PLOS ONE

RESEARCH ARTICLE

A novel family of beta mixture models for the differential analysis of DNA methylation data: An application to prostate cancer

Koyel Majumdar[1], Romina Silva[2,3], Antoinette Sabrina Perry[3,4], Ronald William Watson[2,3], Andrea Rau[5], Florence Jaffrezic[5], Thomas Brendan Murphy[1], Isobel Claire Gormley[1]*

1 School of Mathematics and Statistics, University College Dublin, Dublin, Ireland, **2** School of Medicine, University College Dublin, Dublin, Ireland, **3** Conway Institute of Biomedical and Biomolecular Research, University College Dublin, Dublin, Ireland, **4** School of Biology and Environmental Science, University College Dublin, Dublin, Ireland, **5** INRAE, UMR1313 AgroParisTech, GABI, Université Paris-Saclay, Gif-sur-Yvette, France

* claire.gormley@ucd.ie

Citation: Majumdar K, Silva R, Perry AS, Watson RW, Rau A, Jaffrezic F, et al. (2024) A novel family of beta mixture models for the differential analysis of DNA methylation data: An application to prostate cancer. PLoS ONE 19(12): e0314014. https://doi.org/10.1371/journal.pone.0314014

Editor: Nikos Kavallaris, Karlstad University: Karlstads Universitet, SWEDEN

Received: March 19, 2024

Accepted: November 4, 2024

Published: December 11, 2024

Data Availability Statement: All data and code are available via the R package called betaclust which is freely available at https://github.com/koyelucd/betaclust and on CRAN at https://cran.r-project.org/web/packages/betaclust/index.html.

Funding: KM was funded with the financial support of Science Foundation Ireland (www.sfi.ie) under Grant number 18/CRT/6049. The funder played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Identifying differentially methylated cytosine-guanine dinucleotide (CpG) sites between benign and tumour samples can assist in understanding disease. However, differential analysis of bounded DNA methylation data often requires data transformation, reducing biological interpretability. To address this, a family of beta mixture models (BMMs) is proposed that (i) objectively infers methylation state thresholds and (ii) identifies differentially methylated CpG sites (DMCs) given untransformed, beta-valued methylation data. The BMMs achieve this through model-based clustering of CpG sites and by employing parameter constraints, facilitating application to different study settings. Inference proceeds via an expectation-maximisation algorithm, with an approximate maximization step providing tractability and computational feasibility. Performance of the BMMs is assessed through thorough simulation studies, and the BMMs are used for differential analyses of DNA methylation data from a prostate cancer study. Intuitive and biologically interpretable methylation state thresholds are inferred and DMCs are identified, including those related to genes such as GSTP1, RASSF1 and RARB, known for their role in prostate cancer development. Gene ontology analysis of the DMCs revealed significant enrichment in cancer-related pathways, demonstrating the utility of BMMs to reveal biologically relevant insights. An R package `betaclust` facilitates widespread use of BMMs.

Introduction

Epigenetics is the study of heritable changes in gene activity that do not involve explicit changes to the DNA sequence [1]. DNA methylation is an epigenetic process where a methyl group is added to or removed from the 5′ carbon of the cytosine ring [2]. This process assists in regulating gene expression and is essential for the development of an organism, but irregular changes in DNA methylation patterns can lead to damaging health effects [3]. A cytosine-

Competing interests: The authors have declared that no competing interests exist.

guanine dinucleotide (CpG) site is hypomethylated if neither of the DNA strands in a diploid individual are methylated, hemimethylated if either of the DNA strands are methylated or hypermethylated if both the strands are methylated. A differentially methylated CpG site (DMC) is a CpG site that has different methylation states between DNA samples collected from different biological conditions, which may have been taken from tissues of an individual over time, different tissues from the same individuals or distinct individuals.

The DNA methylation process has been extensively studied in the context of cancer, and its treatment [4]. CpG islands that remain unmethylated in normal cells can become methylated in abnormal cells such as cancer cells [5], and it has been shown that tumour suppressor genes are silenced by hypermethylation of their promoter regions [6, 7]. For example, in prostate cancer, the fifth major cause of cancer-related mortality globally [8], hypermethylation of certain tumour suppressor genes, such as GSTP1, RARB, APC and RASSF1, has been observed during the early stages of the disease [9–11]. A better understanding of disease can therefore be achieved by identifying regions that are differentially methylated between benign and tumour samples.

The Illumina MethylationEPIC BeadChip microarray [12] is used to interrogate over 850,000 CpG sites and retrieve methylation profiling of the CpG sites in the human genome. The Illumina microarray produces a sample of methylated (*Methylated*) and unmethylated (*Unmethylated*) light signal intensities, and the level of methylation, or the *beta* value, is $beta = \max(Methylated)/(\max(Methylated) + \max(Unmethylated) + \chi)$, where χ is a constant offset added for regularisation in case of very low *Methylated* and *Unmethylated* values [13]. The methylation level at a CpG site is quantified by this *beta* value and is constrained to lie between 0 and 1. The *beta* values are continuous with a value close to 1 suggesting that a site is hypermethylated, while values close to 0 represent hypomethylation. The two probe intensities are assumed to be gamma-distributed as they can take only positive values, and their ratio results in beta distributed variables. Thus, the *beta* values can be modelled using a beta distribution.

The *beta* values in general have higher variance in the center of the [0,1) interval than towards its endpoints. This leads to heteroscedasticity, which imposes challenges for analyses as assumptions for the ubiquitous Gaussian models are violated. Hence, *beta* values are usually converted to *M*-values using a logit transformation as these values are statistically more convenient; Gaussian models can be used as the transformed data lie within $(-\infty, \infty)$ [13]. However, such transformations make inference less biologically interpretable and hence there is a need to model the *beta* values in their innate form.

In many methylation array studies, thresholds of *beta* values are subjectively selected to identify the three methylation states. For instance, [14] deemed a CpG site to be hypomethylated if its *beta* value was <0.2 and hypermethylated if its *beta* value was >0.8, while [15] employed 0.3 and 0.7 as thresholds. Such subjective selection of thresholds may increase the likelihood of false positives and negatives, leading to incorrect inference and necessitating an objective approach to determining methylation state thresholds.

Mixture models for transformed *beta* values have been proposed in several studies to find biologically meaningful clusters. For instance, [16, 17] use mixture models to model a subset of CpG sites and cluster samples into latent groups of biologically related samples. Additionally, methods such as the variational Bayes beta mixture model [18] and the Dirichlet process beta mixture model [19] analyse untransformed *beta* values; the former addresses the feature selection problem in the context of DNA methylation data, whereas the latter models the beta values to identify DNA methylation subgroups. The `Methylmix` [20] R [21] package uses a univariate beta mixture model to uncover patient subgroups with similar DNA methylation levels for a specific CpG site, with Wilcoxon rank sum tests used to establish hypermethylated and hypomethylated genes. Beta mixture models have also been proposed for intra-array

quantile normalization [22] and for clustering individual DNA samples into the three methylation states which can then be used to classify cancer tissue type [23]. The use of a beta mixture model has been extended for classifying the methylation states of CpG sites; the approach accounts for boundary values and employs a method-of-moments approach to inference, but considers only small numbers of CpG sites [24]. While mixture models for DNA methylation data have been used for a range of purposes, they have not been utilised to uncover differential methylation across the genome using untransformed *beta* values.

Several methods have been developed for detecting DMCs in different DNA sample types. For instance, the `PanDM` method [25] leverages joint modeling to perform methylation site clustering, differential methylation detection, and pan-cancer pattern discovery by modelling the transformed p-values associated with each CpG site for a given cancer type. A principal component analysis and tensor decomposition approach involving unsupervised feature selection [26] was proposed where principal component scores were associated with each CpG site and used to identify DMCs. Another approach, termed `FastDMA` [27], employs an analysis of covariance to perform both single probe analysis and differentially methylated region scanning while modelling the *M*-values. The popular `limma` method [28] identifies DMCs by modelling the *M*-values using an empirical Bayesian approach. Other studies identify the DMCs by modelling the *beta* values via multiple moderated t-tests or Wilcoxon rank-sum tests [29, 30]. Additionally, a multiple hypothesis testing approach, combined with multivariate permutation tests, has been proposed to detect group differences in epigenetic data [31], as has a nonparametric test to identify DMCs between multiple treatments [32]; while this approach can analyze smaller arrays with e.g., 28,000 CpG sites, it is computationally intensive for modern, larger-scale arrays. These approaches to DMC identification use subjective thresholds, transformed values, moderated *t*-tests and/or nonparameteric methods [33]. Crucially, such approaches lack biological interpretability and often face reproducibility and computational scalability challenges when considering data from different studies, of the scale resulting from current microarray technologies.

Several mixture models for bounded data are available. For instance, in the context of semiparametric density estimation, [34] fit a Gaussian mixture model to range-power transformed bounded data, from which the density for the original data is obtained. Mixture models of bounded Laplace distributions also allow for modelling bounded data by truncation of the Laplace distribution, but are computationally expensive for large datasets [35]. Similarly, bounded support asymmetric generalized Gaussian mixture models are adaptable to different distributional shapes but can be computationally expensive as inference requires numerical optimisation [36]. A beta mixture model is an appropriate choice for bounded DNA methylation data: the support of the beta distribution is congruent with the *beta* values, its flexibility allows for skew and symmetric distributional shapes, and it is computationally feasible to work with given its parsimony. Importantly, the beta distribution parameters provide relevant biological interpretations enabling biologically intuitive and meaningful inference.

Here we propose a family of beta mixture models (BMMs), which address specific research questions arising in the context of differential analysis of DNA methylation data, by introducing a range of constraints on the parameters of a BMM. The resulting novel family of BMMs facilitates a model-based approach to clustering CpG sites given their innate beta-valued methylation data to (i) objectively identify methylation state thresholds and (ii) identify DMCs between different sample types. The BMMs are capable of clustering the entire microarray of CpG sites, from DNA samples collected from multiple tissues from each of several patients, in a computationally efficient manner. Performance is assessed through simulation studies, and the BMMs are used to analyse a motivating prostate cancer (PCa) dataset. The capability of the BMMs is demonstrated to appositely model the *beta* values, to objectively identify thresholds

and to identify existing and novel DMCs, including those related to genes implicated in prostate cancer, such as GSTP1, RARB and RASSF1. An R package, `betaclust`, freely available on github and CRAN, facilitates widespread use of the BMMs.

Methods

Prostate cancer data

A prostate cancer study [37], which involved collection of DNA methylation samples from four patients with metastatic prostate cancer disease, motivated the development of the BMMs. Tissue samples from matched biopsy cores (tumour and histologically matched normal—herein benign) were collected from each patient, and DNA was extracted from the samples. Methylation profiling of the DNA samples was conducted using the Infinium MethylationEPIC Beadchip [38]. The raw DNA methylation data are freely available for download on the Gene Expression Omnibus (GEO) repository (GSE119260); datasets GSM3362390-GSM3362397 were analysed here and were accessed on 26th of January, 2021 for research purposes. The authors had no access to information that could identify individual participants.

Observed *beta* values for each of 694,923 CpG sites for the two DNA sample types were collected from each of the four patients. Raw methylation array data was quality controlled and pre-processed as in [37], where the data were normalized, and probes overlapping with SNPs, probes with the highest fraction of unreliable measurements, probes lying outside of CpG sites and those on the sex chromosome were removed. The resulting dataset had 103 CpG sites ($< 0.014\%$ of the total number of CpG sites) with missing *beta* values. While imputation techniques exist for DNA methylation data, missing values were not imputed here due to their very low percentage, and the high uncertainty associated with imputed values in diseased samples due to their heterogeneity [39]; here the 103 CpG sites with missing data were therefore removed. No observed *beta* values were equal to 0. The resulting dataset contained *beta* values for $C = 694,820$ CpG sites from each of $R = 2$ DNA sample types collected from each of $N = 4$ patients. Here, these data are appositely modeled in their innate *beta* form to (i) objectively identify methylation state thresholds and (ii) uncover DMCs between two sample types using a model-based clustering approach.

A beta distribution

The beta distribution has support on [0, 1] and is parameterized by two positive shape parameters, α and δ. Given the properties of the *beta* values, the beta distribution is used here to appositely model the methylation level x_{cnr} of the c^{th} CpG site ($c = 1, \ldots, C$), from the n^{th} patient ($n = 1, \ldots, N$), from their r^{th} DNA sample type ($r = 1, \ldots, R$) i.e.,

$$f(x_{cnr}|\alpha, \delta) \sim \text{Beta}(x_{cnr}|\alpha, \delta) = \frac{{x_{cnr}}^{\alpha-1}(1 - x_{cnr})^{\delta-1}}{\text{B}(\alpha, \delta)},$$

for $0 \leq x_{cnr} \leq 1$, where $\text{B}(\alpha, \delta) = (\Gamma(\alpha)\Gamma(\delta))/\Gamma(\alpha + \delta)$, where $\Gamma(\cdot)$ is the gamma function. The DNA methylation data are collected in the $C \times NR$ dimensional dataset $\mathbf{X}$ where each of the NR columns contains the methylation levels of the C CpG sites in one of the R sample types from each of the N patients.

A beta mixture model

A mixture model assumes the observed data have been generated from a heterogeneous population composed of K groups or clusters. In the context of DNA methylation data, there are

$G = 3$ possible methylation states: hypomethylation, hemimethylation or hypermethylation. Hence, when analysing methylation data from a single DNA sample type (i.e., where $R = 1$) each CpG site exhibits one of $K = G^R = 3$ methylation states characterised by each of the K clusters. Here, interest lies in objectively inferring thresholds between these $K = 3$ methylation states.

When analysing methylation data across multiple (i.e., $R > 1$) DNA sample types to identify DMCs, each CpG site will exhibit one of a possible $K = G^R$ combinations of methylation states, here characterised by each of K clusters in a mixture model. For example, given the three methylation states and considering a CpG site across $R = 2$ sample types (e.g., across benign and tumour samples), the CpG site can potentially exhibit any of the $K = 3^2 = 9$ combinations of these three states (for example hypermethylated in both samples, hypermethylated in one sample and hypomethylated in the other, etc). Therefore, in this scenario, $K = 3^2 = 9$ with each cluster characterising one of the possible methylation state combinations.

We propose a beta mixture model for the methylation data for all cases $R \geq 1$. Here $\boldsymbol{\theta}$ is used to denote the shape parameters in a beta mixture model, i.e., $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\alpha}_K, \boldsymbol{\delta}_K)$, where $\boldsymbol{\alpha}_k$ and $\boldsymbol{\delta}_k$ are the shape parameters of cluster k. The shape parameters are allowed to vary among the clusters, patients and sample types. The mixing proportions $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ lie between 0 and 1, $\sum_{k=1}^{K} \tau_k = 1$, and denote the probability of belonging to cluster $k\ \forall k = 1, \ldots, K$. Independence is assumed across patients and samples, given a CpG site's cluster membership, leading to the probability density function for such a beta mixture model (BMM):

$$f(\mathbf{X}|\boldsymbol{\tau}, \boldsymbol{\theta}) = \prod_{c=1}^{C} \sum_{k=1}^{K} \tau_k f(\mathbf{X}|\boldsymbol{\alpha}_k, \boldsymbol{\delta}_k) = \prod_{c=1}^{C} \sum_{k=1}^{K} \tau_k \prod_{n=1}^{N} \prod_{r=1}^{R} \text{Beta}(x_{cnr}|\alpha_{knr}, \delta_{knr}) \quad (1)$$

Computation of maximum likelihood estimates (MLEs) of $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ from the associated log likelihood function is complex, and an incomplete data approach is therefore used here. The latent vector $\mathbf{z}_c = (z_{c1}, \ldots, z_{cK})$ is introduced for each CpG site c, where z_{ck} is 1 if CpG site c belongs to the k^{th} group and 0 otherwise. The $C \times K$ matrix $\mathbf{Z}$ is combined with the *beta* values to form the complete data ($\mathbf{X}$, $\mathbf{Z}$). The complete data log-likelihood function is

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C} \sum_{k=1}^{K} z_{ck} \{\log \tau_k + \sum_{n=1}^{N} \sum_{r=1}^{R} \log[\text{Beta}(x_{cnr}|\alpha_{knr}, \delta_{knr})]\}. \quad (2)$$

The complete data log-likelihood function (2) can be used to find the MLEs $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\theta}}$ using the expectation-maximisation (EM) algorithm [40]; on convergence a probabilistic clustering solution is also available from the expected value of z_{ck}, the posterior probability of CpG site c belonging to cluster k.

A family of BMMs. The most generalised BMM is defined in (1) which models the CpG sites as belonging to K latent groups. By introducing a variety of constraints on the parameters of this generalised BMM, a family of three beta mixture models is proposed. Each model serves a specific purpose e.g., to cluster the CpG sites into the 3 methylation states allowing objective inference of methylation state thresholds, or to facilitate the identification of DMCs between different sample types.

The K·· model. The K·· model facilitates objective inference of thresholds between methylation states by clustering C CpG sites into one of $K = G = 3$ methylation states, based on a single sample type ($R = 1$) from each of N patients. Under the K·· model the shape parameters of each cluster are constrained to be equal for each patient, but allowed to vary across clusters.

The complete data log-likelihood function is therefore

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{1} \log[\text{Beta}(x_{cnr}|\alpha_{k\cdot\cdot}, \delta_{k\cdot\cdot})]\}.$$

The KN· model. The KN· model facilitates objective inference of methylation state thresholds by clustering each of the C CpG sites into one of $K = G = 3$ methylation states, based on data from a single sample type ($R = 1$) from each of N patients. While the KN· model has a similar purpose to the K·· model, it differs in that it is less parsimonious as it allows cluster and patient-specific shape parameters. The complete data log-likelihood function is therefore

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{1} \log[\text{Beta}(x_{cnr}|\alpha_{kn\cdot}, \delta_{kn\cdot})]\}.$$

The K·R model. The K·R model facilitates identification of differentially methylated CpG sites between $R > 1$ DNA sample types collected from each of N patients. The K·R model assumes conditional independence between CpG sites from paired samples from the same patient, given the CpG sites' cluster membership. The K·R model also assumes each of the K clusters characterises a different combination of the G methylation states across the R biological conditions where $K = G^R = 9$ here. Under the K·R model the shape parameters are allowed to vary for each sample type and for different clusters but are constrained to be equal for each patient. The complete data log-likelihood function for the K·R model is therefore

$$\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = \sum_{c=1}^{C}\sum_{k=1}^{K} z_{ck}\{\log \tau_k + \sum_{n=1}^{N}\sum_{r=1}^{R} \log[\text{Beta}(x_{cnr}|\alpha_{k\cdot r}, \delta_{k\cdot r})]\}.$$

This family of beta mixture models enables the objective inference of methylation state thresholds (via the K·· and/or KN· models) and the identification of DMCs between R DNA sample types (via the K·R model), as illustrated in the simulation studies and applications that follow.

Parameter estimation

The parameters of the BMMs are estimated and the cluster membership for each CpG site inferred using the EM algorithm. Here, we delineate this for the generalised BMM (1). Derivations for the K··, KN· and K·R models are detailed in Appendices S1–S3 in S1 File.

The EM algorithm consists of two steps: in the expectation step the expected value of the complete data log-likelihood function is obtained, conditional on the observed data and current parameter estimates. The maximisation step maximises the expected complete data log-likelihood with respect to the parameters. To obtain $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\theta}}$, the expectation and maximisation steps are iterated until convergence to at least a local optimum of the log-likelihood function.

An initial clustering of CpG sites is obtained using k-means clustering and the method of moments is used to calculate initial values of $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$. The two steps proceed as follows:

- Expectation-step: the expected value of z_{ck} is calculated, i.e., the posterior probability of CpG site c belonging to cluster k, conditional on current parameter estimates. At iteration $t + 1$

$$\hat{z}_{ck} = \mathbf{E}[z_{ck}|\mathbf{X}, \boldsymbol{\tau}^{(t)}, \boldsymbol{\theta}^{(t)}] = \frac{\tau_k^{(t)} \prod_{n=1}^{N} \prod_{r=1}^{R} \text{Beta}(x_{cnr}|\alpha_{knr}^{(t)}, \delta_{knr}^{(t)})}{\sum_{k'=1}^{K} \left[\tau_{k'}^{(t)} \prod_{n=1}^{N} \prod_{r=1}^{R} \text{Beta}(x_{cnr}|\alpha_{k'nr}^{(t)}, \delta_{k'nr}^{(t)})\right]}.$$

- Maximisation-step: estimates of the parameters $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ are calculated by maximising the expected complete data log-likelihood function, given the $\hat{\mathbf{Z}}$ values from the E-step.

For the maximisation-step, the expected complete data log-likelihood function is maximised by differentiating it w.r.t the parameters. Closed form solutions for the mixing proportions are available as $\hat{\tau}_k = \sum_{c=1}^{C} \hat{z}_{ck}/C, \ \forall \ k = 1, \ldots, .K$. For the shape parameters, the expected complete data log-likelihood function to be maximized is

$$\begin{aligned} \ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}}) = \quad & \sum_{c=1}^{C} \sum_{k=1}^{K} \hat{z}_{ck} \{\log \tau_k + \sum_{n=1}^{N} \sum_{r=1}^{R} [(\alpha_{knr} - 1)\log x_{cnr} + \\ & (\delta_{knr} - 1)\log(1 - x_{cnr}) - \log \text{B}(\alpha_{knr}, \delta_{knr})]\}. \end{aligned} \tag{3}$$

Differentiating (3) w.r.t α_{knr} yields

$$\frac{\partial \ell_C}{\partial \alpha_{knr}} = \sum_{c=1}^{C} \hat{z}_{ck} \{\log x_{cnr} - [\psi(\alpha_{knr}) - \psi(\alpha_{knr} + \delta_{knr})]\} \tag{4}$$

where ψ is the logarithmic derivative of the gamma function known as the digamma function, $\psi(\alpha_{knr}) = \partial \log \Gamma(\alpha_{knr})/\partial \alpha_{knr}$. Similarly, the derivative of $\ell_C(\boldsymbol{\tau}, \boldsymbol{\theta}|\mathbf{X}, \hat{\mathbf{Z}})$ w.r.t δ_{knr} is

$$\frac{\partial \ell_C}{\partial \delta_{knr}} = \sum_{c=1}^{C} \hat{z}_{ck} \{\log(1 - x_{cnr}) - [\psi(\delta_{knr}) - \psi(\alpha_{knr} + \delta_{knr})]\}. \tag{5}$$

Closed form solutions for $\hat{\alpha}_{knr}$ and $\hat{\delta}_{knr}$ are not available due to the presence of the digamma function. To obtain the MLEs, numerical optimisation algorithms such as BFGS [41] and BHHH [42] could be used. However, for the large datasets considered here, use of these algorithms proved to be computationally infeasible.

A digamma approximation. Here an approximation to the digamma function is used to allow for closed form solutions for the shape parameters. The lower bound for the digamma function for all $y > 1/2$ is $\psi(y) > \log(y - 1/2)$ [43]. Given the context, we assume that the beta distributions in the family of BMMs are unimodal and bounded, meaning the shape parameters are >1. Thus, the lower bound approximation holds and was empirically observed to be a very close approximation of the digamma function. The lower bound is used in (4) and (5) to give

$$\frac{\partial \ell_C}{\partial \alpha_{knr}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N} \sum_{r=1}^{R} \left[\log x_{cnr} - \log \frac{\alpha_{knr} - 1/2}{\alpha_{knr} + \delta_{knr} - 1/2}\right] \tag{6}$$

and

$$\frac{\partial \ell_C}{\partial \delta_{knr}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N} \sum_{r=1}^{R} \left[\log(1 - x_{cnr}) - \log \frac{\delta_{knr} - 1/2}{\alpha_{knr} + \delta_{knr} - 1/2} \right]. \quad (7)$$

Equating Eqs (6) and (7) to zero, we get closed-form, approximate estimates as

$$\hat{\alpha}_{knr} = 0.5 + \frac{0.5\exp(-y_2)}{\{[\exp(-y_2) - 1][\exp(-y_1) - 1]\} - 1}$$

and

$$\hat{\delta}_{knr} = \frac{0.5\exp(-y_2)[\exp(-y_1) - 1]}{\{[\exp(-y_2) - 1][\exp(-y_1) - 1]\} - 1},$$

where $y_1 = (\sum_{c=1}^{C} \hat{z}_{ck} \log x_{cnr})/(\sum_{c=1}^{C} \hat{z}_{ck})$ and $y_2 = (\sum_{c=1}^{C} \hat{z}_{ck} \log(1 - x_{cnr}))/(\sum_{c=1}^{C} \hat{z}_{ck})$.

Utilising the digamma function approximation brings notable computational gains with run times, for example, reducing from 65 hours (when using numerical optimisation at the maximisation step) to 15 minutes (when using the digamma approximation) when analysing the PCa data on a computer equipped with an Intel Core i7 CPU with 2.70GHz speed, 6 physical cores and 16 GB of RAM.

Inferring methylation state thresholds

To objectively infer thresholds between methylation states, without loss of generality, we denote by clusters 1 and 2 the clusters representing hypomethylated and hypermethylated CpG sites respectively. The ratio of fitted density estimates ω_j for cluster $j = 1, 2$ is

$$\omega_j = \frac{\tau_j f(\mathbf{X}|\boldsymbol{\alpha}_j, \boldsymbol{\delta}_j)}{\sum_{k \neq j} \tau_k f(\mathbf{X}|\boldsymbol{\alpha}_k, \boldsymbol{\delta}_k)}.$$

The threshold separating e.g., the hypomethylated and hemimethylated clusters is calculated as the minimum *beta* value at which $\omega_1 \geq 1$. Similarly, the threshold dividing the hemimethylated and hypermethylated clusters is the maximum *beta* value at which $\omega_2 \geq 1$.

In the K·· model, as the shape parameters are constrained to be equal for each patient, a single set of thresholds is calculated for all patients. In the KN· model, the shape parameters vary for each patient, so a set of thresholds is calculated for each patient. Unless one model is appropriate given the question of interest, to choose the optimal model between the K·· and KN· models, here the well utilised model selection tools of the Akaike information criterion (AIC) [44], Bayesian information criterion (BIC) [45] and the integrated complete log-likelihood criterion (ICL) [46] are examined.

Identifying the most differentially methylated clusters

While the K·R model clusters the CpG sites into K clusters, subsequent quantification of the degree of differential methylation of CpG sites in each cluster is required. Here, this is quantified by comparing the $R = 2$ beta distributions associated with the two sample types in each cluster (e.g., benign and tumour) using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve as a measure of separability, which has an advantage of being widely used in biology.

The ROC curve for a cluster is generated by drawing 1000 samples from each of the two fitted beta distributions, corresponding to two sample types, after convergence of the EM algorithm. By varying threshold values from 0 to 1, the sampled *beta* values are used to compute sensitivity and specificity for discriminating between sample types, which are in turn used to construct the ROC curve and associated AUC. Higher AUC values indicate greater separation between the fitted beta distributions for two sample types within a cluster, and therefore that the CpG sites within that cluster are more differentially expressed between the two. In the case where $R > 2$, the beta distributions linked to each sample type within a cluster are compared to one another, and the maximum AUC across pairs of sample types is selected as the cluster's AUC. A large AUC would then indicate distinct methylation patterns between at least two sample types. We also consider the Wasserstein distance (WD) [47], which computes the disparity between cumulative distributions, as an additional approach to quantifying the degree of differential methylation of CpG sites between sample types in each cluster.

A process flow diagram that summarises the overall approach to inferring subjective thresholds and identifying DMCs using the proposed BMMs is given in Fig 1.

Results

Simulated data results

Simulated data. One hundred simulated datasets consisting of methylation values for $C = 600,000$ CpG sites are generated using R [21]. Each simulated dataset consists of *Beta* values from two biological sample types (sample A and sample B) from each of $N = 4$ patients. Hypomethylated CpG sites were generated from a *Beta*(2,20) distribution, with hemi- and hyper- values generated from *Beta*(4,3) and *Beta*(20,2) distributions respectively. To emulate the noisy data observed in real settings, zero-centred Gaussian noise with standard deviation 0.01 was added to the beta-generated data; resulting values outside [0, 1] were replaced by the closest minimum or maximum value from the beta-generated data. Reflecting typical behaviour in DNA methylation data, 35%, 35% and 30% of CpG sites in a single sample were simulated as hypomethylated, hemimethylated and hypermethylated, respectively. This resulted in, on average, 64% of the CpG sites being differentially methylated between the two sample types. Of these DMCs, on average, 30% were hypomethylated in one sample and hypermethylated in the other, or vice versa; such highly differentially methylated CpG sites are of prime interest. Fig 2 illustrates a single simulated data set, with clusters of CpG sites ordered from most to least differentially methylated between samples, according to AUC.

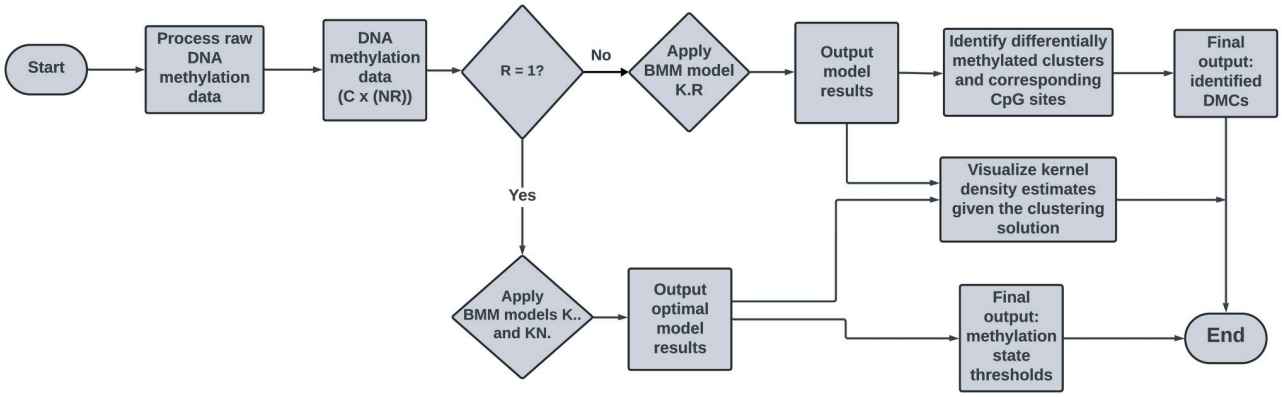

Fig 1. Process flow diagram for the differential analysis of beta-valued DNA methylation data using a novel family of beta mixture models.

https://doi.org/10.1371/journal.pone.0314014.g001

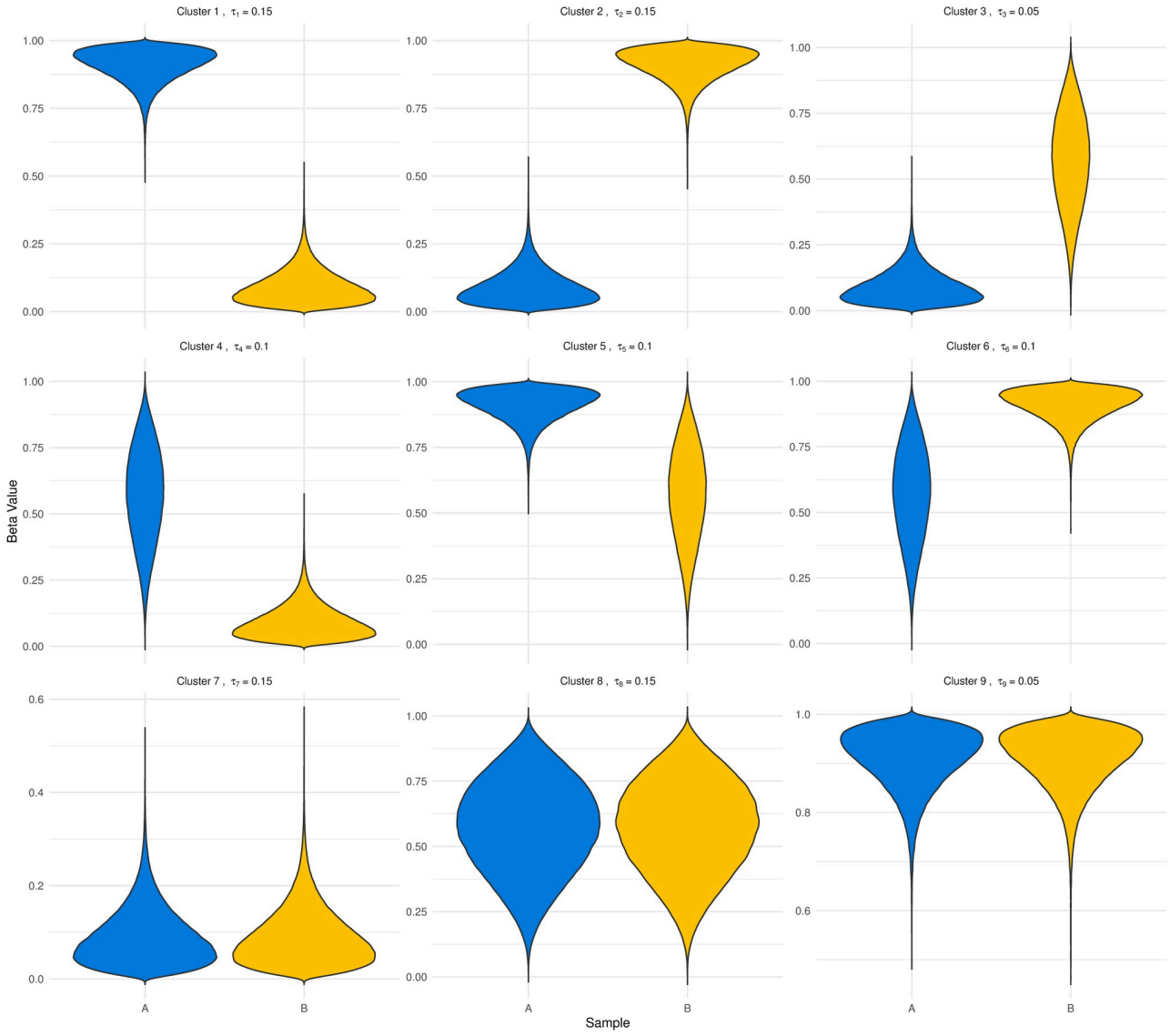

Fig 2. Violin plots of a simulated dataset. Each panel illustrates the simulated beta-distributed values in samples A and B. The proportion of CpG sites in each cluster is detailed in the panel title. Clusters are ordered numerically from most to least differentially methylated, according to AUC.

https://doi.org/10.1371/journal.pone.0314014.g002

Estimating methylation state thresholds. To cluster the CpG sites in a sample type into 3 clusters representing the 3 methylation states and to infer the thresholds between these states, the K·· and KN· models were fitted to the data from sample type A in each of the 100 simulated datasets. The true K·· generating model was selected by AIC, BIC and ICL to be optimal in each case. Fig 3 illustrates the density estimates under the clustering solution of the K·· model for a single simulated dataset. The hemimethylated CpG sites are clustered in cluster 1, while the hypomethylated and hypermethylated CpG sites are in clusters 2 and 3 respectively. The estimated mixing proportion of CpG sites for each cluster (see Fig 3) are notably similar to the true mixing proportions. As parameters are constrained to be equal for each patient in the K·· model, a single set of thresholds is inferred for all 4 patients. The threshold (see Fig 3) of 0.258 indicates that any CpG site with a lower *beta* value is likely to be hypomethylated. Similarly

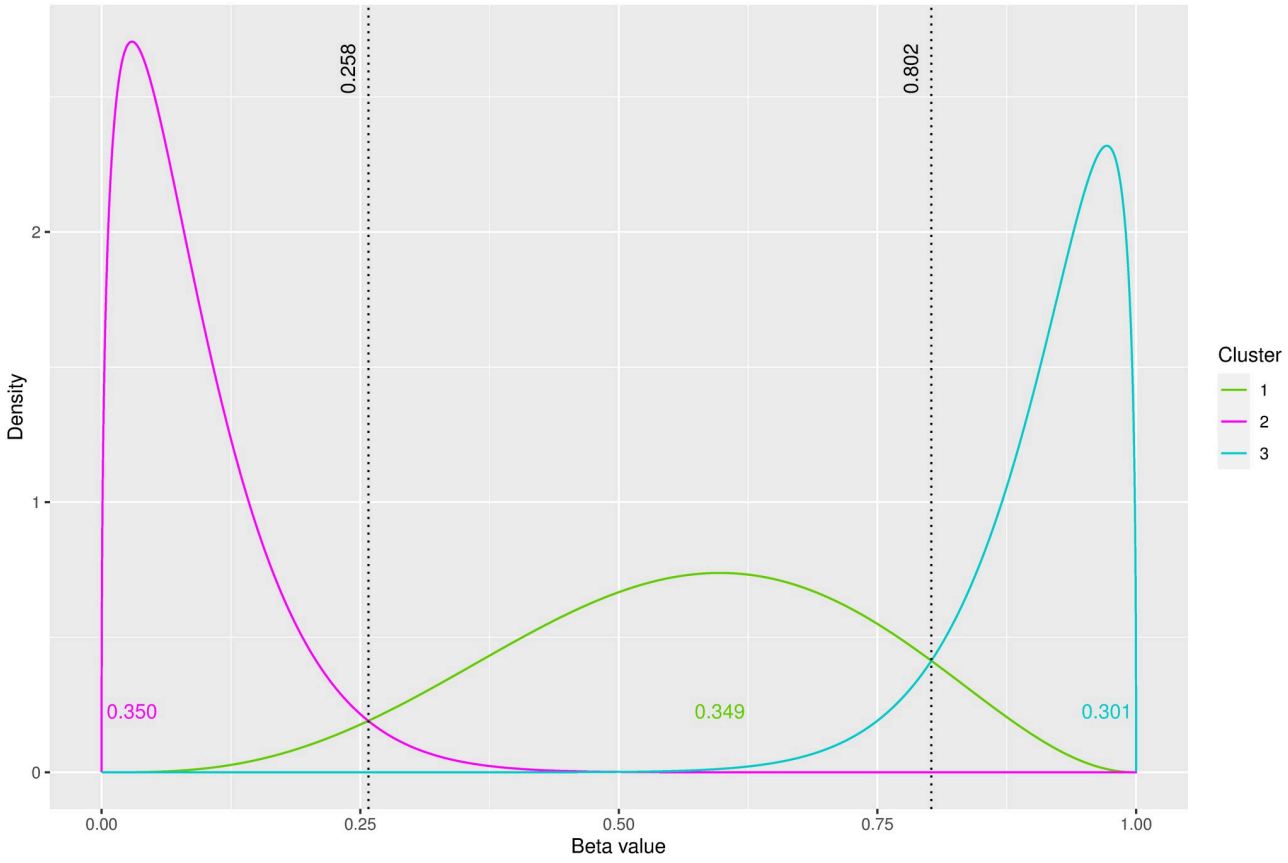

Fig 3. Fitted density estimates under the K·· model on a simulated dataset from sample type A. The thresholds between methylation states are illustrated by black dotted lines. The estimated mixing proportions are also displayed.

https://doi.org/10.1371/journal.pone.0314014.g003

any CpG site with a *beta* value greater than the second threshold of 0.802 is likely to be hypermethylated. These objectively inferred thresholds are very close to the true thresholds of 0.244 and 0.808.

The adjusted Rand index (ARI) [48] gives a measure of agreement between two clustering solutions, where an ARI of 1 indicates full agreement. The mean ARI across the 100 simulated datasets for the K·· model was 0.9949 (s.d. 0.0002) and for the KN· model was 0.9949 (s.d. 0.0002), demonstrating accurate and stable clustering solutions. The mean ARI when comparing the K·· and KN· clustering solutions was 0.999 (s.d. 0.00001). A summary of the parameter estimates and kernel density plots under the K·· model are available in Appendices S4–S5 in S1 File.

Identifying DMCs. To identify differentially methylated CpG sites between multiple DNA sample types in the simulated data, the K·R model is fitted to each of the $C \times NR$ dimensional datasets. For each CpG site, as there are $R = 2$ sample types, $G^R = 9$ different combinations of the three methylation states are possible across sample types A and B. The CpG sites that are e.g., hypomethylated in one sample type and hypermethylated in the other are of interest as they indicate potential epigenetic changes in the genome.

Under the K·R model, for each simulated data set, the AIC, BIC and ICL criteria were non-informative as they consistently decreased for $K = 2, \ldots, 30$ (see Appendix S6 in S1 File). Thus, the biologically motivated K·R model with $K = G^R = 9$ was considered here. The AUC and WD

Table 1. Mean and standard deviation (s.d.) of the AUC and WD metrics for each cluster across the 100 simulated datasets.

		Cluster								
		1	2	3	4	5	6	7	8	9
AUC	Mean	1.0000	1.0000	0.9951	0.9940	0.9726	0.9456	0.5225	0.5139	0.5041
	S.D.	0.0000	0.0000	0.0009	0.0009	0.0028	0.0508	0.0147	0.0115	0.0076
WD	Mean	0.8187	0.8185	0.4796	0.4797	0.3377	0.3162	0.0043	0.0040	0.0005
	S.D.	0.0001	0.0001	0.0005	0.0004	0.0012	0.0456	0.0091	0.0089	0.0015

https://doi.org/10.1371/journal.pone.0314014.t001

metrics were employed to assess the similarity between the $R = 2$ probability distributions within each cluster, giving insight to the degree of differential methylation within clusters. Table 1 shows the mean and standard deviation of the AUC and WD values for each cluster across the 100 simulated datasets. Throughout, clusters are presented in descending order of their degree of differential methylation, based on decreasing AUC and, in the case of ties, WD values. The six differentially methylated clusters i.e., those in which the methylation state was different between the two sample types, are correctly highlighted as the most differentially methylated.

The graph in Fig 4 shows the fitted density estimates of the clustering solution under the K·R model for a single simulated dataset. The associated dissimilarity metrics correctly indicate clusters 1–6 as the most differentially methylated clusters, with 65.1% of CpG sites

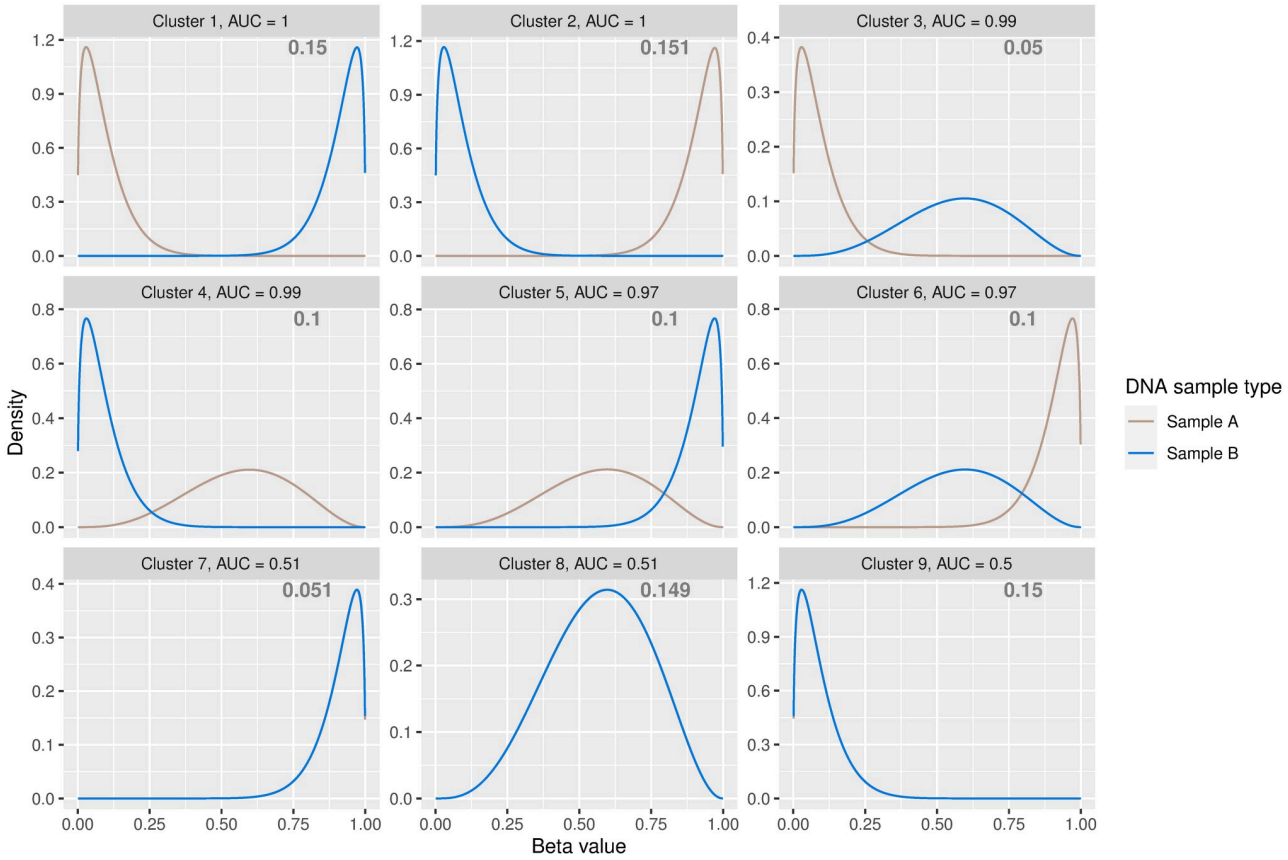

Fig 4. Fitted density estimates under the K·R model on a simulated dataset. The estimated mixing proportions are displayed in the relevant panel.

https://doi.org/10.1371/journal.pone.0314014.g004

belonging to these clusters, which is very close to the true mixing proportions. The density estimates show that, for example, cluster 1 captures DMCs which are hypomethylated in sample type A and hypermethylated in sample type B while cluster 2 contains DMCs which are hypermethylated in sample type A and hypomethylated in sample type B. Mean standard performance metrics across the 100 simulated datasets signify an accurate and stable clustering process, with a mean false discovery rate (FDR) of 0.0041 (s.d. 0.0121), mean sensitivity of 0.9742 (s.d. 0.0563), mean specificity of 0.9921 (s.d. 0.0244) and mean ARI of 0.9758 (s.d. 0.0370). A summary of parameter estimates under the K·R model is available in Appendix S4 in S1 File with kernel density estimates in Appendix S7 in S1 File.

All computations were conducted using R [21] on a Windows 11 operating system equipped with an Intel Core i7 CPU with 2.70GHz speed and 16GB RAM. In terms of computational cost, for example, fitting the K·R model to a single simulated dataset took 55.86 seconds on a computer with 6 cores. To explore the impact of an increasing value of N on the computational cost, further simulation studies in which $N = \{8, \ldots, 60\}$, where N increased in increments of 4, demonstrated a linear increase in computational cost. This is intuitive given the form of the model's likelihood function with respect to N. Further details on computational cost are provided, for all three BMM models, in Appendix S8 in S1 File.

The ability of the K·R model to detect DMCs was compared with that of the state-of-the-art `limma` method [28], which requires the beta values to be transformed into M-values for analysis. Results indicate `limma` also performs well but with lower accuracy than the BMM approach: `limma` had a mean FDR of 0.0080 (s.d. 0.0001), mean sensitivity of 0.9074 (s.d. 0.0008), mean specificity of 0.9864 (s.d. 0.0003) and mean ARI of 0.7562 (s.d. 0.0018). The lower mean sensitivity value in particular suggests that `limma` identified fewer DMCs than the BMM. Boxplots displaying the performance metrics across the 100 simulations for the K·R model and `limma` are available in Appendix S9 in S1 File.

Finally, to explore the robustness of the BMM approach to model misspecification, the same simulation settings were considered but where the data were simulated from a t-distribution with 8 degrees of freedom. Both the K·R model and `limma` demonstrated mixed ability to detect DMCs when applied to the expit-transformed and logit-transformed data respectively. The BMM and `limma` approaches attained, respectively, a mean FDR of 0.4918 (s.d. 0.076) and 0.7018 (s.d. 0.0006), a mean sensitivity of 0.4558 (s.d. 0.0908) and 0.9995 (s.d. 0.0002) and a mean specificity of 0.875 (s.d. 0.0203) and 0.3279 (s.d. 0.0019). The BMM's low sensitivity and high specificity suggests that the BMM identified the non-DMCs correctly but failed to detect a large portion of true DMCs, while `limma` demonstrated the opposite ability. Boxplots of the performance metrics are provided in Appendix S10 in S1 File.

Prostate cancer data results

Estimating methylation state thresholds. For the PCa data, to cluster the CpG sites into the 3 methylation states and objectively infer the methylation state thresholds in the benign and tumour sample types, the K·· and KN· models were fitted. The AIC, BIC and ICL suggest the KN· model as optimal for both the benign and tumour sample types. This is intuitive, particularly for the tumour sample types where the degree of disease varies for each patient, as the KN· model allows for patient specific shape parameters.

The fitted density estimates and inferred thresholds for patient 1 are discussed here; those for patients 2, 3 and 4 are available in Appendix S11 in S1 File.

In the benign sample type from patient 1, the estimated mixing proportions were 0.244 for hypomethylation, 0.363 for hemimethylation, and 0.393 for hypermethylation. The inferred methylation state thresholds are 0.258 and 0.747 for the benign sample type, and 0.19 and

0.751 for the tumour sample type. The hypermethylation state thresholds in the benign and tumour sample types are very close; in contrast, the hypomethylation state thresholds are quite different. While these objective thresholds are close to the subjective values suggested in the literature of 0.2 and 0.8, the difference results in more hypo- and hypermethylated CpG sites being identified by the BMM as DMCs.

Patient 1 was known to have a greater degree of disease severity than the other patients with a matched normal DNA methylation profile more tumour-like than benign. For patient 1, the hypermethylation threshold (0.747) was lower than that for the other patients (0.774, 0.766 and 0.814 for patients 2, 3 and 4 respectively) suggesting more hypermethylated CpG sites in patient 1's benign sample type than in the other patients' samples. A similar pattern was observed in the methylation thresholds inferred from the patients' tumour sample types (see Appendix S11 in S1 File) in that the threshold was lower for patient 1.

The ARIs of 0.94 and 0.96 between the KN· and K·· solutions for the benign and tumour sample types respectively indicate good clustering agreement. Summaries of parameter estimates and the kernel density estimates under the KN· model are available in Appendices S4 and S12 in S1 File.

Identifying DMCs in the PCa data. To identify differentially methylated CpG sites in the PCa data, the K·R model was fitted to the $C \times NR$ dimensional dataset. Similar to the simulation study, the AIC, BIC and ICL were non-informative and consistently decreased across models with $K = 2, \ldots, 30$ (see Appendix S13 in S1 File). Thus, motivated by the $G^R = 9$ unique methylation state combinations that could be biologically present across the benign and tumor sample types, a model with $K = G^R = 9$ was fitted to the PCa data.

Fig 5 illustrates fitted density estimates of the clustering solution, ordered by AUC or WD in the case of equal AUC values. Kernel density estimates are in Appendix S14 in S1 File. Table 2 summarises the parameter estimates and details the AUC and WD metrics for each cluster. As the extent of disease progression varied across patients, the PCa data were noisier than the simulated data, and the AUC and WD metrics were lower in general. Both the AUC and WD metrics suggest that clusters 1 and 2 contain the CpG sites that are most differentially methylated in nature. Inspection of the density and parameter estimates of clusters 1 and 2 provides insight: cluster 1 captures CpG sites exhibiting a downward trend in methylation values for tumor samples with increased methylation values in the benign samples. On the other hand, CpG sites in cluster 2 tend to have higher methylation levels in tumor samples than in benign samples. While there are visual differences between the benign and tumor density estimates in clusters 3–5, the density estimates in later clusters are almost visually indistinguishable between the two samples, particularly in the case of clusters 6, 7 and 9. This is intuitive as clusters with smaller AUC (and WD) values contain the least differentially methylated CpG sites between the benign and tumour samples and their respective density estimates within such clusters will be very similar. Fig 6 shows the empirical cumulative distribution functions (ECDFs) for the DMCs within clusters 1 and 2, for both benign and tumor sample types. The ECDF for cluster 1 shows an increase in *beta* values within the benign samples, relative to the tumor samples while the ECDF of the CpG sites in cluster 2 indicates an elevation in *beta* values in the tumor samples, compared to the benign samples. The K·R model identifies 102,757 CpG sites, belonging to clusters 1 and 2, as being mostly differentially methylated.

On performing gene ontology analysis [49], CpG sites in cluster 2 were found to be related to known genes e.g., RARB, GSTP1, RASSF1, SFRP2, which are implicated in prostate cancer. For example, hypermethylation of RARB promoter genes is a significant biomarker in diagnosing prostate cancer [50]. The methylation levels of the DMCs in cluster 2 that belong to the RARB genes suggest the median *beta* value is higher in the tumour sample type than in the benign sample type for all patients (see Appendix S15 in S1 File). Through non-parametric

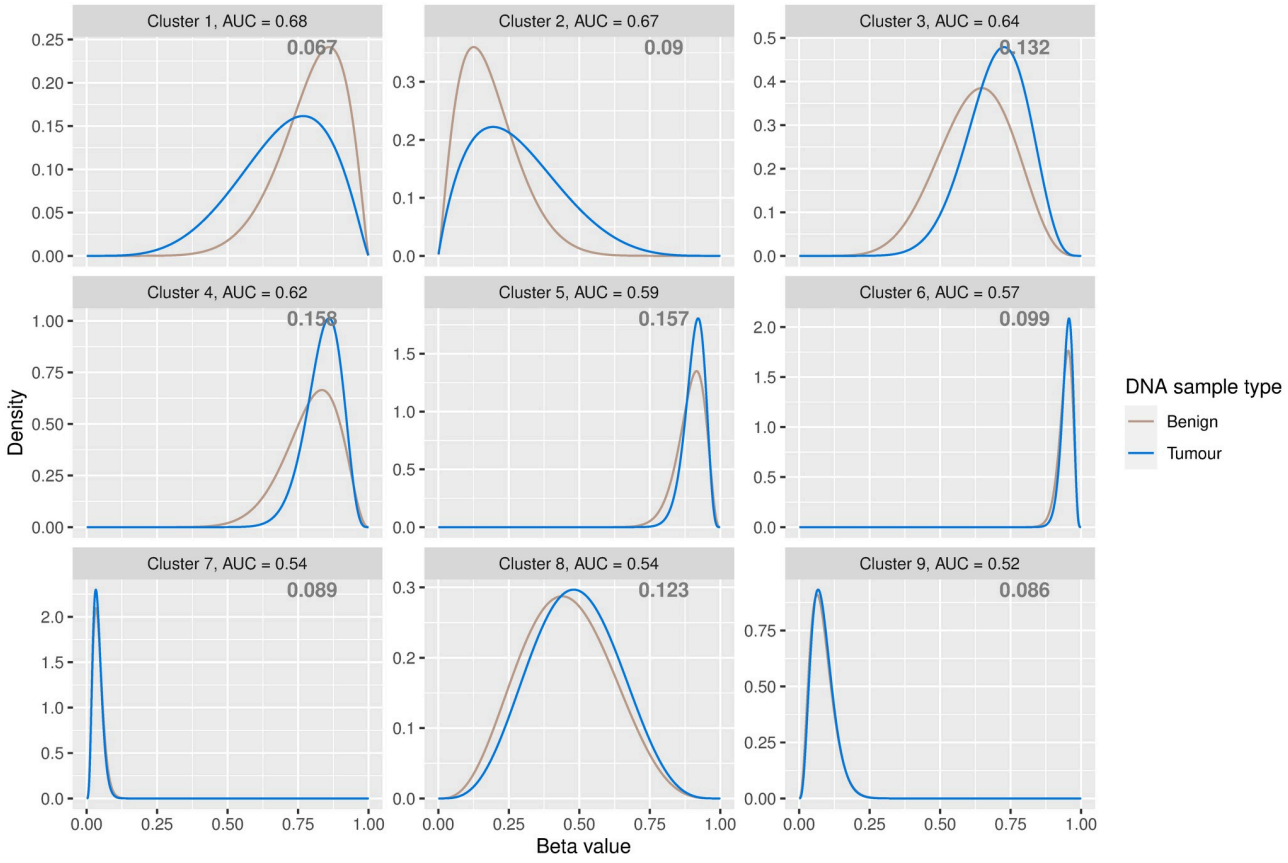

Fig 5. Fitted density estimates under the clustering solution of the K·R model. The model estimates parameters for $K = 9$ clusters for the DNA methylation data from benign and tumour prostate sample types. The estimated mixing proportions are displayed in the relevant panel.

https://doi.org/10.1371/journal.pone.0314014.g005

tests, the *beta* values were shown to be significantly higher in the tumour samples than in the benign samples for the CpG sites related to these genes ($p < 0.05$). Further, the ECDF for DMCs related to the RARB genes for benign and tumour sample types illustrated that the DMCs have increased *beta* values in the tumour samples compared to the benign samples (see

Table 2. Beta distributions' parameter estimates for the benign and tumour samples, and the AUC and WD metrics, for the PCa data under the K·R model.

	Benign				Tumour					
Cluster	$\hat{\alpha}$	$\hat{\delta}$	Mean	S.D.	$\hat{\alpha}$	$\hat{\delta}$	Mean	S.D.	AUC	WD
1	8.815	2.277	0.795	0.116	5.076	2.231	0.695	0.160	0.683	0.100
2	2.324	10.223	0.185	0.106	1.975	5.040	0.282	0.159	0.667	0.096
3	8.005	4.810	0.625	0.130	12.058	5.170	0.700	0.107	0.639	0.075
4	13.006	3.387	0.793	0.097	27.000	5.249	0.837	0.064	0.624	0.044
5	33.720	4.006	0.894	0.050	56.506	5.734	0.908	0.036	0.593	0.015
6	84.926	5.023	0.944	0.024	111.727	5.978	0.949	0.020	0.572	0.005
7	4.842	112.897	0.041	0.018	5.455	133.043	0.039	0.016	0.542	0.002
8	4.071	4.924	0.453	0.157	4.686	4.990	0.484	0.153	0.537	0.032
9	3.749	41.317	0.083	0.041	4.194	45.197	0.085	0.039	0.523	0.002

* Standard deviation is denoted as S.D.

https://doi.org/10.1371/journal.pone.0314014.t002

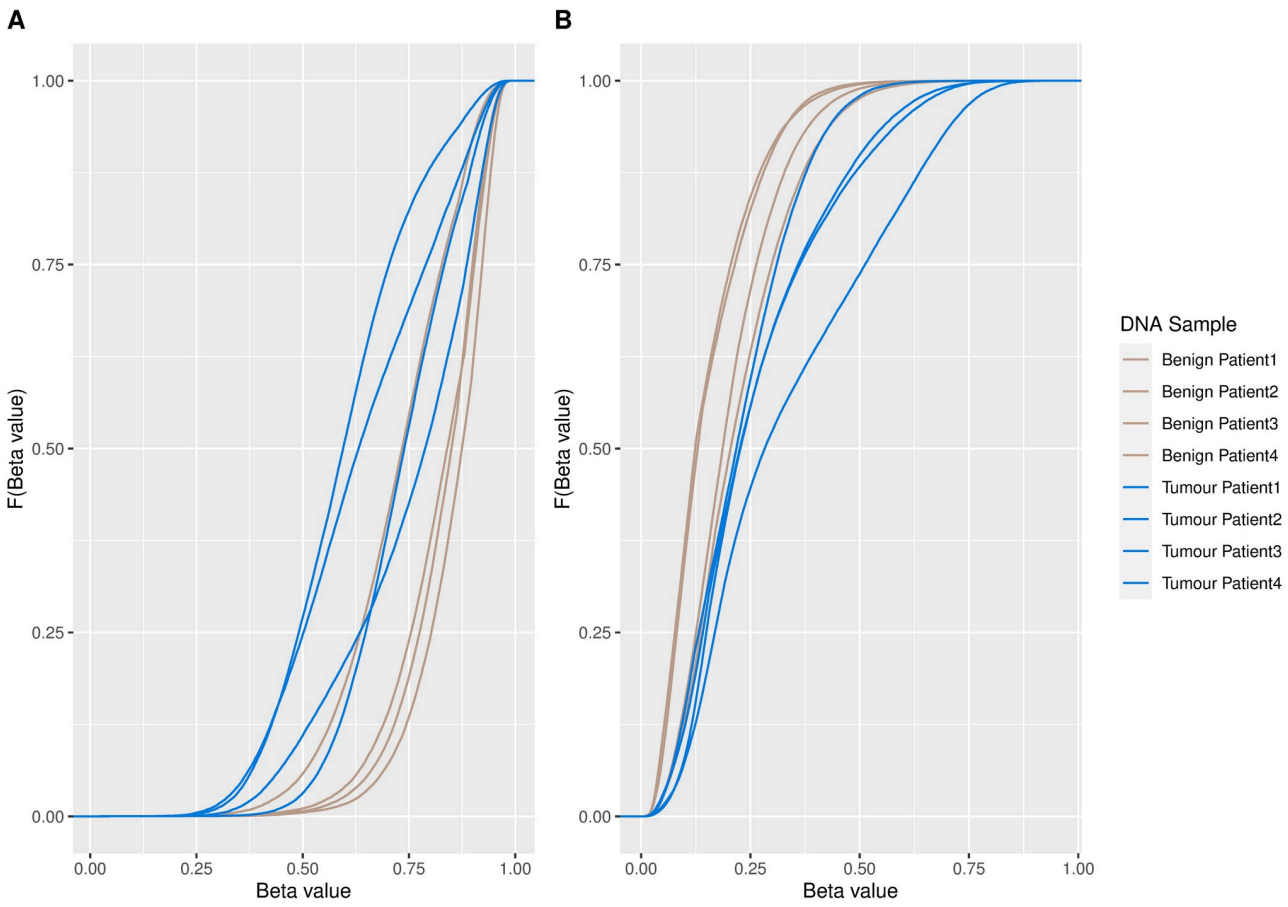

Fig 6. Empirical cumulative distribution functions for DMCs. Empirical cumulative distribution functions for DMCs in (A) cluster 1 and (B) cluster 2 for all patients and sample types.

https://doi.org/10.1371/journal.pone.0314014.g006

Appendix S16 in S1 File). Analysis of genes linked to CpG sites in clusters 3–9, the less differentially methylated clusters, did identify some genes previously implicated in prostate cancer e.g., AKT1 [51]. However, non-parametric tests also suggested no statistically significant difference ($p > 0.05$) in *beta* values between benign and tumour samples for CpG sites in clusters 3–9 linked to the gene AKT1, as did associated box and ECDF plots (see Appendix S17 in S1 File).

Gene ontology analysis of the DMCs in cluster 1 also unveiled approximately 16 noteworthy biological processes. These processes, distinct from cancer-related pathways, encompass vital functions such as nervous system processes. A substantial count of 1001 significant biological processes were revealed among the DMCs in cluster 2 (FDR-adjusted p-value <0.05). Further, considering the KEGG pathways, the DMCs in cluster 1 were associated with one significant pathway, while the DMCs within cluster 2 exhibited involvement in a noteworthy 61 significant pathways. Of these significant pathways, many were cancer related e.g., the proteoglycans in cancer pathway was the second most enriched pathway.

Given the BMM's model-based approach to clustering, the uncertainty in CpG site c's clustering is available as $1 - \max_{k=1,\ldots,K} (\hat{z}_{ck})$, with a maximum possible uncertainty of $1 - 1/K = 8/9$. All CpG sites have clustering uncertainties well below this maximum, demonstrating that the CpG sites are clustered with high certainty (see Appendix S18 in S1 File).

To demonstrate the general applicability of the approach, the BMMs were also fitted to a publicly available DNA methylation dataset from an esophageal squamous cell carcinoma study (ESCC); full details are available in Appendix S19 in S2 File.

Discussion

DNA methylation is widely studied for disease diagnosis and treatment. Technology advancements have led to the development of microarrays that can assay e.g., 850,000 CpG sites from a DNA sample [12], but the analysis of these large arrays has been limited by a lack of appropriate statistical methods for the bounded and heteroskedastic nature of *beta*-valued DNA methylation data. The methylation states of CpG sites are often of interest and are typically identified using thresholds which are defined in the literature based on intuition [14] rather than using an objective approach. Additionally, to detect DMCs, it is common practice to apply a logit transformation to *beta* values, and subsequently model them as Gaussian-distributed [13, 28]. Alternatively, comparisons between untransformed methylation levels among sample types are often conducted using multiple moderated t-tests or Wilcoxon rank sum tests [29, 30]. The approach proposed here advocates against transforming the data and instead proposes modelling the data in its innate form when inferring methylation state thresholds and DMCs.

In the context of prostate cancer, a family of beta mixture models is proposed which employs novel constraints on the model parameters to cluster CpG sites based on untransformed *beta* values to objectively identify methylation state thresholds and DMCs between benign and tumour samples. The BMMs use a model-based clustering approach and inference is computationally efficient through the use of a digamma approximation. The objective inference of methylation thresholds demonstrated that the thresholds of 0.2 and 0.8 or 0.3 and 0.7 defined in literature are not appropriate for every scenario. The thresholds inferred from each patient's data showed variability, reflecting the different stages of disease among patients. The proposed K·R model clusters CpG sites from multiple DNA sample types to determine the CpG sites with differential methylation. Gene ontology enrichment analysis of the genes associated with CpG sites in the most differentially methylated clusters revealed several significant biological processes, cancer-related pathways and genes implicated in prostate cancer, opening new avenues of research. The results illustrate the ability of the BMMs to analyse large microarrays consisting of samples from multiple conditions from several patients and to reveal biologically relevant methylation patterns, thus contributing to advances in the field of quantitative DNA methylation analysis.

In terms of the family of BMMs developed here, there are several potential future research directions. For example, while DNA methylation can be influenced by environmental and clinical variables, the proposed BMMs do not incorporate such covariates. However, the BMMs could be extended, for example using a mixture of experts approach [52] where the parameters of the BMM are modeled as functions of the covariates, to offer a richer modelling framework. Further, a key assumption of the proposed BMMs is that the methylation states of adjacent CpG sites are conditionally independent given their cluster membership. However, methylation levels of adjacent CpG sites are often highly correlated [53]. This phenomenon gives rise to the emergence of biologically meaningful regions with discernible patterns. Expanding the scope of the BMM family to encompass the spatial dependencies within the data would present an opportunity to incorporate these structural nuances and ultimately facilitate the identification of particularly relevant differentially methylated regions. While the scale of missing data in the prostate cancer data considered here was almost negligible, it could be more prevalent in other settings. Such cases would motivate the development of imputation approaches that

are cognisant of the heterogeneity typical of DNA methylation datasets. Finally, the methylation state of a human genome changes over time depending on clinical conditions. Longitudinal methylation data are often collected to study the effect of environmental changes or treatments on disease progression. Such data are vast and current approaches struggle to handle these extensive data in their innate form. In order to analyze methylation changes over time in multiple patients, similar to [54], the BMMs could be further enhanced to model dependency over time.

Supporting information

S1 File. Supporting information 1 for 'A novel family of beta mixture models for the differential analysis of DNA methylation data: An application to prostate cancer' data by Majumdar et al. Supporting Information (Appendices S1 to S18) includes BMM derivations, parameter estimates and density estimate plots when BMM is applied to simulated and prostate cancer datasets.
(PDF)

S2 File. Supporting information 2 for 'A novel family of beta mixture models for the differential analysis of DNA methylation data: An application to prostate cancer' data by Majumdar et al. Supporting Information (Appendix S19) contains analysis of BMMs applied to the additional esophageal squamous cell carcinoma dataset.
(PDF)

Acknowledgments

The authors wish to thank members of the Working Group in Model-based Clustering for their discussions on this work.

Author Contributions

Conceptualization: Koyel Majumdar, Romina Silva, Antoinette Sabrina Perry, Thomas Brendan Murphy, Isobel Claire Gormley.

Data curation: Romina Silva, Antoinette Sabrina Perry.

Formal analysis: Koyel Majumdar, Romina Silva, Antoinette Sabrina Perry, Thomas Brendan Murphy, Isobel Claire Gormley.

Investigation: Romina Silva, Antoinette Sabrina Perry.

Methodology: Koyel Majumdar, Andrea Rau, Florence Jaffrezic, Thomas Brendan Murphy, Isobel Claire Gormley.

Software: Koyel Majumdar, Thomas Brendan Murphy, Isobel Claire Gormley.

Validation: Koyel Majumdar, Antoinette Sabrina Perry, Ronald William Watson, Andrea Rau, Florence Jaffrezic, Thomas Brendan Murphy, Isobel Claire Gormley.

Visualization: Koyel Majumdar, Thomas Brendan Murphy, Isobel Claire Gormley.

Writing – original draft: Koyel Majumdar, Thomas Brendan Murphy, Isobel Claire Gormley.

Writing – review & editing: Koyel Majumdar, Romina Silva, Antoinette Sabrina Perry, Ronald William Watson, Andrea Rau, Florence Jaffrezic, Thomas Brendan Murphy, Isobel Claire Gormley.

References

1. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes & Development. 2009 Apr 1; 23: 781–783. https://doi.org/10.1101/gad.1787609 PMID: 19339683
2. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013 Jan; 38: 23–38. https://doi.org/10.1038/npp.2012.112 PMID: 22781841
3. Jin Z, Liu Y. DNA methylation in human diseases. Genes & Diseases. 2018; 5(1): 1–8. https://doi.org/10.1016/j.gendis.2018.01.002 PMID: 30258928
4. Das PM, Singal R. DNA methylation and cancer. Journal of Clinical Oncology. 2004 Nov 15; 22(22): 4632–4642. https://doi.org/10.1200/JCO.2004.07.151 PMID: 15542813
5. Bird A. DNA methylation patterns and epigenetic memory. Genes & Development 2002 Jan 1; 16: 6–21. https://doi.org/10.1101/gad.947102 PMID: 11782440
6. de Almeida BP, Apolónio JD, Binnie A, Castelo-Branco P. Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers. BMC Cancer. 2019 Mar; 19: 219. https://doi.org/10.1186/s12885-019-5403-0 PMID: 30866861
7. Kim JH, Dhanasekaran SM, Prensner JR, Cao X, Robinson D, Kalyana-Sundaram S, et al. Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Research. 2011 Jul 1; 21: 1028–1041. https://doi.org/10.1101/gr.119347.110 PMID: 21724842
8. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2021 May; 71(3): 209–249. PMID: 33538338
9. Li LC, Okino ST, Dahiya R. DNA methylation in prostate cancer. Biochimica et Biophysica Acta (BBA)—Reviews on Cancer. 2004 Sep 20; 1704(2): 87–102. https://doi.org/10.1016/j.bbcan.2004.06.001 PMID: 15363862
10. Daniunaite K, Jarmalaite S, Kalinauskaite N, Petroska D, Laurinavicius A, Lazutka JR, et al. Prognostic value of RASSF1 promoter methylation in prostate cancer. The Journal of Urology. 2014 Dec 1; 192(6): 1849–1855. https://doi.org/10.1016/j.juro.2014.06.075 PMID: 24980613
11. Moritz R, Ellinger J, Nuhn P, Haese A, Müeller SC, Graefen M, et al. DNA hypermethylation as a predictor of PSA recurrence in patients with low-and intermediate-grade prostate cancer. Anticancer Research. 2013 Dec 1; 33(12):5249–5254. PMID: 24324057
12. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology. 2016 Oct; 17: 208. https://doi.org/10.1186/s13059-016-1066-1 PMID: 27717381
13. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010 Nov; 11: 587. https://doi.org/10.1186/1471-2105-11-587 PMID: 21118553
14. Chen X, Zhang Q, Chekouo T. Filtering high-dimensional methylation marks with extremely small sample size: an application to gastric cancer data. Frontiers in Genetics. 2021 Jul 12; 12: 705708. https://doi.org/10.3389/fgene.2021.705708 PMID: 34322159
15. Men C, Chai H, Song X, Li Y, Du H, Ren Q. Identification of DNA methylation associated gene signatures in endometrial cancer via integrated analysis of DNA methylation and gene expression systematically. Journal of Gynecologic Oncology. 2017 Sep; 28(6):e83. https://doi.org/10.3802/jgo.2017.28.e83 PMID: 29027401
16. Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics. 2004 Aug; 20(12): 1896–1904. https://doi.org/10.1093/bioinformatics/bth176 PMID: 15044245
17. Koestler DC, Christensen BC, Marsit CJ, Kelsey KT, Houseman EA. Recursively partitioned mixture model clustering of DNA methylation data using biologically informed correlation structures. Statistical Applications in Genetics and Molecular Biology. 2013 Mar 5; 12(2): 225–240. https://doi.org/10.1515/sagmb-2012-0068 PMID: 23468465
18. Ma Z, Teschendorff AE. A variational Bayes beta mixture model for feature selection in DNA methylation studies. Journal of Bioinformatics and Computational Biology. 2013 Aug 14; 11(4): 1350005. https://doi.org/10.1142/S0219720013500054 PMID: 23859269
19. Zhang L, Meng J, Liu H, Huang Y. A nonparametric Bayesian approach for clustering bisulfate-based DNA methylation profiles. BMC Genomics. 2012 Oct 26; 13 (Suppl 6): S20. https://doi.org/10.1186/1471-2164-13-S6-S20 PMID: 23134689
20. Gevaert O, Tibshirani R, and Plevritis SK. Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biology. 2015 Jan; 16: 17. https://doi.org/10.1186/s13059-014-0579-8 PMID: 25631659

21. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2024; https://www.R-project.org/.

22. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013 Jan 15; 29(2): 189–196. https://doi.org/10.1093/bioinformatics/bts680 PMID: 23175756

23. Laurila K, Oster B, Andersen CL, Lamy P, Orntoft T, Yli-Harja O, et al. A beta-mixture model for dimensionality reduction, sample classification and analysis. BMC Bioinformatics. 2011 May; 12: 215. https://doi.org/10.1186/1471-2105-12-215 PMID: 21619656

24. Schröder C, Rahmann S. A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. Algorithms for Molecular Biology. 2017 Aug; 12: 21. https://doi.org/10.1186/s13015-017-0112-1 PMID: 28828033

25. Shi M, Tsui SK, Wu H, Wei Y. Pan-cancer analysis of differential DNA methylation patterns. BMC Medical Genomics. 2020 Oct; 13 (Suppl 10): 154. https://doi.org/10.1186/s12920-020-00780-3 PMID: 33087120

26. Taguchi YH, Turki T. Principal component analysis-and tensor decomposition-based unsupervised feature extraction to select more suitable differentially methylated cytosines: Optimization of standard deviation versus state-of-the-art methods. Genomics. 2023 Mar; 115(2):110577. https://doi.org/10.1016/j.ygeno.2023.110577 PMID: 36804268

27. Wu D, Gu J, Zhang MQ. FastDMA: an infinium humanmethylation450 beadchip analyzer. PloS ONE. 2013 Sep 5; 8(9):e74275. https://doi.org/10.1371/journal.pone.0074275 PMID: 24040221

28. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015 Apr 20; 43(7): e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

29. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, et al. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics. 2012 Mar; 28(5): 729–730. https://doi.org/10.1093/bioinformatics/bts013 PMID: 22253290

30. Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD, et al. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Research. 2013 Jun 1; 41(11): e117. https://doi.org/10.1093/nar/gkt242 PMID: 23598999

31. Schildknecht K, Olek S, Dickhaus T. Simultaneous statistical inference for epigenetic data. PLoS ONE. 2015 May 12; 10(5): e0125587. https://doi.org/10.1371/journal.pone.0125587 PMID: 25965389

32. Chen Z, Huang H, Liu Q. Detecting differentially methylated loci for multiple treatments based on high-throughput methylation data. BMC Bioinformatics. 2014 May; 15: 142. https://doi.org/10.1186/1471-2105-15-142 PMID: 24884464

33. Wang Z, Wu X, Wang Y. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC Bioinformatics. 2018 Apr 11; 19:115. https://doi.org/10.1186/s12859-018-2096-3 PMID: 29671397

34. Scrucca L. A transformation-based approach to Gaussian mixture density estimation for bounded data. Biometrical Journal. 2019 Jul; 61(4): 873–888. https://doi.org/10.1002/bimj.201800174 PMID: 30983031

35. Azam M, Bouguila N. Multivariate bounded support Laplace mixture model. Soft Computing. 2020 Sep; 24: 13239–13268. https://doi.org/10.1007/s00500-020-04737-7

36. Azam M, Bouguila N. Multivariate bounded support asymmetric generalized Gaussian mixture model with model selection using minimum message length. Expert Systems with Applications. 2022 Oct 15; 204:117516. https://doi.org/10.1016/j.eswa.2022.117516

37. Silva R, Moran B, Russell NM, Fahey C, Vlajnic T, Manecksha RP, et al. Evaluating liquid biopsies for methylomic profiling of prostate cancer. Epigenetics. 2020 Jan; 15(6-7): 715–727. https://doi.org/10.1080/15592294.2020.1712876 PMID: 32000564

38. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016; 8(3): 389–399. https://doi.org/10.2217/epi.15.114 PMID: 26673039

39. Di Lena P, Sala C, Prodi A, Nardini C. Missing value estimation methods for DNA methylation data. Bioinformatics. 2019 Oct; 35(19): 3786–3793. https://doi.org/10.1093/bioinformatics/btz134 PMID: 30796811

40. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological). 1977; 39(1): 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

41. Nocedal J, Wright SJ. Quasi-Newton methods. Numerical Optimization. Springer Series in Operations Research and Financial Engineering. 2006; Chapter 6: 136–163.

42. Berndt EK, Hall BH, Hall RE, Hausman JA. Estimation and inference in non-linear structural models. Annals of Economic and Social Measurement. 1974; 3(4): 653–665.

43. Diamond HG, Straub A. Bounds for the logarithm of the Euler gamma function and its derivatives. Journal of Mathematical Analysis and Applications. 2016 Jan 15; 433(2): 1072–1083. https://doi.org/10.1016/j.jmaa.2015.08.034

44. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E., Tanabe K., Kitagawa G. (eds) Selected papers of Hirotugu Akaike. New York, NY: Springer New York. 1998; 199–213.

45. Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978 Mar; 6(2): 461–464. https://doi.org/10.1214/aos/1176344136

46. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000 Jul; 22(7): 719–725. https://doi.org/10.1109/34.865189

47. Ponti A, Giordani I, Mistri M, Candelieri A, Archetti F. The "Unreasonable" Effectiveness of the Wasserstein Distance in Analyzing Key Performance Indicators of a Network of Stores. Big Data and Cognitive Computing. 2022 Nov 15; 6(4): 138. https://doi.org/10.3390/bdcc6040138

48. Hubert L, Arabie P. Comparing Partitions. Journal of Classification. 1985; 2: 193–218. https://doi.org/10.1007/BF01908075

49. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. Genome Biology. 2021 Jun; 22: 173. https://doi.org/10.1186/s13059-021-02388-x PMID: 34103055

50. Ameri A, Alidoosti A, Hosseini Y, Parvin M, Emranpour MH, Taslimi F, et al. Prognostic value of promoter hypermethylation of Retinoic Acid Receptor Beta (RARB) and CDKN2 (p16/MTS1) in prostate cancer. Chinese Journal of Cancer Research. 2011 Dec; 23: 306–311. https://doi.org/10.1007/s11670-011-0306-x PMID: 23358881

51. Herberts C, Murtha AJ, Fu S, Wang G, Schönlau E, Xue H, et al. Activating AKT1 and PIK3CA mutations in metastatic castration-resistant prostate cancer. European Urology. 2020 Dec; 78(6): 834–844. https://doi.org/10.1016/j.eururo.2020.04.058 PMID: 32451180

52. Gormley IC, Frühwirth-Schnatter S. Mixture of experts models. In Handbook of mixture analysis. 2018 Dec; 1st ed.: 271–307.

53. Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, et al. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Research. 2009 Sep 1; 19(9): 1593–1605. https://doi.org/10.1101/gr.095190.109 PMID: 19581485

54. Nyamundanda G, Gormley IC, Brennan L. A dynamic Probabilistic principal components model for the analysis of longitudinal metabolomics data. Journal of the Royal Statistical Society, Series C (Applied Statistics). 2014 Nov; 63(5): 763–782. https://doi.org/10.1111/rssc.12060