining a ground-truth differential methylation using cell-sorted data. In this case study, each of the methods described in the previous section were used to predict which loci were differentially methylated for each cell-type from mixed-cell data. The cell-sorted data available provided an independent and accurate means of detecting differential methylation, and in consequence the differential methylation status detected via cell-sorted data were used as the ground-truth against which the predictions from mixed-cell data were compared.

For each cell-type and locus, the beta-values were grouped into two classes based on sex, and an F test of difference of means was performed. The resulting p-values were adjusted using the Benjamini-Hochberg procedure (29) to control the false discovery rate. Using the cut-off α, a cell-type was labelled as differentially methylated at a given site if the associated adjusted p-value was less than α. In this case study we used the value $\alpha = 1 \times 10^{-4}$. While this implies a degree of false positives and false negatives in this set, we considered this to be a reasonable substitute for a perfect ground-truth, especially since the false positive rate should be low given the value of α, and that it is unlikely that whole blood methods would detect cell-type differential methylation without it also being detected in cell-sorted data.

C.2. Validation. When the actual locations of Cell-type differential methylation for each cell-type are known, it is pos-

Kennedy *et al.* | Cell-type specific analysis of heterogeneous methylation bioRχiv | 5

sible to perform a Receiver-Operator Curve (ROC) analysis for each cell-type, as well as for general differential methylation prediction. The relationship between panel size and False Discovery Rate (FDR) was explored and compared between methods, and panel sizes were chosen so the FDR was $10\% \pm 1\%$.

Results

D. Simulation Study. The marginal densities and tabulated summary statistics for the proportion estimates on which the simulated proportion values are based are given in Suppl. Table 1.

As shown in Figure 2 The Orthogonal Effect-Linear Regression (OE-LR) predicted Differential Methylation for all simulation runs of Scenarios 1 and 3, but only 76% and 44% of runs in Scenarios 2 and 4 respectively. OE-LR exhibited poor performance in Scenario 4, indicating that when differential methylation only occurs in a single uncommon cell-type, this is difficult to detect as an orthogonal effect, since the apparent orthogonal effect of a truly cell-type specific effect is proportional to the average cell-type fraction.

The LASSO method predicted differential methylation at a higher frequency than the other methods in all four scenarios, except for Bayes-CDM at a cut-off value of 0.2. However, the LASSO displayed a tendency to incorrectly predict differential methylation for cell-types where no differential methylation was present. This resulted in a low rate of correct subset prediction. The LASSO performed comparatively well for predicting differential methylation in Scenario 4, indicating that it may be suitable where only a single cell-type is differentially methylated.

The Aggregating Linear Regression (Agg-LR) performed the most poorly in terms of predicting differential methylation and correct subset prediction (excluding OE-LR which is not designed to predict the correct subset). While Agg-LR rarely incorrectly predicted differential methylation, it lacked the sensitivity to predict differential methylation in cell-types where it was present. As a result, Agg-LR predicted the correct subset in 18% of simulation runs in Scenario 3 and $\leq 8\%$ for the other scenarios. The Agg-LR method was computationally efficient, so it may be useful as a first-pass analysis when the signal is particularly large and from multiple cell-types as in Scenario 1.

The PSEA method tended only to predict differential methylation in a single cell-type if any, which reduced its performance in the scenarios where more than one cell-type was present (1,2, and 4). The method had a very low rate of predicting cell-type level differential methylation incorrectly.

The best method for correct subset prediction was Bayes-CDM, however the optimal cut-off value differed between scenarios. In Scenario 1, a cut-off value of 0.05 was optimal, but in Scenario 4 the cut-off value 0.1 was optimal, and 0.15 was optimal for Scenario 2 and 3. While these values did not correlate with the optimal values for prediction of differential methylation in each scenario, they tracked well with the correct subset measure. This indicates that the prediction of differential methylation increases with cut-off both because of an increase in prediction of truly differentially methylated cell-types, and a higher rate of incorrectly predicting differential methylation. As a result, the rate of correct subset prediction increases as a result of correct prediction, but decreases as the rate of incorrect prediction increases.

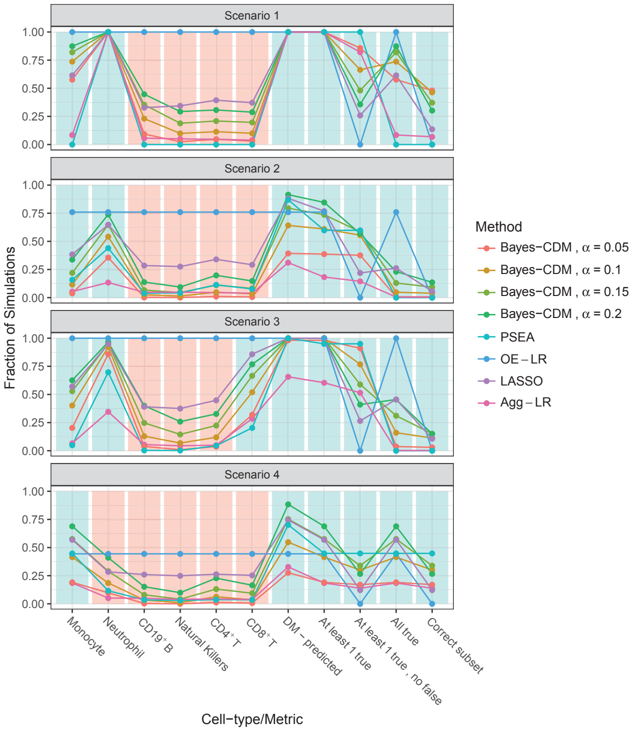


Fig. 2. Results for four scenarios with 100 simulation runs each. Bayes-CDM, PSEA, Agg-LR and LASSO methods were used to detect cell-type differential methylation in each of the 6 cell-types, as well as the four metrics. For Bayes-CDM, several different cut-off values are given. Numbers indicate the proportion of simulation runs where the Cell-type differential methylation was detected or multi-cell-type metric was met. The rows are coloured blue or red, depending on whether a large or small value should be seen, given which cell-types are differentially methylated in each scenario.

For a single simulation run on each scenario, the model in Bayes-CDM was fitted using Hamiltonian Monte Carlo (HMC), and the posterior densities compared with the Laplace approximations. While the true densities displayed some skewness where the Laplace approximations did not, the mean values were close and the overall shapes overlapped well. Therefore we concluded that the Laplace approximations were a sufficiently accurate representation of the true posterior densities (see Suppl. Figure 2).

E. Case Study: Sex Associations. Based on the cell-sorted data, it was determined that 5813 (1.3%) of CpG loci exhibited differential methylation in at least one cell-type. Of these loci, 2553 (43.9%) exhibited differential methylation in all six cell-types, and were found to be almost all located on the X and Y chromosomes (2538 and 14 respectively; one exception on Chr 3). Of the DMLs with less than 6 differentially methylated cell-types, most of these were also located on the X and Y, although 27 (0.8%) were located on other chromosomes.

A ROC analysis (Figure 3) showed that all methods except for the Agg-LR showed very high discrimination (AUC >

99%) between CpGs with and without differential methylation.

The FDR (Figure 3 B) for Agg-LR and LASSO methods grew rapidly compared to the other methods as CpGs were added to the panel, while the FDR for Bayes-CDM, OE-LR and PSEA grew very slowly, indicating these methods were better able to discriminate between differentially methylated sites and non-differentially methylated loci.

Investigating these panels showed that almost all loci which exhibited differential methylation in four or more cell-types were contained within the panels, with the exception of the Agg-LR panel. All panels tended to include fewer loci where less than four cell-types were differential methylated. For all methods, there was indication of a relationship between the number of cell-types with differential methylation at a locus, and the chance of predicting differential methylation at this locus.

F. Cell-type-specific Differential Methylation Prediction. The second task was to predict differential methylation at loci for specific cell-types. The ROC analysis showed that the Bayes-CDM performed quite consistently over all cell-types. It was the top-performing method for the CD8$^+$T and CD19$^+$B cell-types, while the OE-LR method performed the best for the Monocyte and Neutrophil cell-types, which tended to be differentially methylated together, resulting in a large effect. For the Bayes-CDM and OE-LR methods, the Neutrophil cell-type appeared to be the most difficult to detect based on the AUC values, despite being the most common cell-type. The PSEA method showed very poor performance for determining cell-type differential methylation for any of the cell-types. The LASSO was generally a poor to moderate performer for the all cell-types except CD4$^+$T. The Agg-LR method showed poor comparative performance except for Neutrophil, where it outperformed the Bayes-CDM method. For predicting differential methylation at specific cell-types, the Bayes-CDM and OE-LR performances were almost equal best, although the AUC indicated OE-LR was a slightly better performer.

For panels of the same size the Bayes-CDM and OE-LR methods generally had lower FDR values, and the relationship between FDR and panel size was monotonic, with the exception of Neutrophils where the minimum was at a panel size around 2000. The FDR curve for the PSEA method was not monotonic or smooth, and did not drop below 10% except for Natural Killers.

Panels were selected for a $10 \pm 1\%$ FDR where possible. This produced panels of varying size, including several cases where the FDR was never close enough. Notably no method could produce a panel of any size for Neutrophil with a FDR near 10%. While panels from both Bayes-CDM and OE-LR contained most loci with 5 or more other cell-types also exhibiting differential methylation, Bayes-CDM tended to detect more of the CpG loci with smaller subsets of cell-types exhibiting DM for CD8T, NK, and CD19B, while OE-LR was able to detect more of these loci in Monocytes.

Discussion

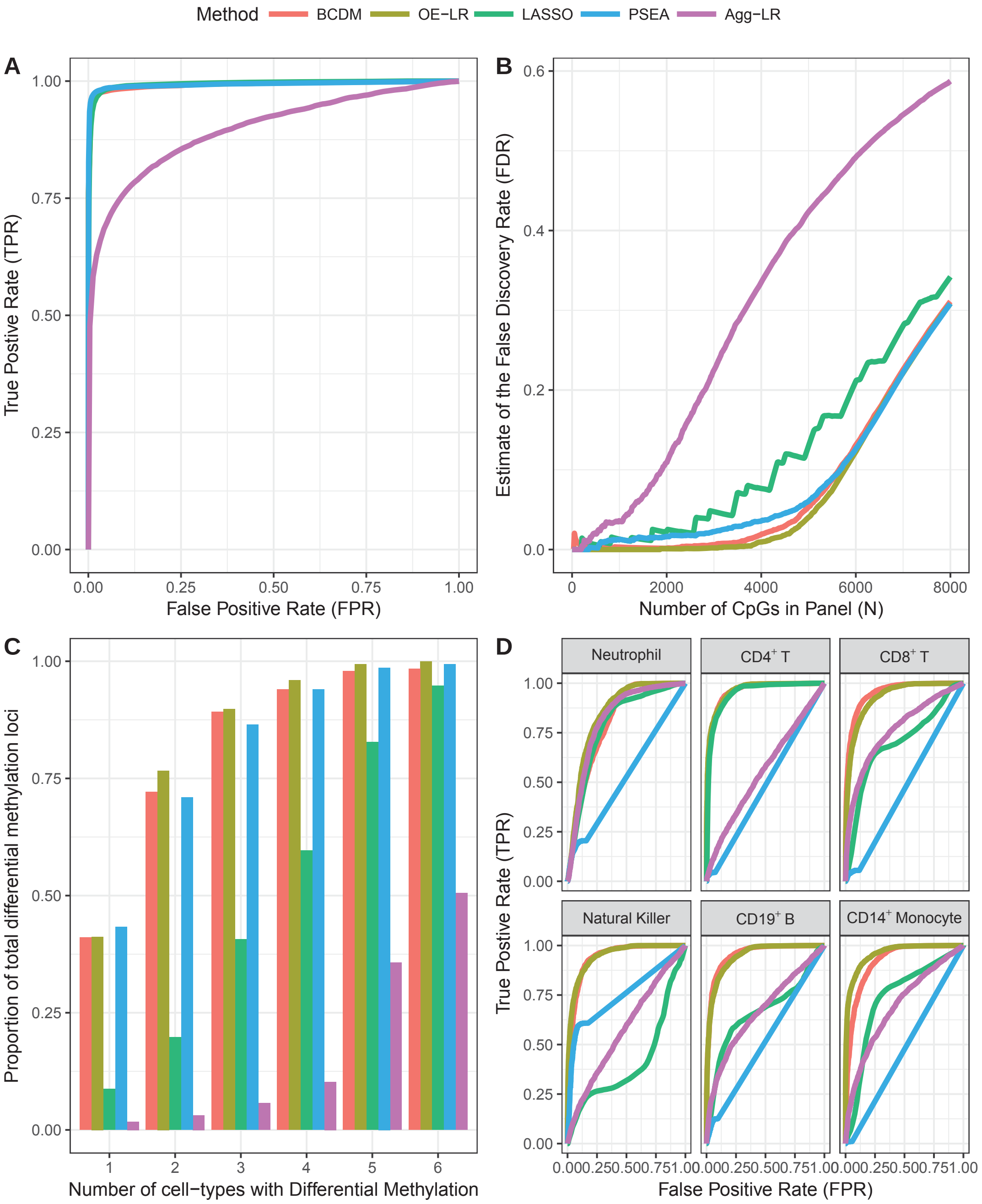
In this paper, we present a new method for predicting differential methylation in specific cell-types when only heterogeneous or mixed cell data are available. The method is based on a Bayesian model which takes account of the inherent constraints on both methylation data and cell-type methylation levels through logit link functions, whilst preserving the linear mixing process assumption. We also incorporate prior knowledge of cell-type relatedness by specifying independent Normal priors on a set of contrast parameters, which act to shrink related cell-type methylation levels together. The method performed relatively well at detecting differentially methylated loci in comparison to other methods, and demonstrated consistent performance in identifying differential methylation associated with phenotype for all cell-types.

Almost all methods for predicting differential methylation at the cell-type level require *a priori* cell-type proportion information. The PSEA method (25) uses the mean values of *a priori* known cell-type specific expression as a surrogate for proportion estimates, while the method by Montaño et al. (18) uses reference-based proportion estimate (27) as covariate inputs. An exception is the recent unsupervised method MeDeCom by Lutsik et al. (19), which applies non-negative matrix factorisation in addition to a boundary-weighted regularisation penalty to estimate so-called latent methylation components bounded between 0 and 1, and proportion estimates for each of these components. While this method presents a useful procedure for estimating cell-type methylation level, it does not include phenotype covariate information, and so it is not make clear how to use it for identifying phenotype related differential methylation.

A significant hurdle in modelling the heterogeneous data is that both the beta-value data and the cell-type methylation levels are restricted to the unit interval. A general remedy to the bounded methylation data is to apply a logit-transform, the result being referred to as an M-value. The M-value is unrestricted, thus allowing standard analysis tools with Normal assumptions to be used more effectively (30). However, a critical issue in the case of mixed cell data is that if methylation is characterized in terms of M-values, then the linear mixing process assumption used in previous methods needs to be redefined.

Methods which consider cell-type heterogeneity can model phenotype effects either orthogonal to proportion estimates such as RefFreeEWAS (21), or as interactions between phenotype and proportion, such as PSEA Kuhn et al. (25) and the method by Montaño et al. (18). Kuhn et al. (25) and Montaño et al. (18) give similar mathematical arguments for this conclusion.

Previous methods have used a Normal distribution for beta-values without any constraints (18) or used constrained optimisation (19). (30) showed empirically that standard statistical tests for differential methylation with normal distribution assumptions were more powerful when the input data was the M-value version rather than beta-value. This method uses a logit-Normal model, which is equivalent to using a normal distribution with M-value data, but the linear mixing pro-



Method	DM	CD19⁺B	CD4⁺T	CD8⁺T	CD14⁺Mono	Neutrophil	Natural Killer
Bayes-CDM	99.2	95.4	95.8	95.2	92.1	80.8	94.0
OE-LR	99.4	94.6	96.2	93.1	95.9	84.9	94.6
LASSO	99.4	66.9	95.3	69.8	73.5	79.3	42.2
PSEA	99.3	51.7	48.7	47.1	47.7	53.1	75.10
Agg-LR	89.6	67.0	55.8	76.6	69.4	82.2	57.5

Fig. 3. (A) ROC curves for finding Differentially Methylated Loci, showing True Positive Rate (TPR) on the y-axis and False Positive Rate on the x-axis. Area Under the Curve (AUC) metric given for each method in the legend. (B) False Discovery Rate for panels of predicted DMLs, as it changes with increasing panel size for all 5 methods, calculated using the ground-truth from the cell-sorted data. (C) Frequencies of differentially methylated CpGs found in 10% FDR panels from the five different methods, stratified by the number of cell-types differentially methylated. (D) ROC analysis of the five different methods' ability to detect differential methylation for each of the six cell-types. AUC (%) values are given in the table below the graph.

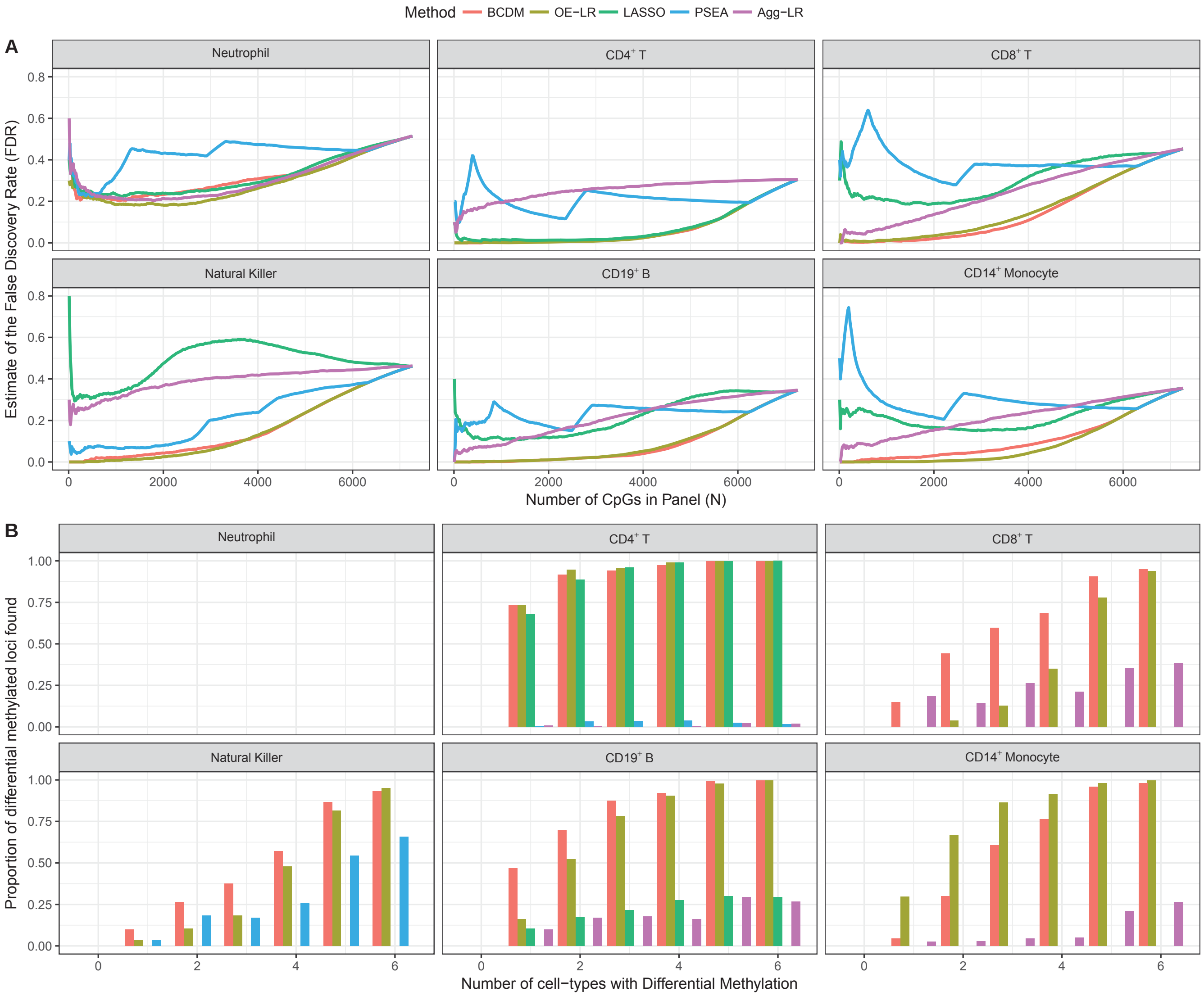


Fig. 4. (A) False Discovery Rate (FDR) curve for each method as a function of panel size. Each method creates a different panel for each cell-type. (B) Distribution of CpGs in 10% FDR panels as a function of the number of cell-types with differential methylation. Values here are a proportion of the total number of CpGs with differential methylation for the given cell-type.

cess assumption is still valid. Additionally, with the exception of White et al. (13), this is the first method to incorporate prior information from the cell-type lineage to improve inference. Whereas White et al. (13) used the cell-type lineage in a model selection context to identify a small set of candidate models based on lineage groupings, here we use a single model and incorporate the lineage as informative prior distributions.

The shrinkage parameter λ_0 controls the informativeness of the prior, and is analogous to the tuning parameter in ridge regression or other penalized regression methods. Because a hierarchical approach was used where λ_0 is conditionally separated from the parameters of interest by λ_1 and λ_2, this should mean the inferences are somewhat sensitive to choice of λ_0. Nevertheless, the value could potentially be optimised by cross-validation or via information criteria such as AIC or BIC. However, these require multiple runs of the model for different tuning values, which has a high computational burden. An alternative could be the Empirical Bayesian Gibbs Sampler proposed by Casella (31), which uses an EM algorithm step before the Monte Carlo sampling algorithm to optimise the hyperparameter. A related algorithm proposed by Atchadé (32) uses stochastic approximation to optimise the hyperparameter and draw MC samples in the same run. These algorithms have subsequently been employed in several Bayesian regularised regression methods, including the Bayesian LASSO (33), the Bayesian elastic net (34), and the Bayesian adaptive LASSO (35). These MCMC-based procedures are much slower than the optimisation method used, and so present a significant computational challenge given the large number of loci. Despite this, the problem remains embarrassingly parallel.

From the results of our case study, it is evident that the ability of any method to predict differential methylation at a locus depends on the number of differentially methylated cell-types and the cell-type of interest. Among the compared methods

Bayes-CDM was close to the best for finding differentially methylated loci, but investigation of the 10% FDR panels showed that Bayes-CDM was more likely to correctly predict differential methylation at loci with a single differentially methylated cell-type for some cell-types. It is therefore recommended that orthogonal methods such as OE-LR or heterogeneity correction methods (e.g. RefFreeEWAS (21) or ReFACTor (23)) are used alongside Bayes-CDM in finding differentially methylated loci. Differentially methylated loci found using the former could subsequently be investigated with Bayes-CDM to predict cell-type differential methylation.

The simulation scenarios showed that Bayes-CDM was broadly better at accurately detecting cell-type differential methylation where differentially methylated cell-types were contained in a lineage grouping. This indicates that the Bayes-CDM model is more suited to loci where the underlying profile of cell-type methylation levels fits well within lineage groupings. The cell-sorted data indicates that a large portion of the methylome is significantly associated with cell-type groupings (13), in comparison with the number associated with specific cell-types. Performance of Bayes-CDM at detecting differential methylation restricted to a single cell-type is limited for small sample size, but as the amount of information from data increases, the bias in the posterior toward cell-type groupings would decrease, resulting in more precision for specific cell-types.

This model-based approach has a great deal of potential for similar applications where the assumption of a linear mixing process is appropriate. For instance, gene expression array data from mixed cell samples is constrained to the positive real line and can be modelled as a linear combination of underlying cell-type expression profiles (25). A similar model to Bayes-CDM using a log-transform rather than a logit-transform could be applied to detect cell-type specific expression associated with a phenotype. Count-based data from a mixed cell sample could be modelled in a similar way with a binomial, beta-binomial, Poisson, or Poisson-Gamma data distribution.

In the context of tissue for which there is no known level of hierarchy, one could still apply a logit-function to enforce the constraint on the cell-type methylation levels and the whole blood methylation. Instead of applying informative priors on contrast parameters, the priors could be applied to the cell-type levels, with a common mean parameter, or another representations based on known external information.

The future work is to expand the model to detect cell-type level differential methylation in multiple neighbouring loci rather than just on a locus-by-locus basis. This will allow investigation of differential methylation regions as well as single loci.

Acknowledgements

The authors wish to acknowledge Professor David Balding and Professor Terry Speed for their advice regarding modelling constrained data.

Funding

This research is supported by an Australian Government Research Training Program (RTP) Scholarship, and the Australian Research Council Australian Centre for Mathematical and Statistical Frontiers (ACEMS).

Bibliography

1. Lovisa E. Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, and Juha Kere. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. *PLoS ONE*, 7(7):e41361, jul 2012. doi: 10.1371/journal.pone.0041361.
2. A. M. Deaton, S. Webb, A. R. W. Kerr, R. S. Illingworth, J. Guy, R. Andrews, and A. Bird. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Research*, 21(7):1074–1086, may 2011. doi: 10.1101/gr.118703.110.
3. B. Zhang, Y. Zhou, N. Lin, R. F. Lowdon, C. Hong, R. P. Nagarajan, J. B. Cheng, D. Li, M. Stevens, H. J. Lee, X. Xing, J. Zhou, V. Sundaram, G. Elliott, J. Gu, T. Shi, P. Gascard, M. Sigaroudinia, T. D. Tlsty, T. Kadlecek, A. Weiss, H. O'Geen, P. J. Farnham, C. L. Maire, K. L. Ligon, P. A. F. Madden, A. Tam, R. Moore, M. Hirst, M. A. Marra, B. Zhang, J. F. Costello, and T. Wang. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the m&m algorithm. *Genome Research*, 23(9): 1522–1540, jun 2013. doi: 10.1101/gr.156539.113.
4. Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14 (10):R115, 2013. doi: 10.1186/gb-2013-14-10-r115.
5. Osman El-Maarri, Tim Becker, Judith Junen, Syed Saadi Manzoor, Amalia Diaz-Lacava, Rainer Schwaab, Thomas Wienker, and Johannes Oldenburg. Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. *Human Genetics*, 122(5):505–514, sep 2007. doi: 10.1007/s00439-007-0430-3.
6. Marco P. Boks, Eske M. Derks, Daniel J. Weisenberger, Erik Strengman, Esther Janson, Iris E. Sommer, René S. Kahn, and Roel A. Ophoff. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS ONE*, 4(8):e6767, aug 2009. doi: 10.1371/journal.pone.0006767.
7. Lindsay M. Webb and Mireia Guerau de Arellano. Emerging role for methylation in multiple sclerosis: Beyond DNA. *Trends in Molecular Medicine*, 23(6):546–562, jun 2017. doi: 10.1016/j.molmed.2017.04.004.
8. M. Iridoy Zulet, L. Pulido Fontes, T. Ayuso Blanco, F. Lacruz Bescos, and M. Mendioroz Iriarte. Epigenetic changes in neurology: DNA methylation in multiple sclerosis. *Neurología (English Edition)*, 32(7):463–468, sep 2017. doi: 10.1016/j.nrleng.2015.03.020.
9. Vardhman K. Rakyan, Huriya Beyan, Thomas A. Down, Mohammed I. Hawa, Siarhei Maslau, Deeqo Aden, Antoine Daunay, Florence Busato, Charles A. Mein, Burkhard Manfras, Kerith-Rae M. Dias, Christopher G. Bell, Jörg Tost, Bernhard O. Boehm, Stephan Beck, and R. David Leslie. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genetics*, 7(9):e1002300, sep 2011. doi: 10.1371/journal.pgen.1002300.
10. Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, Klementy Shchetynsky, Annika Scheynius, Juha Kere, Lars Alfredsson, Lars Klareskog, Tomas J Ekström, and Andrew P Feinberg. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2): 142–147, jan 2013. doi: 10.1038/nbt.2487.
11. Anatoliy Melnikov, Denise Scholtens, Andrew Godwin, and Victor Levenson. Differential methylation profile of ovarian cancer in tissues and plasma. *The Journal of Molecular Diagnostics*, 11(1):60–65, jan 2009. doi: 10.2353/jmoldx.2009.080072.
12. Ekaterina Olkhov-Mitsel, Andrea J. Savio, Ken J. Kron, Vaijayanti V. Pethe, Thomas Hermanns, Neil E. Fleshner, Bas W. van Rhijn, Theodorus H. van der Kwast, Alexandre R. Zlotta, and Bharati Bapat. Epigenome-wide DNA methylation profiling identifies differential methylation biomarkers in high-grade bladder cancer. *Translational Oncology*, 10(2): 168–177, apr 2017. doi: 10.1016/j.tranon.2017.01.001.
13. Nicole M White, Miles C Benton, Daniel W Kennedy, Andrew Fox, Lyn R Griffiths, Rodney A Lea, and Kerrie L Mengersen. Accounting for cell lineage and sex effects in the identification of cell-specific DNA methylation using a bayesian model selection algorithm. apr 2017. doi: 10.1101/124826.
14. Steve Horvath, Vei Mah, Ake T. Lu, Jennifer S. Woo, Oi-Wa Choi, Anna J. Jasinska, José A. Riancho, Spencer Tung, Natalie S. Coles, Jonathan Braun, Harry V. Vinters, and L. Stephen Coles. The cerebellum ages slowly according to the epigenetic clock. *Aging*, 7(5):294–306, may 2015. doi: 10.18632/aging.100742.
15. Steffan D. Bos, Christian M. Page, Bettina K. Andreassen, Emon Elboudwarej, Marte W. Gustavsen, Farren Briggs, Hong Quach, Ingvild S. Leikfoss, Anja Bjølgerud, Tone Berge, Hanne F. Harbo, and Lisa F. Barcellos. Genome-wide DNA methylation profiles indicate CD8+ t cell hypermethylation in multiple sclerosis. *PLOS ONE*, 10(3):e0117403, mar 2015. doi: 10.1371/journal.pone.0117403.
16. Andrew E Jaffe and Rafael A Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15(2):R31, 2014. doi: 10.1186/gb-2014-15-2-r31.
17. E Andres Houseman, Karl T Kelsey, John K Wiencke, and Carmen J Marsit. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics*, 16(1), mar 2015. doi: 10.1186/s12859-015-0527-y.
18. Carolina M Montaño, Rafael A Irizarry, Walter E Kaufmann, Konrad Talbot, Raquel E Gur, Andrew P Feinberg, and Margaret A Taub. Measuring cell-type specific differential methylation in human brain tissue. *Genome Biol*, 14(8):R94, 2013. doi: 10.1186/gb-2013-14-8-r94.
19. Pavlo Lutsik, Martin Slawski, Gilles Gasparoni, Nikita Vedeneev, Matthias Hein, and Jörn

Method	Advantages	Limitations
Bayes-CDM	• implicit boundary constraints on data • cell-type lineage information incorporated. • consistent performance for identifying cell-type differential methylation over different cell-types. • Optimal method for identifying DM in CD8$^+$T, and CD19$^+$B.	• Slightly suboptimal for identifying differentially methylated loci. • necessary to set λ_0 *a priori*.
Orthogonal Effects Linear Regression	• Most powerful method for finding differentially methylated loci. • Optimal method for identifying DM in CD14$^+$ Monocyte, Neutrophil, CD4$^+$T and Natural Killer types.	• Doesn't explicitly make cell-type level differential methylation predictions.
Aggregating Linear Regression		• limited degrees of freedom means method is unable to capture heterogeneity variation in Whole Blood. • Sub-optimal performance for predicting differential methylation in all cell-types and predicting differentially methylated loci.
LASSO	• Powerful method for identifying differentially methylated loci. • effective for predicting CD4$^+$T differential methylation.	• High rate of falsely predicting differential methylation at the cell-type level.
PSEA	• Optimal method for predicting correctly a single differentially methylated cell-type.	• Poor performance when number of differentially methylated cell-types is greater than 1.

Walter. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biology*, 18(1), mar 2017. doi: 10.1186/s13059-017-1182-6.

20. Shahin Mohammadi, Neta Zuckerman, Andrea Goldsmith, and Ananth Grama. A critical survey of deconvolution methods for separating cell types in complex tissues. *Proceedings of the IEEE*, 105(2):340–366, feb 2017. doi: 10.1109/jproc.2016.2607121.
21. Eugene Andres Houseman, Molly L Kile, David C Christiani, Tan A Ince, Karl T Kelsey, and Carmen J Marsit. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. Technical report, jan 2016.
22. Jeffrey T. Leek and John D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007. doi: 10.1371/journal.pgen.0030161.
23. Elior Rahmani, Noah Zaitlen, Yael Baran, Celeste Eng, Donglei Hu, Joshua Galanter, Sam Oh, Esteban G Burchard, Eleazar Eskin, James Zou, and Eran Halperin. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nature Methods*, 13(5):443–445, mar 2016. doi: 10.1038/nmeth.3809.
24. Charles E. Breeze, Dirk S. Paul, Jenny van Dongen, Lee M. Butcher, John C. Ambrose, James E. Barrett, Robert Lowe, Vardhman K. Rakyan, Valentina Iotchkova, Mattia Frontini, Kate Downes, Willem H. Ouwehand, Jonathan Laperle, Pierre-Étienne Jacques, Guillaume Bourque, Anke K. Bergmann, Reiner Siebert, Edo Vellenga, Sadia Saeed, Filomena Matarese, Joost H.A. Martens, Hendrik G. Stunnenberg, Andrew E. Teschendorff, Javier Herrero, Ewan Birney, Ian Dunham, and Stephan Beck. eFORGE: A tool for identifying cell type-specific signal in epigenomic data. *Cell Reports*, 17(8):2137–2150, nov 2016. doi: 10.1016/j.celrep.2016.10.059.
25. Alexandre Kuhn, Doris Thu, Henry J Waldvogel, Richard L M Faull, and Ruth Luthi-Carter. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature Methods*, 8(11):945–947, oct 2011. doi: 10.1038/nmeth.1710.
26. Stan Development Team. RStan: the R interface to Stan, 2016. R package version 2.14.1.
27. Eugene Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012. doi: 10.1186/1471-2105-13-86.
28. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.
29. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246.
30. Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587, 2010. doi: 10.1186/1471-2105-11-587.
31. G. Casella. Empirical bayes gibbs sampling. *Biostatistics*, 2(4):485–500, dec 2001. doi: 10.1093/biostatistics/2.4.485.
32. Yves F. Atchadé. A computational framework for empirical bayes inference. *Statistics and Computing*, 21(4):463–473, jun 2010. doi: 10.1007/s11222-010-9182-3.
33. Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, jun 2008. doi: 10.1198/016214508000000337.
34. Qing Li and Nan Lin. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, mar 2010. doi: 10.1214/10-ba506.
35. Chenlei Leng, Minh-Ngoc Tran, and David Nott. Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244, sep 2013. doi: 10.1007/s10463-013-0429-6.