RESEARCH

betaclust: a family of mixture models for beta valued DNA methylation data

Koyel Majumdar[1], Romina Silva[2,3], Antoinette Sabrina Perry[2,3], Ronald William Watson[3], Thomas Brendan Murphy[1] and Isobel Claire Gormley[1*]

Abstract

Background: The DNA methylation process has been extensively studied for its role in cancer. Hypermethylation of promoter genes has been shown to silence the tumour suppressor genes: the cytosine-guanine dinucleotide (CpG) sites which remain unmethylated in normal cells become methylated in cancer cells. The methylation state of a CpG site is hypermethylated if both the alleles are methylated, hypomethylated if neither of the alleles are methylated and hemimethylated otherwise. Identifying the differentially methylated CpG (DMC) sites between benign and tumour samples can help understand the disease and its treatment.

There is a lack of suitable methods for modelling the methylation values, known as beta values, in their innate form. For this reason, the beta values are usually transformed into M-values for analysis. Typically, arbitrary thresholds are selected and used to identify the methylation state of a CpG site and based on this, the DMCs are identified between different samples. We propose a family of novel beta mixture models (BMMs) to (i) appositely model the innate beta valued data, (ii) objectively identify methylation state thresholds and (iii) identify the DMCs using a model-based clustering approach. The family of BMMs employs different parameter constraints applicable to different study settings. Parameter estimation proceeds via an EM algorithm, with a novel approximation during the M-step providing tractability and ensuring computational feasibility.

Results: The BMMs are used to analyse a simulated dataset, a prostate cancer dataset and an esophageal squamous cell carcinoma dataset. Our approach objectively identifies methylation state thresholds and identifies more DMCs between the benign and tumour samples than conventional methods in a computationally efficient manner. The identified DMCs have been shown to be linked to genes known to be implicated in prostate cancer carcinogenesis. The empirical cumulative distribution function of the DMCs indicates that they have higher methylation values in the tumour tissue than in the benign tissue.

Conclusion: An R package `betaclust` is developed to facilitate the widespread use of the developed BMMs and efficiently cluster the DNA methylation data in its innate form to identify the DMCs, eliminating the need for subjectively choosing thesholds for determining a CpG site's methylation state.

Keywords: DNA methylation; model-based clustering; beta mixture model; digamma function; EM algorithm.

Background

Epigenetics is the study of heritable changes in gene activity that do not involve any explicit change to the DNA sequence [1]. DNA methylation is one of the epigenetic processes where a methyl group is added to or removed from the 5′ carbon of the cytosine ring [2]. This process assist 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]. The DNA methylation process has been extensively studied in the context of cancer, and its treatment [4, 5]. It has been observed that cytosine-guanine dinucleotide (CpG) islands that remain unmethylated in normal cells can become methylated in abnormal cells such as cancer cells [6]. The methylation state of a CpG locus is said to be hypermethylated if both the alleles are methylated, hypomethylated if neither of the alleles is methylated and hemimethylated otherwise. A differentially methylated region (DMR) is a genomic region that has different methylation states between different sam-

*Correspondence: claire.gormley@ucd.ie
[1]School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland
Full list of author information is available at the end of the article

ples, which may have been taken from tissues of an individual over time, different tissues from the same individuals or other individuals [7]. In several cancer studies, it has been observed that tumour suppressor genes are silenced by hypermethylation of promoter genes [8–10]. Identifying DMRs between the benign and tumour samples can help understand disease and its treatment.

The Illumina MethylationEPIC BeadChip microarray [11] 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 (M) and unmethylated (U) light signal intensities, and the level of methylation is measured as a ratio of these intensities. The *beta* value is used to quantify the methylation level and is calculated as $beta = \max(M)/(\max(M) + \max(U) + \chi)$, where χ is a constant offset added to regularise the values for very low M and U values [12]. The methylation level at a CpG site is quantified by this *beta* value and is constrained to lie between 0 and 1 as it measures the proportion of methylated and unmethylated signal intensities. The *beta* values are continuous, and a value close to 1 suggests that the 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 the two endpoints. This results in the data being heteroscedastic, which imposes serious challenges for analysing them as they violate the assumptions for the ubiquity of Gaussian models. Hence, the *beta* values are usually converted to M-values using a logit transformation, $M = \log(beta/(1-beta))$ as these values are statistically more convenient. For the M-values, a positive value represents hypermethylation, whereas a negative value represents hypomethylation of a CpG site. The transformed data can be modelled using Gaussian models as the data are no longer bounded and lie within $(-\infty, \infty)$ [12]. However, transformation make the results less biologically interpretable. Hence there is a need to statistically model the *beta* values in their innate form.

In many studies, thresholds of *beta* values are subjectively selected to identify the three methylation states [13]. A *beta* value < 0.2 is often used to suggest hypomethylation of a CpG site whereas a value > 0.8 is used to suggest hypermethylation. This is because the *beta* values within the interval [0.2,0.8] are approximately linearly related to the M-values [12]. Several studies model a small set of CpG sites and apply unsupervised machine learning techniques to cluster the different sample types into latent groups. A Bernoulli-log normal mixture model [14] was proposed to identify two subgroups of small cell and non-small cell lung cancer by analyzing the DNA methylation states of 7 CpG sites from 87 lung cancer cell lines. Another study modelled the DNA methylation states using beta mixture models using a recursive partitioning algorithm to choose the optimal number of clusters of samples: this recursive-partitioning mixture model (RPMM) [15] was used to identify the association of the samples to the different types of tissue, including bladder, blood, brain, kidney and others. A further improvement of the RPMM model was developed by using Gaussian and beta distributions and by incorporating the knowledge of correlation between neighbouring CpG sites [16]. This approach helped identify the correlation between CpG sites on one gene with another gene and identify biologically meaningful clusters in a dataset comprising of case/control cancer cells. Epiclomal [17] is an R and Python package that has been developed to cluster sparse single-cell DNA methylation data using a probabilistic clustering method and also impute missing values. This method uses a hierarchical mixture model to cluster the DNA methylation data from different cells into an optimal number of clusters. A variational Bayes beta mixture model [18] and a non-parametric beta mixture model [19] were developed to identify latent groups of biologically related samples by analyzing the methylation states of the samples. As a small number of CpG sites related to the genes of interest are selected for these studies, it is difficult to computationally scale these models to the complete microarray. There are model-based clustering approaches to cluster CpG sites in their innate *beta* form to identify DMCs. The Methylmix [20] R package clusters genes into known methylation states using a univariate beta mixture model. The univariate nature of this model limits it from considering related samples from a cohort of similarly diseased patients. Most studies use M-values for identifying DMCs, but some methods use *beta* values for their models. These models use multiple t-tests or Wilcoxon rank-sum tests to identify the DMCs [21, 22].

We propose a novel family of beta mixture models (BMM) and use a model-based clustering approach to (i) identify methylation states of each CpG site in a sample, (ii) objectively infer thresholds and (iii) identify the DMCs between different samples. The DMRs of interest can then be retrieved from the identified DMCs. The models developed are computationally efficient and are capable of analysing the entire microarray. The BMMs are applied to 20 simulated datasets, a prostate cancer dataset and an esophageal squamous

cell carcinoma dataset demonstrating the capability to model the *beta* values and efficiently cluster the DNA methylation datasets to identify DMCs between different samples while objectively identifying thresholds.

Prostate cancer data

Prostate cancer (PCa) is the second most commonly diagnosed cancer in men and is the second major reason of mortality, globally [23]. Prostate specific antigen (PSA) is the most widely used biomarker for cancer detection in the early stages and the presence of cancer tissues is detected based on elevated levels of PSA. However, elevated levels of PSA have been observed in men without cancer and to confirm the occurrence of cancer, a biopsy of prostate tissue is conducted [24]. Triggering of prostate cancer and disease advancement is related to epigenetic changes such as hypermethylation of target genes resulting in gene inactivation. Hypermethylation of certain tumour suppressor genes are observed during the early stages of the disease [25]. The identification of methylation changes in these target genes can help in early cancer diagnosis. The change in methylation pattern during the treatment can assist in understanding the effectiveness of the treatment.

DNA methylation samples were collected for a study of methylation profiling [26]. The study cohort comprised of 4 patients with high-grade prostate cancer disease. Two tissue samples were collected from each patient, from a benign prostate tissue biopsy and a tumour prostate tissue biopsy; DNA samples were extracted from these tissues. Methylation profiling of these DNA samples were done using the Infinium MethylationEpic Beadchip [27]. The methylation array data consisted of 694,923 CpG sites and *beta* values for each CpG site for the 2 DNA samples collected from each of the 4 patients. The raw methylation array data was quality controlled and pre-processed using RnBeads [28] R package [26]. Further, 103 CpG sites with missing *beta* values were removed from the dataset. The resulting dataset had $R = 2$ DNA samples collected from each of $N = 4$ patients and each sample contained *beta* values for $C = 694{,}820$ CpG sites. Here the aim is to detect DMCs, which we do by modelling the untransformed data and objectively infer thresholds to identify the three methylation states in DNA methylation samples and consequently uncover DMCs between the benign and tumour samples.

Method

Beta distributions

The beta distribution has support on the interval $[0, 1]$ and is parameterized by two positive shape parameters, α and δ. If $\alpha > 1$ and $\delta > 1$ then the distribution is unimodal and if the parameters are < 1 then the distribution is bimodal. A uniform distribution is obtained if $\alpha = 1$ and $\delta = 1$.

The methylation level x_{cnr} for each of $c = (1, ..., C)$ CpG sites from r^{th} DNA sample collected from n^{th} patient is considered and probability density function (pdf),

$$f(x_{cnr}, \alpha, \delta) = Beta(x_{cnr}; \alpha, \delta) = \frac{{x_{cnr}}^{\alpha-1}(1 - x_{cnr})^{\delta-1}}{B(\alpha, \delta)}, 0 \leq x_{cnr} \leq 1$$

where $B(\cdot)$ is the beta function, defined in terms of the gamma function ($\Gamma(\cdot)$):

$$B(\alpha, \delta) = \int_0^1 z^{\alpha-1}(1 - z)^{\delta-1}\, dz = \frac{\Gamma(\alpha)\Gamma(\delta)}{\Gamma(\alpha + \delta)}.$$

A dataset $\mathbf{X}$=$\{\boldsymbol{x}_1,...,\boldsymbol{x}_{NR}\}$ consisting of C beta values from R independent and identically distributed (i.i.d.) DNA samples collected from each of N i.i.d. patients is considered. In this dataset, the data recorded on each patient n's sample is a vector of length C, such that $\boldsymbol{x}_n = (x_{n1}, ..., x_{nC})$, resulting in the multivariate dataset $\mathbf{X}$ of dimension $C \times NR$. Assuming independence across each of the N patients' R samples, the density for $\boldsymbol{X}$ is,

$$f(\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\delta}) = \prod_{n=1}^{N} \prod_{r=1}^{R} Beta(x_{cnr}; \alpha_{nr}, \delta_{nr}),$$

where x_{cnr} defines the methylation level of CpG site c, on the r^{th} sample of patient n.

A beta mixture model

A mixture model is a statistical model where the observed data are assumed to have been generated from a heterogeneous population. Each CpG site is assumed to have been generated by one of the K groups or clusters in the heterogeneous population using a probabilistic model. The parameter $\boldsymbol{\theta}$ is used to denote all the shape parameters in the mixture model, i.e. $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1, \boldsymbol{\delta}_1, ..., \boldsymbol{\alpha}_k, \boldsymbol{\delta}_k)$, where α_k and δ_k are the shape parameters of cluster K. The mixing proportions $\boldsymbol{\tau}$ =$(\tau_1,...,\tau_K)$ lie between 0 and 1, $\sum_{k=1}^{K} \tau_k = 1$ and represent the probability of belonging to cluster k $\forall k = 1, ..., K$. The probability density function for this beta mixture model (BMM) is,

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

The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ of the beta mixture model can be estimated by maximising the log-likelihood function. The log-likelihood function for the mixture model with K clusters is defined as,

$$l(\boldsymbol{\theta}, \boldsymbol{\tau}; \boldsymbol{X}) = \sum_{c=1}^{C} \log \Bigg[\sum_{k=1}^{K} \tau_k \prod_{n=1}^{N} \prod_{r=1}^{R} Beta(x_{cnr}; \alpha_{knr}, \delta_{knr}) \Bigg]. \quad (1)$$

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

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

This complete data log-likelihood (2) is used to find the MLEs $\hat{\theta}$ and $\hat{\tau}$ using the Expectation-Maximisation (EM) algorithm [29]. Further, a probabilistic clustering solution is obtained via the expected value of $\boldsymbol{z}_c$, $\mathbf{E}[z_c]$, provided on convergence of the EM algorithm.

A family of BMMs

Each CpG site is known to have one of the $M = 3$ methylation states: hypomethylation, hemimethylation and hypermethylation. The most generalised BMM is defined in (2) which models the CpG sites as belonging to K latent groups. By introducing a variety of constraints on the parameters of the generalised BMM, a family of three beta mixture models is developed to cluster the CpG sites into the 3 methylation states while objectively inferring methylation state thresholds and identifying DMCs between different samples.

The K·· model The K·· model clusters each of the C CpG sites into one of $K = M$ methylation states, based on a single sample ($R = 1$) from each of N patients. Under the K·· model the shape parameters of each cluster are constrained to be equal for each patient. Thus, for the K·· model the complete data log-likelihood function is,

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

This model is used to (i) identify methylation states of each CpG site in a sample and (ii) objectively infer thresholds while eliminating the need for subjective thresholds.

The KN· model This model is similar to the K·· model and is used to cluster each of the C CpG sites into one of $K = M$ methylation states, based on data collected from a single sample ($R = 1$) from each of N patients. The KN· model differs from the K·· model as it is less parsimonious, allowing cluster and patient-specific shape parameters. The complete data log-likelihood function for this model is,

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

This model also aims to provide the solution to (i) identify methylation states of each CpG site in a sample and (ii) objectively infer thresholds between the 3 methylation states.

The K·R model This model allows identification of the differentially methylated CpG sites between the R DNA samples collected from each of the N patients. The model implores $K = M^R$ clusters each identifying a possible combination of the M methylation states in the R samples. The shape parameters vary for each sample type but are constrained to be equal for each patient. The complete data log-likelihood function for this model is,

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

This model enables the identification of DMCs between R DNA samples and seeks to address the 3$^{\text{rd}}$ part of the research question.

For each of these 3 BMMs, the parameters are estimated, and the cluster membership for each CpG site is inferred. We illustrate here how the parameters are estimated using the EM algorithm for the most general BMM (2). The computations for the parameter estimation for the K··, KN· and K·R models are illustrated in the Additional file: Appendix 1 — 3.

Parameter estimation

The Expectation-Maximisation (EM) algorithm [29] is used to obtain the maximum likelihood estimates of the parameters. The EM algorithm consists of two steps which are iterated until convergence. In the expectation step (E-step) the expected value of the complete data log-likelihood function is obtained, conditional on the observed data and the current parameter estimates. The maximisation step (M-step) maximises the expected complete data log-likelihood function with respect to the parameters. To obtain the parameter estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$, the E and M steps are iterated until convergence, to at least a local optimum,

$$\ell(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\tau}^{(t+1)}; \boldsymbol{X}) - \ell(\boldsymbol{\theta}^{(t)}, \boldsymbol{\tau}^{(t)}; \boldsymbol{X}) < \epsilon$$

where ϵ is an arbitrarily chosen small value.

An initial clustering solution is obtained using k-means and the method of moments is then used to calculate the initial values of α_{knr} and δ_{knr}. These initial values are provided as starting values to the EM algorithm. The E and M steps proceed as follows:

- E-step: The expected value of $\boldsymbol{z}_c$ is calculated, i.e. the posterior probability of $\boldsymbol{x}_c$ belonging to the k^{th} cluster, conditional on the current parameter estimates, i.e. at iteration $t+1$.

$$\begin{aligned}\hat{z}_{ck} =& \mathbf{E}[z_{ck}|\boldsymbol{X}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\tau}^{(t)}] \\ =& \frac{\tau_k^{(t)} \prod_{n=1}^{N} \prod_{r=1}^{R} Beta(x_{cnr}; \alpha_{knr}^{(t)}, \delta_{knr}^{(t)})}{\sum_{k=1}^{K} [\tau_k^{(t)} \prod_{n=1}^{N} \prod_{r=1}^{R} Beta(x_{cnr}; \alpha_{knr}^{(t)}, \delta_{knr}^{(t)})]}.\end{aligned}$$

- M-step: Estimates of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ are calculated given the $\hat{Z}$ values from the E-step. The expected complete data log-likelihood function is maximised by differentiating it w.r.t the parameters. Therefore, the mixing probabilities are calculated as,

$$\hat{\tau_k} = \sum_{c=1}^{C} \hat{z}_{ck}/C, \ \forall\, k = 1, ...K.$$

For the shape parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ a numerical optimisation approach is required in the M-step as closed form solutions are not available due to the presence of the gamma function in the beta density function.

The expected complete data log-likelihood function to be optimized is,

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

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

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

where ψ is the logarithmic derivative of the gamma function known as the digamma function,

$$\psi(\alpha_{knr}) = \frac{\partial \log \Gamma(\alpha_{knr})}{\partial \alpha_{knr}}.$$

Similarly, the derivative w.r.t δ_{knr} is,

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

The solutions $\hat{\alpha}_{knr}$ and $\hat{\delta}_{knr}$ are not available in closed form as they contain the digamma function. To obtain the maximised parameter estimates, quasi-newton numerical approximation algorithms like BFGS [30] and BHHH [31] could be used. For large datasets such as that considered here, these approximation methods are not computationally feasible.

A digamma approximation Numerical approximation approaches are used to approximate the digamma function to obtain closed form solutions for the shape parameters. The asymptotic series of the digamma function is,

$$\psi(y) = \log(y) + \frac{1}{2y} - \sum_{n=1}^{\infty} \frac{B_{2n}}{2n y^{2n}},$$

where y is the shape parameter in our case and B_{2n} are the Bernoulli numbers. The lower and upper bounds for the digamma function for all $y > 1/2$ [32] are

$$\psi(y) > \log(y - 1/2) \tag{6}$$

and

$$\psi(y) < \log(y - 1/2) + (y - 1/2)^{-2}/24.$$

A key assumption of our BMMs is that the beta distributions are unimodal hence the shape parameters are greater than 1 and these bounds are valid for the family of BMMs developed here. It was empirically observed that the lower bound approximation (6) is a very close approximation of the digamma function and so it is used in (4) and (5) to derive an approximation of the expected complete data log likelihood, allowing for closed-form solutions at the M-step of the EM algorithm. That is

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

and

$$\frac{\partial \ell_C}{\partial \delta_{knr}} \approx \sum_{c=1}^{C} \hat{z}_{ck} \sum_{n=1}^{N} \sum_{r=1}^{R} [\log(1 - x_{cnr}) - \log \frac{\delta_{knr} - 1/2}{\alpha_{knr} + \delta_{knr} - 1/2}]. \quad (8)$$

Equating equations (7) and (8) to zero, we get the approximate estimates of α_{knr} and δ_{knr} 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 $y_1 = (\sum_{c=1}^{C} z_{ck} \log x_{cnr})/(\sum_{c=1}^{C} z_{ck})$

and

$y_2 = (\sum_{c=1}^{C} z_{ck} \log(1 - x_{cnr}))/(\sum_{c=1}^{C} z_{ck})$.

As our BMMs are assumed to be unimodal, the models perform with shape parameters greater than 1. The models are rerun with new starting values if either of the shape parameters are calculated to be less than 1 by the EM algorithm.

Threshold calculation between the 3 methylation states The ratio of fitted density estimates ω is calculated for the clusters representing hypomethylation and hypermethylation. The C CpG sites in a DNA sample are clustered into $K = 3$ methylation states. The ω value for cluster k_1 is calculated as,

$$\omega = \frac{\tau_{k_1} f(\mathbf{X}; \boldsymbol{\alpha}_{k_1}, \boldsymbol{\delta}_{k_1})}{\sum_{k!=k_1} \tau_k f(\mathbf{X}; \boldsymbol{\alpha}_k, \boldsymbol{\delta}_k)}. \quad (9)$$

The threshold point dividing the hypomethylated and hemimethylated clusters is calculated as the minimum beta value at which ω for the hypomethylated cluster is >= 1. Similarly, the threshold point dividing the hemimethylated and hypermethylated clusters is calculated such that the maximum beta value at which the ω value for the hypermethylated cluster is >= 1.

As the parameters are constrained to be equal for each patient in the K·· model, a single pair of threshold points are calculated for each patient. Whereas the parameters vary for each patient type in the KN· model, hence a pair of threshold points are calculated for each patient.

Optimal model assessment

We compare the developed models by using the Akaike Information Criterion (AIC) [33], the Bayesian Information Criterion (BIC) [34], and integrated complete log-likelihood criterion (ICL) [35, 36]. The model which minimises the AIC, BIC and/or ICL is selected as the optimal model. The AIC and BIC values are defined as,

$$AIC = 2Q - 2\log(\hat{L})$$
$$BIC = Q\log(C) - 2\log(\hat{L}).$$

where, $\hat{L}$ is the maximised value of the likelihood function, Q is the number of parameters in the model, and C is the number of CpG sites in the dataset.

The ICL penalizes the BIC by including an entropy term favouring well separated clusters. The model with lowest ICL value is the optimal model. The ICL is defined as,

$$ICL = BIC + 2\sum_{c=1}^{C}\sum_{k=1}^{K} g_{ck}\log(z_{ck}),$$

where $g_{ck} = 1$ if the c^{th} CpG site belongs to the k^{th} cluster and 0 otherwise and $\hat{z}_{ck}$ is the estimated conditional probability that $\boldsymbol{x}_c$ belongs to the k^{th} cluster.

The adjusted Rand index (ARI) [37] is used for obtaining a measure of agreement between different clustering solutions. A contingency table is constructed for

the two different clustering solutions and the ARI is calculated. A value close to 1 suggests the two clustering solutions are in full agreement on the cluster membership of an observation.

Results

Simulation study

To explore the performance of the developed models 20 dataset were simulated containing beta values for each of the $C = 20,000$ CpG sites from $R = 2$ sample types (sample A and sample B) for each of $N = 4$ patients. The first objective of this simulation study was to cluster the CpG sites into the 3 true groups representing the 3 methylation states in sample A and identify the threshold points defining the methylation states. To achieve this, the K·· and KN· models are used to cluster the 20000 CpG sites into 3 clusters. The K·· model is selected as the optimal model with the lowest BIC score for each dataset. In Figure 1 the density estimates under the clustering solution of the K·· model for a single simulated dataset are displayed. CpG sites with a tendency to be hypomethylated are identified by cluster 1, while those with a tendency to be hypermethylated are placed into cluster 3, and CpG sites with a tendency to be hemimethylated are grouped into cluster 2. The proportion of CpG sites belonging to each cluster is also displayed in the graph. As parameters are constrained to be equal for each patient in the K·· model single pair of thresholds are calculated for all 4 patients. The threshold point of 0.242 indicates any CpG site having a beta value lower than this is likely to be hypomethylated. Similarly any CpG site having a beta value greater than the second threshold point of 0.808 is likely to be hypermethylated.

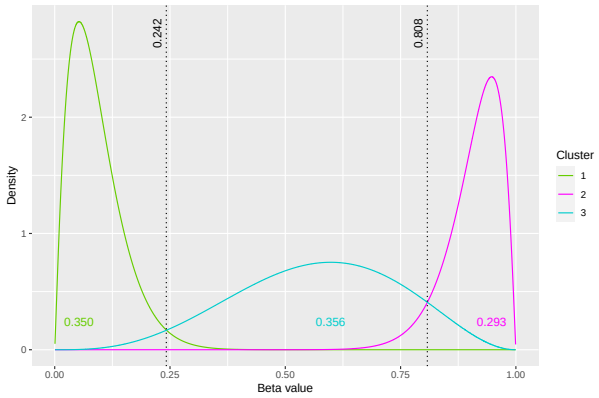

Figure 1: Fitted density estimates under the clustering solution of simulated data from sample A using the K·· model. The 3 clusters identifies the CpG sites in sample A for patient 1, that are likely to be hypomethylated, hemimethylated or hypermethylated. The thresholds displayed are calculated based on the parameters estimated. Similar graphs are obtained for all patients using the KN· model. The proportion of CpG sites belonging to each cluster is displayed.

Table 1: Contingency table for patients in sample A to cluster CpG sites into K=3 methylation states

Methylation state — Cluster	1	2	3
Hypermethylation		5866	12
Hemimethylation		8	7109
Hypomethylation	7004		1

The actual and observed membership is presented in the contingency Table 1 for a single simulated dataset. The ARI obtained based on this table showed 99.7% of agreement between the true groups and the observed clustering for both models. An agreement of 100% was obtained while comparing the clustering solutions between the two models. The mean ARI for the K·· model was obtained as 0.998 (with a standard deviation of 0.002) and the KN· model was calculated to be 0.998 (with a standard deviation of 0.002) for all 20 simulated datasets, exhibiting high accuracy in clustering a single sample into the 3 methylation states by both the models. The mean ARI for comparing the K·· and KN· model was obtained as 0.999 (with a standard deviation of 0.0001), showing approximately similar clustering solutions for the two models. The summary of the parameters estimated by the K·· model is available in Additional file: Appendix 4 and the kernel density estimates graph is available in Appendix 5.

The K·R model is used to achieve the final objective of the family of BMMs, which is to identify the differentially methylated CpG sites between multiple DNA samples. As there are two sample types in the simulated datasets, there can be $K = M^R = 9$ different combinations of the methylation states between sample A and sample B. The CpG sites having hypomethylation state in one sample and hypermethylation in another sample and vice versa are of prime interest as they identify the epigenetic changes in the cancer genome. The density estimates of the clustering solution for this model for a single simulated dataset are shown in Figure 2. The ARI calculated from the contingency Table 2 showing the clustering solution for this simulated dataset illustrated 92.9% accuracy between the actual membership and the observed groupings. The density estimates of the clustering solution show that cluster 2 identifies the CpG sites which have tendency to be hypomethylated in sample A and hypermethylated CpG sites in sample B. Cluster 1 identifies CpG sites which have a tendency of being hyper-

Table 2: Contingency table: identify CpG sites which tend to be differentially methylated between the 2 samples

Methylation state change — Cluster	1	2	3	4	5	6	7	8	9
Hypermethylation - Hemimethylation							1993	14	
Hypermethylation - Hypermethylation						30	975		
Hypermethylation - Hypomethylation	2861		5						
Hemimethylation - Hypomethylation	3		1335		661				
Hypomethylation - Hemimethylation		7		979					
Hemimethylation - Hemimethylation						5	2	2967	
Hypomethylation - Hypomethylation									3060
Hypomethylation - Hypermethylation		2958		1					
Hemimethylation - Hypermethylation						2140		4	

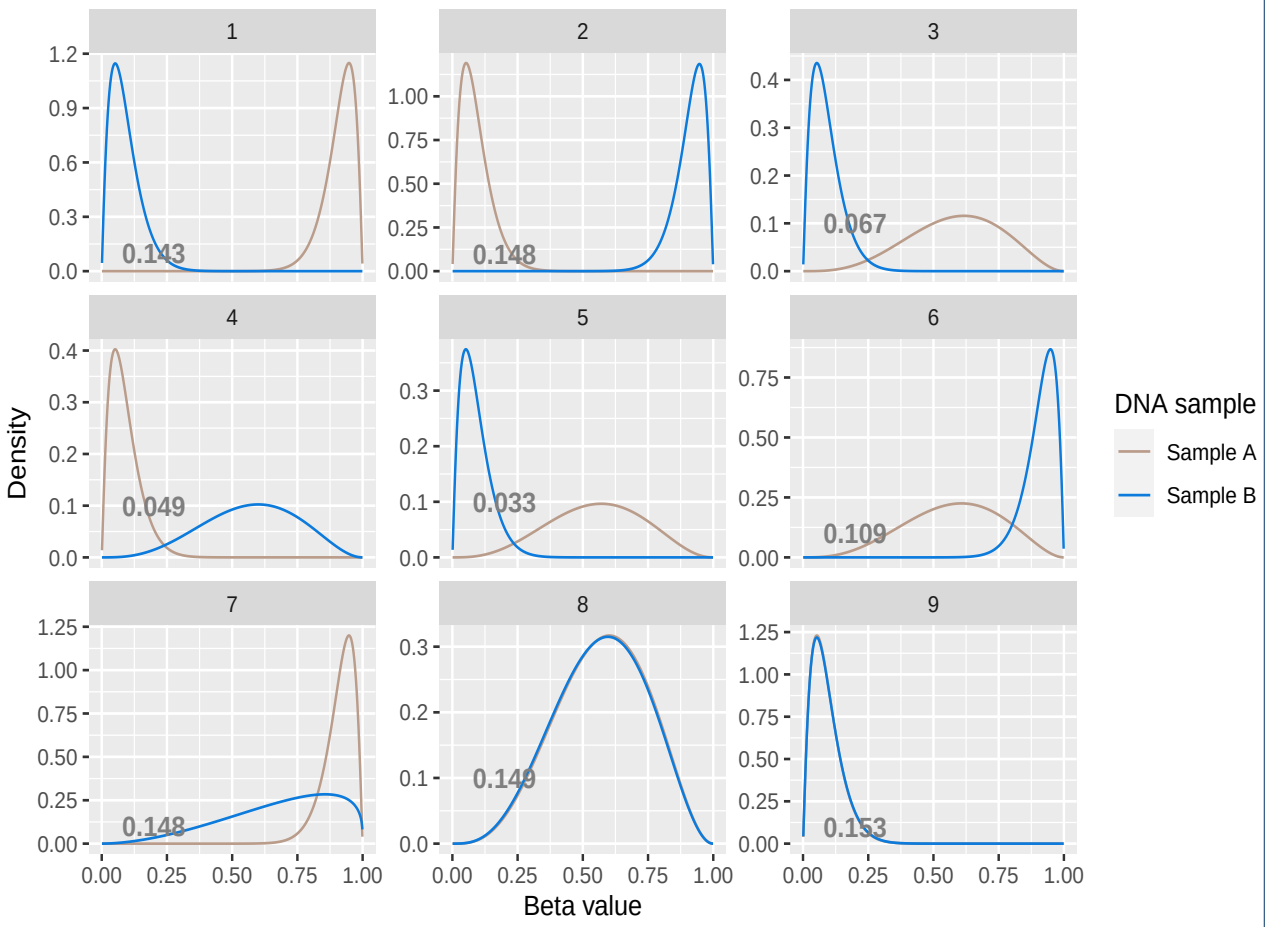

Figure 2: Fitted density estimates under the clustering solution of K·R model on a simulated dataset. Differentially methylated CpG sites between sample A and sample B are identified. The proportion of CpG sites belonging to each of the 9 clusters is displayed in the relevant panel.

methylated in sample A and hypomethylated in sample B. The summary of the parameters estimated by the K·R model is available in Additional file: Appendix 4. The mean ARI for all 20 simulated datasets was obtained as 0.916 (with a standard deviation of 0.041), suggesting very high accuracy in identifying the CpG sites which typically have differential methylation between multiple DNA samples. The kernel density estimate graph for the clustering solution of the K·R model is available in Additional file: Appendix 6.

Prostate cancer data results

The PCa dataset has *beta* values for each of the $C = 694{,}820$ CpG sites from $R = 2$ DNA samples collected from $N = 4$ patients. The first step of analysing the PCa DNA methylation dataset is to cluster the CpG sites in the benign/tumour samples into the 3 known methylation states and objectively infer the thresholds from the clustering solution. To do so, we use the K·· and KN· models from the family of BMM and analyse the clustering solution of the optimal model selected. The BIC score suggested the KN· model to be the optimal model for this dataset while clustering the benign and tumour samples individually. The density estimates for the clustering solution of the KN· model for the benign sample collected from patient 1 are displayed in Figure 3. The summary of the parameters estimated by the KN· model is available in Additional file: Appendix 4. As the parameters are allowed to vary for each patient in the KN· model, similar graphs are obtained and thresholds are calculated from parameters estimated for each patient. The KN· model estimates slightly different parameters for each patient, resulting in slightly different threshold points for each patient, suggesting variation in the degree of each patient's disease. The fitted density estimates graph for the other 3 patients are available in the Additional file: Appendix 7. The threshold points were observed to be 0.258 and 0.747 for the benign sample and 0.19 and 0.751 for the tumour sample collected from patient 1. The threshold point for clustering the CpG sites which tend to be hypermethylated in cluster 3 are almost similar for both samples. In contrast, they are quite different for identifying the CpG sites which are more inclined to be hypomethylated. Although these thresholds are close to the subjectively determined values suggested in the literature, the slight difference is enough to identify more CpG sites belonging to the two main methylation states, hypomethylation and hypermethylation, with the BMM approach than identified using the subjective thresholds. The ARI of 0.94 for the benign sample and 0.96 for the tumour sample when clustered using the KN· and K·· models indicate good agreement between the clustering solutions of the two models. The kernel density estimates graph for the clustering solution of the KN· model is available in Additional file: Appendix 8.

The next step for analysing the PCa dataset is to examine the samples together and identify the differentially methylated CpG sites between the benign and tumour samples. To achieve the said objective, the K·R model from the family of BMMs is used. In Figure 4, the density estimates of the clustering solution are illustrated where clusters 1—4 identify the CpG sites where the methylation state tends to be changing between the benign and tumour samples. The CpG sites

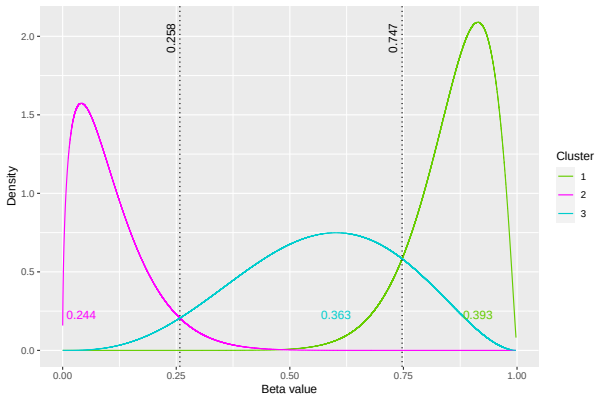

Figure 3: Fitted density estimates under the clustering solution of KN· model for DNA methylation data from the benign sample. The clusters identifies the CpG sites in the benign sample collected from patient 1 which are likely to belong to the groups representing the 3 methylation states. The thresholds are calculated based on the parameters estimated for the patient. The proportion of CpG sites belonging to each cluster is displayed.

that tend to be hypermethylated in the benign sample and hemimethylated in the tumor sample are identified by Cluster 1. Clusters 2—4 show a methylation shift from benign prostate tissue's hypomethylation to tumour tissue's hypermethylation, which shows that the methylation is changing as the cancer progresses. CpG sites in clusters 6 and 7 are likely to be hypermethylated in both samples, whereas those in clusters 8 and 9 are likely to be hypomethylated, and those in cluster 5 are likely to be in a hemimethylated condition in both samples. Cluster 4 identifies the maximum number of CpG sites with the highest value of τ as 0.163, whereas cluster 1 identifies the lowest proportion of CpG sites with τ as 0.058. The K·R model identifies 44.6% of the CpG sites in this dataset as differentially methylated. Table 3 lists the summary of the parameters estimated by the K·R model. The kernel density estimates for the clustering solution of the K·R model are illustrated in Additional file: Appendix 9.

The maximum uncertainty when clustering the CpG sites into K clusters is $1 - 1/K = 8/9$. In Figure 5, all the CpG sites are seen to have clustering uncertainties below 0.5 which demonstrates that the K·R model clusters the CpG sites with low uncertainty.

Using the K·R model, we identify more DMCs related to the genes implicated in prostate cancer carcinogenesis than with conventional methods [26]. The DMCs were mapped to DMRs by defining a DMR to arise when ≥ 3 adjacent CpG sites were identified

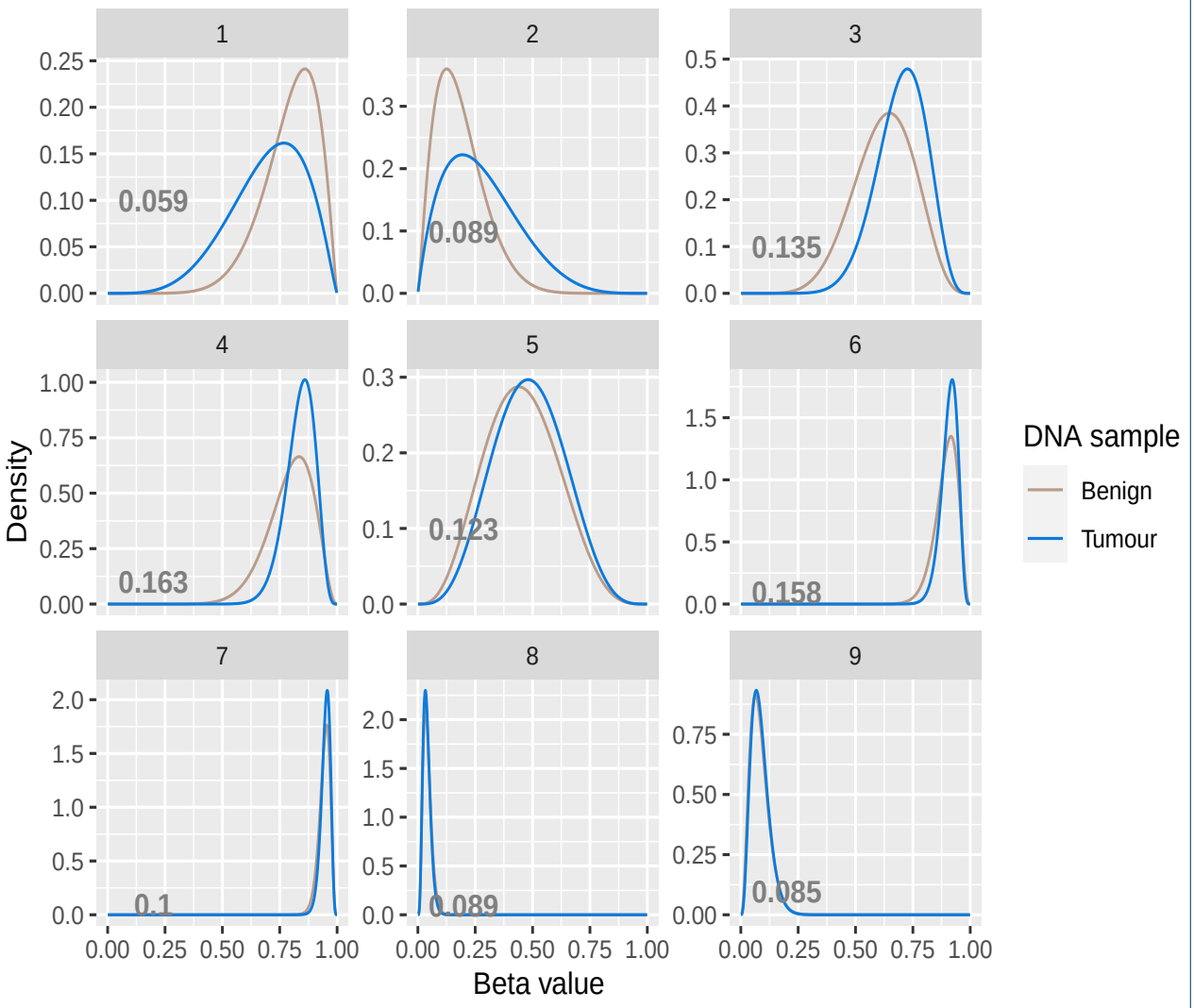

Figure 4: Fitted density estimates under the clustering solution of the K·R model for DNA methylation data from benign and tumour prostate samples. The proportion of CpG sites belonging to each of the 9 clusters is displayed in the relevant panel.

Table 3: Summary of the parameters estimated for the prostate cancer dataset using the K·R model

(a) Benign sample

Clusters	α	δ	mean	std. deviation
Cluster 1	8.815	2.277	0.795	0.116
Cluster 2	2.324	10.223	0.185	0.106
Cluster 3	8.005	4.810	0.625	0.130
Cluster 4	13.006	3.387	0.793	0.097
Cluster 5	4.071	4.924	0.453	0.157
Cluster 6	33.720	4.006	0.894	0.050
Cluster 7	84.926	5.023	0.944	0.024
Cluster 8	4.842	112.897	0.041	0.018
Cluster 9	3.749	41.317	0.083	0.041

(b) Tumour Sample

Clusters	α	δ	mean	std. deviation
Cluster 1	5.076	2.231	0.695	0.160
Cluster 2	1.975	5.040	0.282	0.159
Cluster 3	12.058	5.170	0.700	0.107
Cluster 4	27.000	5.249	0.837	0.064
Cluster 5	4.686	4.990	0.484	0.153
Cluster 6	56.506	5.734	0.908	0.036
Cluster 7	111.727	5.978	0.949	0.020
Cluster 8	5.455	133.043	0.039	0.016
Cluster 9	4.194	45.197	0.085	0.039

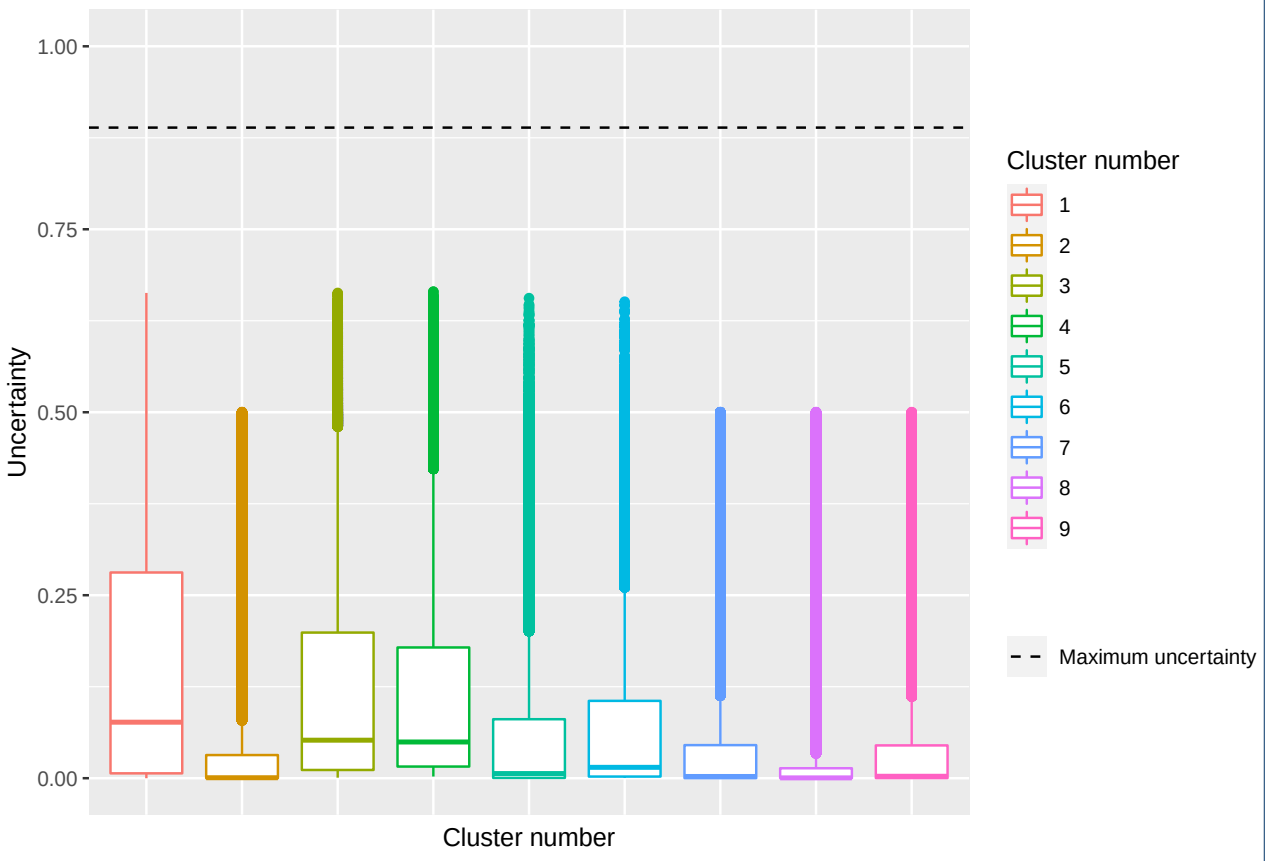

Figure 5: Clustering uncertainties for CpG sites in each clustering group.

as being differentially methylated by the K·R model. Our approach identified more DMRs than conventional methods [26]. Ten differentially methylated CpG sites were mapped to the GSTP1 genes, 35 DMCs mapped to APC genes, 48 DMCs mapped to RARB genes, 31 DMCs mapped to RASSF1 genes and 22 DMCs mapped to SFRP2 genes. Non-parametric tests were performed to test the hypothesis that the *beta* values were higher in the tumour sample than in the benign sample for the CpG sites related to these genes. The p-values obtained were less than the chosen 0.05 significance level for most patients, suggesting the median *beta* value is significantly higher in the tumour sample than in the benign sample. Hypermethylation of RARB promoter genes is a significant biomarker in diagnosing prostate cancer [38]. Figure 6 shows a boxplot displaying the distribution of the differentially methylated CpG sites belonging to the RARB genes for all samples. The boxplot illustrates that the median *beta* value is higher in the tumour sample than in the benign sample for all patients.

Figure 7 shows the empirical cumulative distribution function (ECDF) for the CpG sites identified as differentially methylated related to the RARB genes for all samples. The ECDF illustrates that the identified differentially methylated CpG sites have increased *beta* values in the tumour sample than in the benign sample.

The family of BMMs was also applied to a publicly available esophageal squamous cell carcinoma (ESCC) dataset. Similar results to those obtained from the prostate cancer dataset were obtained for the ESCC dataset. The results are discussed in detail in the Additional file: Appendix 10.

Software

An R package called has been developed to facilitate the widespread use of the BMMs. Each BMM model can be fitted individually or all BMM models can be applied simultaneously in a wrapper function `betaclust`. The *AIC*, *BIC* and *ICL* are calculated to facilitate selection of the optimal model. Statistical

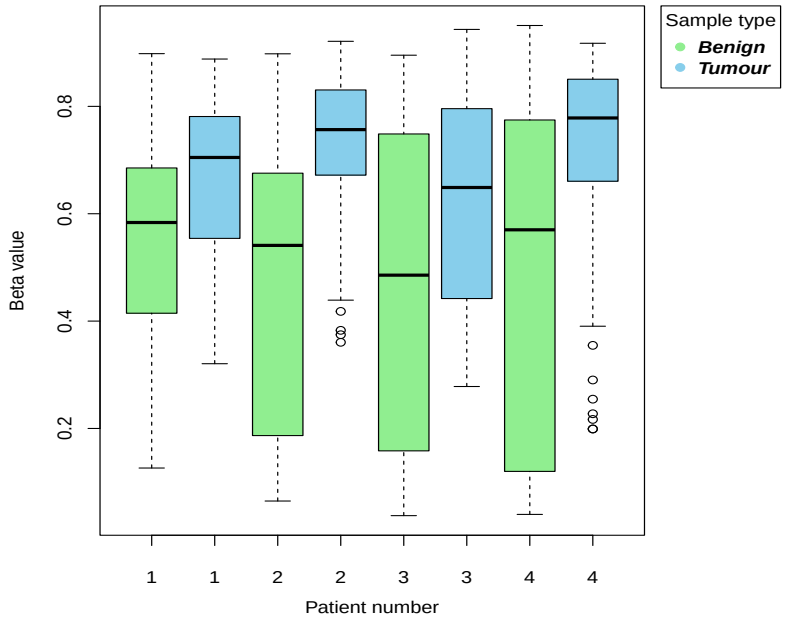

Figure 6: Boxplot of differentially methylated CpG sites related to the RARB genes in the benign and tumour sample.

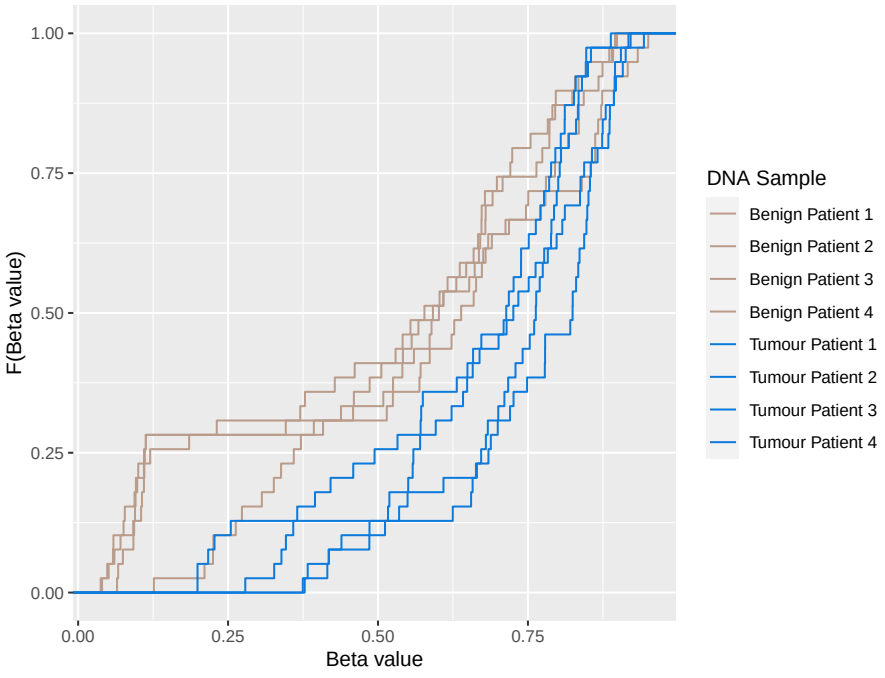

Figure 7: ECDFs for all the differentially methylated CpG sites related to the RARB genes for all the patient samples.

summaries of the fitted models are provided. A variety of plotting options are available for visualization, similar to those shown here. In addition to static plots, there is also the facility to construct interactive density plots. The package has been published in github and will be published in CRAN.

Discussions

DNA methylation process is being studied widely for disease diagnosis and treatment. Technology advancements have led to the development of microarrays that can process 850,000 CpG sites from a DNA sample. But it has been difficult to explore such large arrays, and the analysis has been limited to a specific range as there is a lack of appropriate statistical methods for such large data. The methylation states are identified using thresholds and based on them, DMCs are identified. Such thresholds are defined in literature based on intuition and are not statistically constructed. The *beta* values are usually transformed to Gaussian distributed values using a logit transformation and modelled using Gaussian mixture models. A transformation-based approach to density estimation using the Gaussian mixture model for bounded data has been proposed [39]. The bounded data are first transformed using a range-power transformation, and then the density is estimated using the Gaussian mixture model. The EM algorithm estimates the density parameters for the transformed data. This method makes Gaussian mixture models flexible to handle bounded data. However, our proposed approach advocates against transforming the data and proposes modelling the data in its innate form. Also, the transformation-based approach aims to estimate the density of the data, whereas our approach clusters the data into latent groups based on the estimated densities.

We developed novel mixture models to model these data as a mixture of beta distributions to (i) identify methylation states of each CpG site in a sample, (ii) objectively infer thresholds and (iii) identify the DMCs between different samples. These models remove the dependency of estimating the methylation state of a CpG site on arbitrary thresholds. We evaluated these models on several simulated datasets, a prostate cancer dataset and an esophageal squamous cell carcinoma dataset. The BMMs are computationally efficient, utilising parallel programming and digamma function approximation. The K·· and KN· models accomplish the first two objectives by clustering the CpG sites in a single DNA sample into the known methylation states. The K·R model achieves the end goal by clustering the CpG sites from multiple DNA samples to determine the CpG sites with differential methylation.

An important assumption for the BMMs is that the beta distributions are unimodal, restricting the shape parameters to greater than 1. If either of the shape parameters is estimated to be less than 1, the models are re-executed with new starting values for the EM algorithm. The BMMs can be extended further to consider bimodal beta distributions and not restrict the process to unimodal beta distributions. Even though certain studies may be interested in identifying DMCs, most researchers are more focused on finding DMRs because they are of greater biological interest. After the DMCs are obtained using the BMMs, the DMRs can be identified in predefined regions, such as CpG islands and CpG shores. An important assumption for the BMMs is that the methylation states of the adjacent CpG sites are independent and uncorrelated. On the contrary, the methylation states of adjacent CpG sites have a high spatial correlation suggesting that if a CpG site is hypermethylated, the neighbouring CpG sites tend to be hypermethylated [40]. Even though we uncover more DMCs and DMRs than the conventional methods, we model each CpG site independently, ignoring spatial correlation between the CpG sites. The BMMs can be modified to consider the spatial correlation between adjacent CpG sites. Additionally, the BMMs first locate DMCs before determining DMRs from the located DMCs. The family of BMMs could to be broadened so that the intermediary step of identifying DMCs is removed and the spatial correlation is taken into account to find DMRs of interest.

The methylation state of a human genome changes over time depending on clinical conditions. Longitudinal data of methylation changes in patients are collected to study the effect of environmental changes or treatments on disease progression. Such data are massive and current methods cannot handle such extensive data. We can track how the methylation variability of a group of people varies over time using longitudinal data. In order to analyze the methylation changes over a certain period in multiple patients, the BMMs could be further enhanced to incorporate models specifically for processing longitudinal methylation data.

Conclusions

The BMMs were developed specifically for DNA methylation data and to identify DMCs by modelling the data without transforming it. This work indicates that more CpG sites can be objectively identified which can lead to deeper analysis of the methylation data and finer understanding of the epigenetic changes. This work eliminates the need for arbitrary thresholds for identifying the methylation states. The BMMs are computationally efficient, achieved by the digamma approximation approach. The BMM can be used to

identify the differentially methylated CpG sites, thus embedding this epigenetic analysis in a model-based framework.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KM, TBM and ICG conceived, designed, analysed and implemented the BMMs, the betaclust package and wrote the manuscript. RS and ASP conceived, acquired and processed data. All authors read and approved the final manuscript.

Acknowledgements

This publication has emanated from research supported in part by a Grant from Science Foundation Ireland under Grant number 18/CRT/6049.

Author details

[1]School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland. [2]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland. [3]School of Medicine, UCD Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland.

References

1. Berger, S.L., Kouzarides, T., Shiekhattar, R., Shilatifard, A.: An operational definition of epigenetics. Genes and development **23(7)**, 781–783 (2009). doi:10.1101/gad.1787609
2. Moore, L.D., Le, T., Fan, G.: Dna methylation and its basic function. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology **38(1)**, 23–38 (2013). doi:10.1038/npp.2012.112
3. Jin, Z., Liu, Y.: Dna methylation in human diseases. Genes and diseases **5(1)**, 1–8 (2018). doi.org/10.1016/j.gendis.2018.01.002
4. Das, P., Singal, R.: Dna methylation and cancer. Journal of clinical oncology **22(22)**, 4632–4642 (2004). doi: 10.1200/JCO.2004.07.151
5. Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S.: Epigenome-wide association studies for common human diseases. Nature reviews genetics **12(8)**, 529–541 (2011). doi:10.1038/nrg3000
6. Bird, A.: Dna methylation patterns and epigenetic memory. Genes and development **16(1)**, 6–21 (2002). doi: 10.1101/gad.947102
7. Chen, D.P., Lin, Y.C., Fann, C.S.: Methods for identifying differentially methylated regions for sequence- and array-based data. Briefings in functional genomics **15(6)**, 485–490 (2016). doi: 10.1093/bfgp/elw018
8. P de Almeida, B., Apolónio, J.D., Binnie, A., Castelo-Branco, P.: Roadmap of dna methylation in breast cancer identifies novel prognostic biomarkers. BMC cancer **19** (2019). doi:10.1186/s12885-019-5403-0
9. Cho, J.W., Hong, M.H., Ha, S.J., Kim, Y.J., Cho, B.C., Lee, I., Kim, H.R.: Genome-wide identification of differentially methylated promoters and enhancers associated with response to anti-pd-1 therapy in non-small cell lung cancer. Experimental and molecular medicine **52(9)**, 1550–1563 (2020). doi:10.1038/s12276-020-00493-8
10. Kim, J.H., Dhanasekaran, S.M., Prensner, J.R., Cao, X., Robinson, D., Kalyana-Sundaram, S., Huang, C., Shankar, S., Jing, X., Iyer, M., Hu, M., Sam, L., Grasso, C., Maher, C.A., Palanisamy, N., Mehra, R., Kominsky, H.D., Siddiqui, J., Yu, J., Qin, Z.S., Chinnaiyan, A.M.: Deep sequencing reveals distinct patterns of dna methylation in prostate cancer. Genome research **21(7)**, 1028–041 (2011). doi: 10.1101/gr.119347.110
11. Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., Clark, S.J.: Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. Genome biology **17(1)**, 208 (2016). doi: 10.1186/s13059-016-1066-1
12. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L., Lin, S.M.: Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. BMC bioinformatics **11**, 587 (2010). doi: 10.1186/1471-2105-11-587
13. 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 **12** (2021). 10.3389/fgene.2021.705708
14. Siegmund, K.D., Laird, P.W., Laird-Offringa, I.A.: A comparison of cluster analysis methods using dna methylation data. Bioinformatics **20(12)**, 1896–1904 (2004). doi: 10.1093/bioinformatics/bth176
15. Houseman, E.A., Christensen, B.C., Yeh, R.F., Marsit, C.J., Karagas, M.R., Wrensch, M., Nelson, H.H., Wiemels, J., Zheng, S., Wiencke, J.K., Kelsey, K.T.: Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC bioinformatics **9**, 365 (2008). doi: 10.1186/1471-2105-9-365
16. Koestler, D.C., Christensen, B.C., Marsit, C.J., Kelsey, K.T., Houseman, E.A.: Recursively partitioned mixture model clustering of dna methylation data using biologically informed correlation structures. Statistical applications in genetics and molecular biology **12(2)**, 225–240 (2013). doi: 10.1515/sagmb-2012-0068
17. P E de Souza, C., Andronescu, M., Masud, T., Kabeer, F., Biele, J., Laks, E., Lai, D., Ye, P., Brimhall, J., Wang, B., Su, E., Hui, T., Cao, Q., Wong, M., Moksa, M., Moore, R.A., Hirst, M., Aparicio, S., Shah, S.P.: Epiclomal: Probabilistic clustering of sparse single-cell dna methylation data. PLoS computational biology **16(9)** (2020). doi: 10.1371/journal.pcbi.1008270
18. Ma, Z., Teschendorff, A.E.: A variational bayes beta mixture model for feature selection in dna methylation studies. Journal of bioinformatics and computational biology **11(4)** (2013). doi:10.1142/S0219720013500054
19. Zhang, L., Meng, J., Liu, H., Huang, Y.: A nonparametric bayesian approach for clustering bisulfate-based dna methylation profiles. BMC genomics **13 Suppl 6(Suppl 6):S20** (2012). doi:10.1186/1471-2164-13-S6-S20
20. Gevaert, O., Tibshirani, R., Plevritis, S.K.: Pancancer analysis of dna methylation-driven genes using methylmix. Genome biology **16(1)**, 17 (2015). doi: 10.1186/s13059-014-0579-8
21. Wang, D., Yan, L., Hu, Q., Sucheston, L.E., Higgins, M.J., Ambrosone, C.B., Johnson, C.S., Smiraglia, D.J., Liu, S.: Ima: an r package for high-throughput analysis of illumina's 450k infinium methylation data. Bioinformatics **28(5)**, 729–730 (2012). doi:10.1093/bioinformatics/bts013
22. Warden, C.D., Lee, H., Tompkins, J.D., Li, X., Wang, C., Riggs, A.D., Yu, H., Jove, R., Yuan, Y.C.: Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis. Nucleic acids research [published correction appears in Nucleic acids research. 2019 Sep 5;47(15):8335-8336] **41(11)**, 117 (2013). doi:10.1093/nar/gkt242
23. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71(3)**, 209–249 (2021). doi: 10.3322/caac.21660
24. P, R.: Epidemiology of prostate cancer. World journal of oncology **10(2)**, 63–89 (2019). doi:10.14740/wjon1191
25. Li, L.-C., Okino, S.T., Dahiya, R.: Dna methylation in prostate cancer. Biochimica et Biophysica Acta (BBA) - reviews on cancer **1704(2)**, 87–102 (2004). doi:10.1016/j.bbcan.2004.06.001
26. Silva, R., Moran, B., Russell, N.M., Fahey, C., Vlajnic, T., Manecksha, R.P., Finn, S.P., Brennan, D.J., Gallagher, W.M., Perry, A.S.: Evaluating liquid biopsies for methylomic profiling of prostate cancer. Epigenetics **15(6-7)**, 715–727 (2020). doi: 10.1080/15592294.2020.1712876
27. 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 **8(3)**, 389–399 (2016). doi:10.2217/epi.15.114
28. Müller, F., Scherer, M., Assenov, Y., Lutsik, P., Walter, J., Lengauer, T., Bock, C.: Rnbeads 2.0: comprehensive analysis of dna methylation data. Genome biology **20(1)**, 55 (2019). doi:10.1186/s13059-019-1664-9
29. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society **39(1)**, 1–38 (1977). doi:10.1111/j.2517-6161.1977.tb01600.x

30. Nocedal, J., Wright, S.J.: Quasi-newton methods. Numerical optimization. Springer series in operations research and financial engineering, 192–221 (1999). doi:10.1007/0-387-22742-3_8
31. Berndt, E., Hall, B., Hall, R., Hausman, J.: Estimation and inference in non-linear structural models. Annals of economic and social measurement **3**, 653–665 (1974)
32. Diamond, H.G., Straub, A.: Bounds for the logarithm of the euler gamma function and its derivatives. Journal of mathematical analysis and applications **433(2)**, 1072–1083 (2016). doi:10.1016/j.jmaa.2015.08.034
33. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, and F. Csaki (Eds.), 2nd international symposium on information theory, 267–281 (1973). doi:10.1007/978-1-4612-1694-0_15
34. Schwarz, G.: Estimating the dimension of a model. Annals of statistics **6**, 461–464 (1978). doi: 10.1214/aos/1176344136
35. 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 **22(7)**, 719–725 (2000). 10.1109/34.865189
36. Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: clustering, classification and density estimation using gaussian finite mixture models. The R journal **8(1)**, 289–317 (2016). PMCID: PMC5096736
37. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**, 193–218 (1985). doi:10.1007/BF01908075
38. Ameri, A., Alidoosti, A., Hosseini, S.Y., Parvin, M., Emranpour, M.H., Taslimi, F., Salehi, E., Fadavip, P.: Prognostic value of promoter hypermethylation of retinoic acid receptor beta (rarb) and cdkn2 (p16/mts1) in prostate cancer. Chinese journal of cancer research **23(4)**, 306–311 (2011). doi:10.1007/s11670-011-0306-x
39. Scrucca, L.: A transformation-based approach to gaussian mixture density estimation for bounded data. Biometrical journal **61(4)**, 873–888 (2019). doi:10.1002/bimj.201800174
40. Hodges, E., Smith, A.D., Kendall, J., Xuan, Z., Ravi, K., Rooks, M., Zhang, M.Q., Ye, K., Bhattacharjee, A., Brizuela, L., McCombie, W.R., Wigler, M., Hannon, G.J., Hicks, J.B.: High definition profiling of mammalian dna methylation by array capture and single molecule bisulfite sequencing. Genome research **19(9)**, 1593–1605 (2009). doi:10.1101/gr.095190.109