Bioinformatics, 32, 2016, i405–i412
doi: 10.1093/bioinformatics/btw432
ECCB 2016

Higher order methylation features for clustering and prediction in epigenomic studies

ChantrioInt-Andreas Kapourani[1] and Guido Sanguinetti[1,2,]*

[1]IANC, School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK and [2]Synthetic and Systems Biology, University of Edinburgh, Edinburgh EH9 3JD, UK

*To whom correspondence should be addressed.

Abstract

Motivation: DNA methylation is an intensely studied epigenetic mark, yet its functional role is incompletely understood. Attempts to quantitatively associate average DNA methylation to gene expression yield poor correlations outside of the well-understood methylation-switch at CpG islands.
Results: Here, we use probabilistic machine learning to extract higher order features associated with the methylation profile across a defined region. These features quantitate precisely notions of shape of a methylation profile, capturing spatial correlations in DNA methylation across genomic regions. Using these higher order features across promoter-proximal regions, we are able to construct a powerful machine learning predictor of gene expression, significantly improving upon the predictive power of average DNA methylation levels. Furthermore, we can use higher order features to cluster promoter-proximal regions, showing that five major patterns of methylation occur at promoters across different cell lines, and we provide evidence that methylation beyond CpG islands may be related to regulation of gene expression. Our results support previous reports of a functional role of spatial correlations in methylation patterns, and provide a mean to quantitate such features for downstream analyses.
Availability and Implementation: https://github.com/andreaskapou/BPRMeth
Contact: G.Sanguinetti@ed.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA methylation is a well-studied, heritable epigenetic modification that plays an important role in gene regulatory mechanisms. It is associated with a broad range of biological processes of direct clinical relevance, including X-chromosome inactivation, genomic imprinting, silencing of repetitive DNA and carcinogenesis (Baylin and Jones, 2011; Feinberg and Vogelstein, 1983; Li *et al.*, 1993). Methylation occurs when a methyl group is attached to a DNA nucleotide. In vertebrate genomes, methylation is observed almost exclusively on 5-methylcytosine (5-mC) residues in the context of CpG dinucleotides. Due to increased vulnerability of 5-mC to randomly deaminate into thymine, most of the genome is depleted from CpG dinucleotides, except from small CpG-rich regions, termed CpG islands (CGIs) (Bird, 2002). Hyper-methylation of CGIs near promoter regions is generally associated with transcriptional repression; however, outside of this well documented case, the association between DNA methylation across promoter-proximal regions and transcript abundance is considerably weaker and poorly understood (Jones, 2012).

Recent advances in high-throughput sequencing technology have made it possible to measure the methylation level of cytosines on a genome-wide scale with single nucleotide resolution. Sodium bisulphite treatment of DNA followed by sequencing (BS-seq) efficiently converts unmethylated cytosines to uracils (which are subsequently amplified as thymines by PCR) and leaves the 5-mCs unmodified (Krueger *et al.*, 2012). To obtain DNA methylation levels, reads are aligned to a reference genome allowing changes of cytosines to thymines during the mapping procedure. A variant of BS-seq technology, termed Reduced Representation Bisulphite Sequencing (RRBS) (Meissner *et al.*, 2005), uses methylation-sensitive restriction enzymes to cleave the DNA at specific loci before bisulphite treatment. This results in measuring in greater coverage and at lower cost the methylation level of CpG-rich regions genome-wide.

Despite the widespread take up of BS-seq technology, statistical modeling of such data is still challenging, yet it is crucial in order to uncover biological regulatory mechanisms. Analysis of BS-seq data has mainly focused on identifying differentially methylated regions (DMRs) across different conditions. Some notable DMR methods are BSmooth (Hansen *et al.*, 2012), Bi-Seq (Hebestreit *et al.*, 2013) and M^3D (Mayo et al., 2015). While DMR detection methods are often crucial ingredients in exploratory data analysis pipelines, they do not provide a clear platform to quantitatively understand the relationship between DNA methylation and gene expression. Most studies use DMR detection as a pre-filter, and then simply correlate mean methylation levels across each region (often taken to be promoter-proximal) with gene expression. Adopting this simple approach, genome-wide studies (Bock *et al.*, 2012; Hansen *et al.*, 2011) have reported only modest correlation between average DNA-methylation and gene expression (Pearson's correlation coefficient $r \approx -0.3$).

In this article, we argue that part of the difficulty in quantitatively associating methylation levels with gene expression resides in the simplistic encoding of DNA methylation across a region as a simple average. DNA methylation often displays reproducible, spatially correlated patterns (*profiles*); Figure 1 shows two examples from an ENCODE datasets (Dunham et al., 2012). This spatial reproducibility was exploited by Mayo *et al.* (2015) to provide more powerful tests for DMR, and by Vanderkraats *et al.* (2013) to group genes with similar differential methylation patterns and corresponding expression changes. These results suggest that a precise quantification of the spatial variability in the DNA methylation mark may aid the quest to quantitatively understand the interplay between methylation and transcription. We propose a probabilistic model of methylation profiles, based on latent variable models, which allow us to associate with each region of interest a set of features capturing precisely the methylation profile across the region. We then show that, using such features, we can construct an accurate machine learning predictor of gene expression from DNA methylation, achieving test correlations twice as large as previously reported.

The rest of the paper is organized as follows: we start off by providing a high-level description of our approach. We then describe precisely the statistical methodology we propose. We illustrate our approach on three ENCODE datasets, showing that higher order features allow much more accurate predictions of gene expression. We also show how such features can be used to cluster regions according to their methylation profiles, and show that five prototypical methylation profiles appear to explain most variability in promoter-proximal methylation in human cell lines.

2 Approach

In this article, we propose a novel probabilistic machine learning methodology to quantify the profile of DNA methylation across genomic regions from BS-seq data. Our motivation is practical: inspection of many BS-seq datasets reveals that methylation levels across promoter-proximal regions often show reproducible, spatially correlated profiles. Figure 1 shows two example promoter-proximal regions which clearly display such spatial correlations, resulting in characteristic methylation shapes. We propose a method to quantitate such qualitative information.

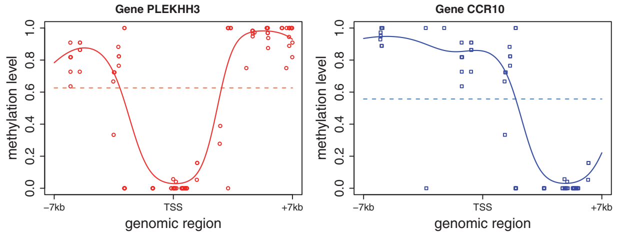

Fig. 1. Methylation patterns for the PLEKHH3 and CCR10 genes from the K562 cell line over ±7*kb* promoter region. Each point represents the relative CpG location w.r.t. TSS and the corresponding DNA methylation level. The dashed horizontal lines show the average methylation level. The shapes of methylation profiles are very different, however, the average methylation level cannot explain them. Also, note that there are no CpG measurements in the (−6 kb, −4 kb) region for the CCR10 gene, and the learned methylation profiles can be thought as imputing the missing values by taking into consideration the spatial co-dependence of nearby CpGs

The method is based on a Generalized Linear Model of basis function regression coupled with a Binomial observation likelihood, and allows us to associate each region with a set of basis function coefficients which capture the methylation profile. We show how such higher order features can then be used in downstream analysis to yield a significantly improved estimate of the correlation between methylation and gene expression, and to identify prototypical methylation profiles across promoter regions.

3 Methods

3.1 Modeling DNA methylation profiles

As in most HTS-based assays, the output of a BS-seq experiment is a set of reads aligned to the genome; the main difference is that the bisulphite treatment changes to thymine any unmethylated cytosine. Thus, the same base on the genome will appear as cytosine on some reads, and as thymine on others; the ratio of reads containing a cytosine readout to total reads gives a measurement of the sample methylation level. This measurement process at a single cytosine can be naturally modeled with a Binomial distribution, where the number of successes represents the number of reads on which the cytosine actually appears as C, and the number of attempts is the total number of reads mapping to the specific site. Let t be the total number of reads that are mapped to a specific CpG site, and let m of these reads to contain methylated cytosines. Then, for each CpG site we assume that $m \sim \mathcal{B}inom(t, p)$, where p is the unknown methylation level.

In this article, and in many practical studies, we are interested in learning the methylation patterns of fixed-width genomic regions, e.g. promoters. Hence, each genomic region $i (i = 1, \dots, N)$ can be represented as a vector of CpG locations $\mathbf{x}_i$, where each entry corresponds to the location of the CpG in the genomic region, relative to a reference point such as the Transcription Start Site (TSS). It should be noted that the vector lengths L_i may vary between different genomic regions, since they depend on the number of actual CpG dinucleotides found in each region. For each region i, we also have a vector of observations $\mathbf{y}_i$, containing the methylation levels of the corresponding CpG sites; each entry consists of the tuple $y_{il} = (m_{il}, t_{il})$, where, m_{il} is the number of 5-mC reads mapped to the l-th CpG site in region i, and t_{il} corresponds to the total number of reads.

Direct comparison of the observation vectors $\mathbf{y}_i$ for different regions is complicated due to the variability in the vector lengths. To enable comparisons between these regions, we formulate our problem as a regression problem, where the methylation profile of each genomic region is modeled as a linear combination of a set of latent basis functions. Let $f(\mathbf{x}_i)$ be a latent function representing the methylation profile for genomic region i. Since the observed methylation data contain the proportion of methylated reads out of the total reads for each CpG site, each entry of the vector $\mathbf{y}_i$ takes values in the $[0, 1]$ interval. Thus, we introduce an unconstrained latent function $g(\mathbf{x}_i)$ defined so that $f(\mathbf{x}_i)$ is the probit transformation of $g(\mathbf{x}_i)$: $f(\mathbf{x}_i) = \Phi(g(\mathbf{x}_i))$, where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the standard normal distribution. Let $\mathbf{f}_i = f(\mathbf{x}_i)$ and $\mathbf{g}_i = g(\mathbf{x}_i)$ be shorthand for the values of the latent functions.

Given the values of the latent function $\mathbf{f}_i$ for region i, the observations y_{il} for each CpG site are independent and identically distributed Binomial variables, so we can define the joint log-likelihood for region i in factorized form:

$$\log\, p(\mathbf{y}_i|\mathbf{f}_i) = \sum_{l=1}^{L_i} \log\{\mathcal{B}inom(m_{il}|t_{il}, \Phi(g_{il}))\} \tag{1}$$

From its final form, we refer to this observation model as the Binomial distributed Probit Regression (BPR) likelihood function. Notice that the BPR model explicitly accounts for the coverage variability across CpG sites through the use of the Binomial observation model: as the variance of a Binomial distribution decreases rapidly with the number of attempts, the model will be very strongly constrained by highly covered sites. Hence, it handles in a principled way the uncertainty present in low coverage reads during the analysis of BS-seq data.

3.2 Feature extraction

To constrain the latent function $\mathbf{g}_i$ we assume it is given as a linear combination of fixed non-linear basis functions $h_j(\cdot)$ of the input space $\mathbf{x}_i$, of the form:

$$\mathbf{g}_i(\mathbf{x}_i, \mathbf{w}_i) = \sum_{j=0}^{M-1} w_j h_j(\mathbf{x}_i) = \mathbf{H}_i \mathbf{w}_i \tag{2}$$

where $\mathbf{w}_i = (w_{i,0}, \ldots, w_{i,M-1})^T$, $\mathbf{H}_i$ is the $L_i \times M$ design matrix, whose elements are given by $\mathbf{H}_{ilj} = h_j(x_{il})$, and M denotes the total number of basis functions. Hence, its probit transformation $\mathbf{f}_i$ is given by:

$$\mathbf{f}_i(\mathbf{x}_i, \mathbf{w}_i) = \Phi(\mathbf{g}_i(\mathbf{x}_i, \mathbf{w}_i)) = \Phi(\mathbf{H}_i \mathbf{w}_i) \tag{3}$$

One should note that even though the function $\mathbf{g}_i$ is linear with respect to the parameters $\mathbf{w}_i$, the latent function $\mathbf{f}_i$ is non-linear due to the presence of the probit transformation. In this study, we consider Radial Basis Functions (RBFs) since they are local functions of the input variable, so that changes in one region of the input space do not affect all other regions. For a single input variable x, the RBF takes the form $h_j(x) = exp(-\gamma||x - \mu_j||^2)$, where μ_j denotes the location of the j^{th} basis function in the input space and γ controls the spatial scale.

Learning the methylation profiles $\mathbf{f}_i$ for each genomic region, is equivalent to optimizing the model parameters $\mathbf{w}_i$. The parameters $\mathbf{w}_i$ can be considered as the extracted features which quantitate precisely notions of shape of a methylation profile. Optimizing $\mathbf{w}_i$ involves maximizing Equation (1) for each genomic region; however, by increasing the number of basis functions, we also increase the resolution for the shape of the methylation profiles, which might lead to overfitting. To ameliorate this issue, we maximize a penalized version of Equation (1), by adding a regularization term $\mathcal{E}$ $(\mathbf{w}_i)$ to the log-likelihood function which will encourage the weights to decay to zero:

$$J(\mathbf{w}_i) = \log p(\mathbf{y}_i|\mathbf{f}_i, \mathbf{w}_i) - \mathcal{E}(\mathbf{w}_i) \tag{4}$$

where $\mathcal{E}(\mathbf{w}_i) = \frac{1}{2}\mathbf{w}_i^T\mathbf{w}_i$ is the squared two-norm. This approach is known as ridge regression or weight decay. Direct maximization of $J(\mathbf{w}_i)$ w.r.t parameters $\mathbf{w}_i$ is intractable due to presence of the probit transformation, hence, we perform numerical optimization using the Conjugate Gradient method (see Section 1 of the Supplementary Material).

3.3 Predicting gene expression

The extracted higher-order methylation features across promoter-proximal regions can be used for downstream analysis, such as predicting transcript abundance, or performing clustering in order to learn prototypical methylation patterns that occur at promoters across different cell lines.

To quantitatively predict expression at each promoter region, we construct a regression model by taking as input the higher-order methylation features extracted from each promoter-proximal region. The performance of the regression model is evaluated by computing the root-mean squared error (RMSE) and the Pearson's correlation coefficient (r) between the predicted and the measured (log-transformed) gene expression levels. We compare our proposed model's performance with the standard approach (Bock *et al.*, 2012; Hansen *et al.*, 2011) which uses the average methylation level across a region as input feature (this approach can be thought of as fitting a constant function across each genomic region). We have tested both a linear regression model and a variety of non-linear models, such as SVM regression, Random Forests and Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991). For the rest of our analysis we use the SVM regression since it is consistently better than the other models (see Section 3 of the Supplementary Material).

In addition to the methylation profile features, we consider two supplementary sources of information which could plausibly act as confounders in the predictions. The first feature accounts for the goodness of fit of each methylation profile to the observed methylation data using the RMSE as error measure, intuitively quantitating the noisiness in the methylation profile. The second feature considers the number of CpG dinucleotides present in each promoter region. It is thought that CpG density may play a functional role in controlling gene expression, with the main evidence being the existence of CpG islands (Deaton and Bird, 2011).

3.4 Clustering methylation profiles

To cluster methylation profiles, we consider a mixture modeling approach (McLachlan and Peel, 2004). We assume that the methylation profiles $\mathbf{f}$ can be partitioned into at most K clusters, and each cluster k can be modeled separately using the BPR likelihood as our observation model. The log-likelihood for the mixture model is defined as:

$$p(\mathbf{y}|\boldsymbol{\Theta}) = \sum_{i=1}^{N} \log\left\{\sum_{k=1}^{K} \pi_k p(\mathbf{y}_i|\mathbf{f}_i, \mathbf{w}_k, z_i = k)\right\} \tag{5}$$

where $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_k, \mathbf{w}_1, \ldots, \mathbf{w}_k)$, π_k are the mixing proportions (with $\pi_k \in (0, 1)\forall k$ and $\sum_k \pi_k = 1$), $\mathbf{w}_k$ are the methylation profile parameters and z_i are the latent variables denoting to which cluster each genomic region belongs. To avoid cluttering the notation, we will omit the dependence of the observation model on the latent variables z_i.

3.4.1 Parameter estimation

To estimate the model parameters $\boldsymbol{\Theta} = (\pi_1, \ldots, \pi_k, \mathbf{w}_1, \ldots, \mathbf{w}_k)$, the Expectation Maximization (EM) algorithm (Dempster *et al.*, 1977) is considered. EM is a general iterative algorithm for computing maximum likelihood estimates when there are missing or latent variables, as in the case of mixture models. EM alternates between inferring the latent variables given the parameters (E-step), and optimizing the parameters given the posterior statistics of the latent variables (M-step). Formally, during the E-step we compute the responsibility that component k takes for explaining observations $\mathbf{y}_i$:

$$\gamma(z_{ik}) = \frac{\pi_k p(\mathbf{y}_i|\mathbf{f}_i, \mathbf{w}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{y}_i|\mathbf{f}_i, \mathbf{w}_j)} \tag{6}$$

The M-step consists of updating the model parameters so as to maximize the expected complete data log-likelihood. The mixing proportions π_k are updated as follows:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N}\gamma(z_{ik}) \tag{7}$$

To re-estimate the observation model parameters $\mathbf{w}_k$, we need to optimize the following quantity:

$$\ell(\mathbf{w}_k) = \sum_i \gamma(z_{ik}) \sum_l \log\{\mathcal{B}inom(m_{il}|t_{il}, \Phi(g_{il}; \mathbf{w}_k))\} \tag{8}$$

However, direct optimization of $\ell(\mathbf{w}_k)$ w.r.t parameters $\mathbf{w}_k$ is intractable, thus, we resort again to numerical optimization strategies. This variant of EM algorithm is known as Generalized EM, or GEM, and it is proved to converge to the maximum likelihood estimate (Wu, 1983). It should be noted that the penalized version of the BPR likelihood, given in Equation (4), can be easily incorporated in the clustering approach.

4 Datasets

To evaluate the performance of the proposed methodology we use real datasets that are publicly available from the ENCODE project consortium (Dunham et al., 2012). More specifically, the following three Tier 1 cell lines are used:

1. K562 immortalized cell line, coming from a human female with chronic myelogenous leukemia.
2. GM12878 lymphoblastoid cell line, produced from the blood of a female donor with northern and western European ancestry by EBV transformation.
3. H1-hESC embryonic stem cells, coming from a human male.

The RRBS data for all three cell lines are produced by the Myers Lab at HudsonAlpha Institute for Biotechnology (GEO: GSE27584). The data are already pre-processed and aligned to the *hg19* human reference genome, and can be downloaded from the web accessible database at UCSC. For our analysis, we use the resulting BED files and we ignore strand information. To obtain more accurate methylation level estimates, we pool together all available replicates. To investigate the correlation between DNA methylation profiles and transcript abundance, we use the corresponding paired-end RNA-seq data produced by Caltech (GEO: GSE33480). The RNA-seq data are pre-processed and mapped to the *hg19* human reference genome using TopHat and transcription quantification, in FPKM (Fragments Per Kilobase transcript per Million mapped reads), is produced using Cufflinks (Trapnell *et al.*, 2012). The RNA-seq data are filtered in order to keep only protein-coding genes.

To define promoter regions, we extract the TSS from the corresponding RNA-seq data, which are annotated based on both versions v3c and v4 of GENCODE GRCh37. Then, we consider N base pairs upstream and downstream from each TSS, resulting in promoter regions of length 2N base pairs. Since the cell lines are coming from different genders, the sex chromosomes are discarded from further analysis (see Section 4 of the Supplementary Material for a detailed description).

5 Results

5.1 Methylation profiles are highly correlated with gene expression

Initially, we examine whether gene expression levels might be predictable from DNA methylation patterns alone. We therefore extract higher-order features from promoter regions of $\pm 7kb$ around the TSS by learning the corresponding methylation profiles using the Binomial Probit Regression (BPR) observation model. To ensure that the promoter-proximal regions will have enough data to learn reasonable methylation profiles, we discard regions with less than 15 CpGs, and restrict our attention to regions which exhibit spatial variability in methylation levels. We applied the same pre-processing steps for the three ENCODE cell lines, which resulted in 7093 promoters for K562, 6022 for GM12878 and 5753 for H1-hESC cell line.

We model the methylation profiles using nine RBFs, which results in ten extracted features including the bias term. In addition to these features, we use the goodness of fit in RMSE and the CpG density across each region. We then train the SVM model on the resulting 12 features using a random subset of 70% of the promoter-proximal regions. We test the model's ability to quantitatively predict expression levels on the remaining 30% of the data. Our results show a striking improvement in prediction accuracy when compared to using the mean methylation level as input feature.

Figure 2A shows a scatter plot of the predicted and measured expression values for the K562 cell line, with Pearson's $r = 0.7$ (P-value of t-test $< 2.2\text{e}{-16}$) and RMSE $= 2.63$, demonstrating that the shape of methylation patterns across promoter-proximal regions is well correlated to mRNA abundance. Figure 2B shows the performance of the regression model when using the mean methylation level as input feature. It is evident that this approach cannot capture the diverse patterns present across the promoter regions, leading to poor prediction accuracy ($r = 0.31$ and RMSE $= 3.52$). Notice that the mean methylation approach erroneously predicts gene expression values only in the $(-2, 4)$ interval, whereas the BPR model captures more accurately the dynamic range of expression. Interestingly, the mean approach erroneously predicts the majority of genes to have expression value around -1, clearly indicating that summarizing DNA methylation by a single average is insufficient to capture the complex relationship with expression. Finally, one should observe the horizontal stripe around -3 on both figures: these are genes whose lack of expression cannot be attributed to

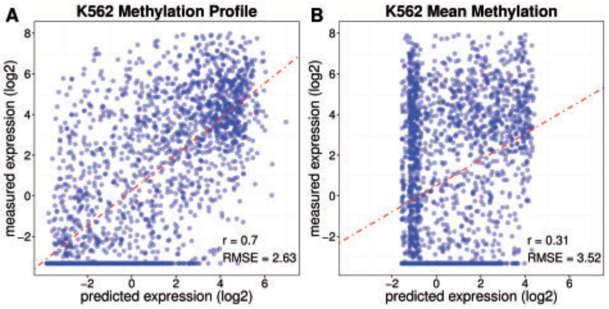

Fig. 2. Quantitative relationship between DNA methylation patterns and expression. (**A**) Scatter plot of predicted gene expression using the BPR model on the x-axis versus the measured (log-transformed) gene expression values for the K562 cell line on the y-axis. Each methylation profile is modeled using nine RBFs. In addition to these features, the SVM regression model uses as input the goodness of fit in RMSE and the CpG density. Each shaded blue dot represents a different gene and the darker the color, the higher the density of points. The red dashed line indicates the linear fit between the predicted and measured expression values, which are highly correlated (Pearson's $r = 0.7$, P-value $< 2.2\text{e}{-16}$), indicating a quantitative relationship between methylation profiles across promoter-proximal regions and transcript abundance. The model performance is also assessed by RMSE, which is 2.63. (**B**) Scatter plot of predicted and measured gene expression values when using the average methylation level as input feature in the SVM model; correlation has decreased substantially ($r = 0.31$ and RMSE $= 3.52$)

DNA methylation patterns, possibly implicating other regulating mechanisms (e.g. histone marks, binding of transcription factors, etc.), or difficulties in the measurement process of RNA-seq experiments (e.g. due to genes having relatively non-unique transcript sequences or multiple promoters).

We then consider the relative importance of the various features in predicting gene expression: in particular, we are interested in determining whether including goodness of fit or CpG density as covariates has any impact on predictive performance. For each cell line, we learn five SVM regression models, each having a different number of input features. The first four models consider as input the extracted higher-order methylation features with a combination of the two additional features we described in the previous section, whereas the last model takes the average methylation level as input feature. To statistically assess our results, we perform 20 random splits in training and test sets and evaluate the model performance on the corresponding test sets. Figure 3 shows boxplots of Pearson's r for the three ENCODE cell lines, where each boxplot indicates the performance of the prediction model on the 20 random splits of the data. The results demonstrate that by considering higher-order features we can build powerful predictive models of gene expression; and in the case of K562 and GM12878 we have more than 2-fold increase in correlation.

Concentrating on the importance of the additional features for the prediction process, we observe that the addition of CpG density does not have a significant prediction improvement compared to using only the shape of methylation profiles as input features (paired Wilcoxon test P-value = 0.22, 0.18 and 0.02 for K562, GM12878 and H1-hESC, respectively). On the other hand, the goodness of fit of the methylation profile in RMSE has a positive impact on the prediction performance (paired Wilcoxon test P-value = 4.8e−05, 4.8e−05 and 0.0001 for K562, GM12878 and H1-hESC, respectively). Finally, we explore the importance of considering different promoter region windows. Table 1 shows Pearson's r when considering various length promoter-proximal regions around the TSS. In general, the BPR model maintains its high predictive power across all cell lines for all different-length regions.

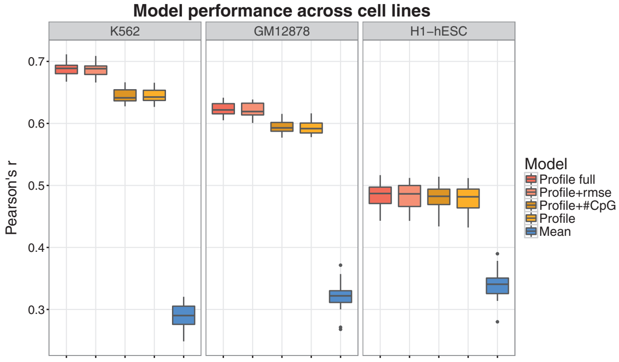

Fig. 3. Boxplot of Pearson's correlation coefficients r for the three ENCODE cell lines (K562, GM12878 and H1-hESC) with different input features for the SVM regression. The 'Profile full' model corresponds to the extracted BPR features plus the two additional features. Each boxplot indicates the performance using 20 random splits of the data in training and test sets. Paired Wilcoxon test shows that the high quantitative relationship between the shape of DNA methylation and expression exists in various cell lines, and is significantly better predictor than using the average methylation level (P-value = 8.4e−12). Regarding the two additional features, we observe that the goodness of fit measured in RMSE has a positive impact in correlation, whereas the CpG density does not improve the prediction performance. Paired Wilcoxon tests between K562 and other cell lines, show that K562 has significantly higher prediction accuracy (P-value = 4.8e−05 for both GM12878 and H1-hESC)

Table 1. Pearson's correlation coefficient r when considering different promoter region windows

Cell line	$\pm 2kb$	$\pm 3kb$	$\pm 4kb$	$\pm 5kb$	$\pm 6kb$	$\pm 7kb$	$\pm 8kb$	$\pm 9kb$
K562	0.63	0.69	0.69	0.67	0.67	**0.70**	0.67	0.67
GM12878	0.62	0.62	0.64	0.61	0.62	**0.61**	0.61	0.61
H1-hESC	0.46	0.49	0.48	0.43	0.49	**0.50**	0.47	0.49

For various length promoter-proximal regions, we show the performance (in Pearson's r) of methylation profiles in accurately predicting gene expression. The BPR model has high correlation across all different-length regions for all cell lines considered in this study. Bold values denote the selected promoter region window for demonstrating the results of this article.

5.2 Methylation profiles are predictive of gene expression across different ENCODE cell lines

We showed that gene expression is effectively predicted from the BPR model by using higher-order methylation features among various cell lines. Next, we further explore if the proposed model maintains predictive power across different cell lines. That is, we apply the regression model trained on one cell line to predict expression levels in another cell line, by using the learned methylation profiles in those cell lines as input features to the regression model. Figure 4A and B shows confusion matrices of correlation coefficients for the cross-cell line prediction process, using the BPR model and the mean methylation level approach, respectively. Figure 4C shows an example of applying the model learned from GM12878 methylation patterns to predict expression levels of the K562 cell line. The BPR model effectively predicts gene expression (r = 065 and 0.49 predicting K562 and H1-hESC, respectively), while, the mean methylation approach provides a poor estimate of correlation (r = 0.28 and 0.22 for predicting K562 and H1-hESC, respectively).

The results indicate that the quantitative relationship between DNA methylation profiles and mRNA abundance is not cell line specific, but that the model captures patterns of association between methylation and expression which hold across different cell lines. Although the proposed models have high prediction accuracy across all cell lines, the H1-hESC cell line shows consistently weaker correlations. This finding is in line with recent studies, reporting weaker correlations of gene expression and chromatin features for the H1-hESC cell line (Dong *et al.*, 2012), and with observations that mRNA-encoding genes in stem cells are transcriptionally paused during cell differentiation (Min *et al.*, 2011).

5.3 Clustering DNA methylation profiles across promoter-proximal regions

We next use the higher order methylation features to cluster DNA methylation patterns across promoter-proximal regions and examine whether distinct methylation patterns across different cell lines are associated to gene expression levels. We apply the same preprocessing steps described in the previous sections and we consider genomic regions of ±7kb around the TSS. We use the Bayesian Information Criterion (BIC) to set the number of clusters to five. We model the methylation profiles at a slightly lower spatial resolution, using four RBFs, as we are interested in capturing broader similarities between profiles, rather than fine details. Figure 5A shows the five distinct methylation profiles that were learned from each cell line after applying the EM algorithm. To investigate the association of promoter methylation profiles and transcription, in Figure 5B we show

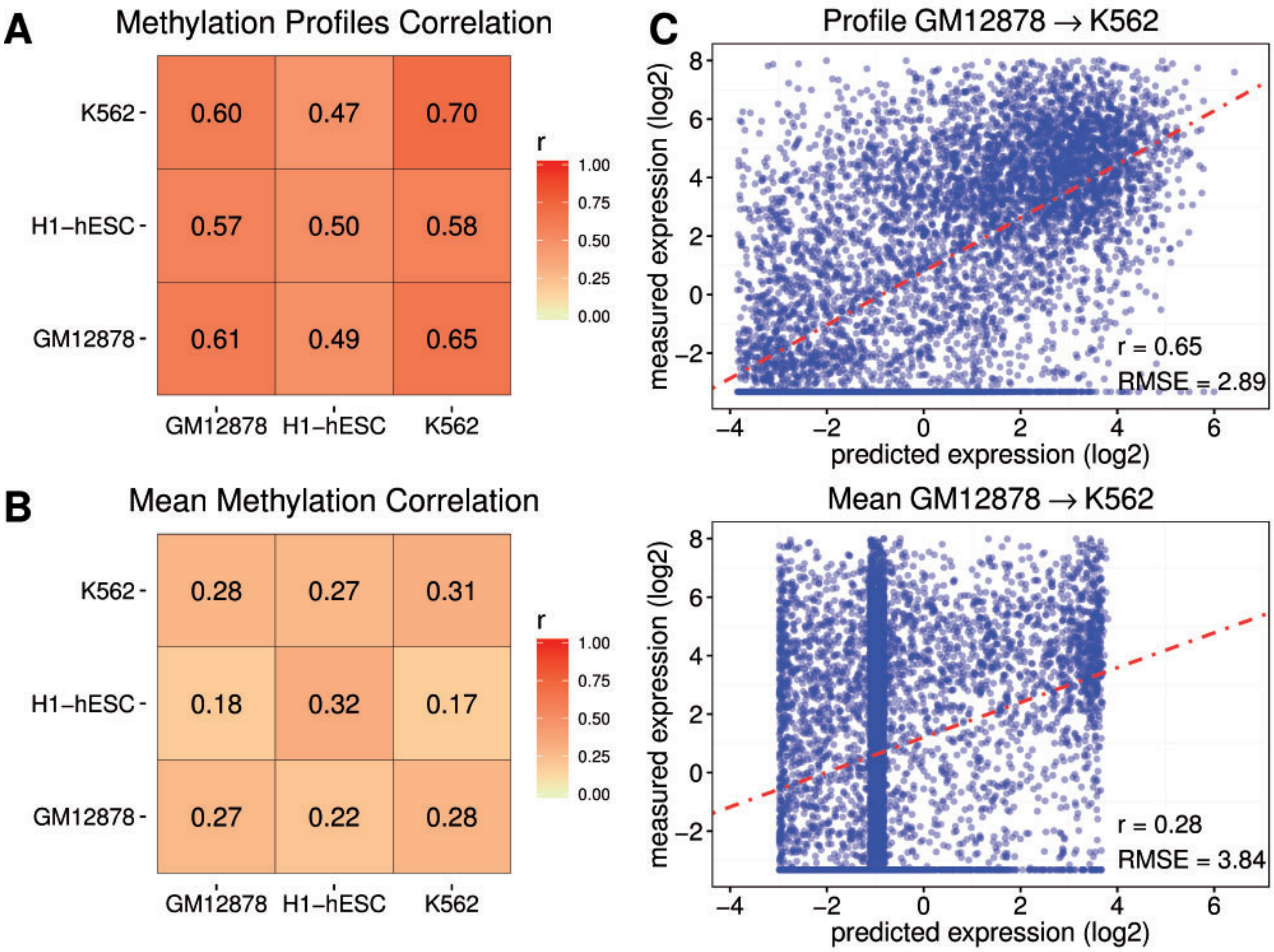

Fig. 4. Prediction accuracy across different cell lines. (**A**) Confusion matrix of Pearson's correlation coefficients *r* across cell-lines when using the BPR model with nine RBFs as input features to the regression model. Each (*i*, *j*) entry of the confusion matrix, corresponds to training a regression model from *i*th cell line and predicting gene expression levels for the *j*th cell line. The color of the confusion matrix corresponds to Pearson's *r* value, the darker the color the higher the correlation. (**B**) The corresponding correlation coefficients when using the mean methylation level as input feature to the regression model. Comparing both confusion matrices, it is evident that the methylation profile approach is more powerful in predicting expression levels across different cell lines. (**C**) Application of the model learned from GM12878 cell line to predict expression levels of the K562 cell line, using methylation profiles (top) and mean methylation levels (bottom) as input features

boxplots with the corresponding mRNA expression values that are assigned to each cluster for each cell line. From the resulting methylation profile clusters, we seek to characterize the common features that are responsible for the corresponding mRNA abundance.

As expected, clusters corresponding to hyper-methylated regions (Cluster 4, green) are associated with repressed genes across all cell lines, confirming the known relationship of DNA methylation around TSSs with gene repression. Also, two distinct patterns emerge: an S-shape profile (Cluster 5, yellow) with hypo-methylated CpGs upstream of TSS, which become gradually methylated at the gene body, and the reverse S-shape pattern (Cluster 3, orange). Genes associated with these profiles have intermediate expression levels for K562 and GM12878, and relatively high expression for H1-hESC. The most interesting pattern is the U-shape methylation profile (Cluster 2, blue), with a hypo-methylated region around the TSS surrounded by hyper-methylated domains. These profiles are associated with high transcriptional activity at their associated genes across all cell lines (t-test P-value $< 2.2\text{e}{-16}$ for all paired cluster comparisons across cell lines). Surprisingly, uniformly low-methylated domains (Cluster 1, red) seem in general to be repressed, except from the H1-hESC cell line, suggesting a different type of relationship between DNA methylation and expression in embryonic stem cells. The clustering analysis confirms, in a complementary way, that DNA methylation profiles and transcriptional process are tightly connected to each other, and this relationship can be generalized across all cell lines considered in this study.

To provide a biological insight on the potential methylation mechanisms that regulate transcription, we consider the purity of the clustering across different cell lines, i.e. which fraction of genes assigned to a certain cluster in a certain cell line are assigned to the same cluster in the other cell lines. Surprisingly, around 68% of the genes assigned to the U-shape profile are present in all three cell lines, while the intersection of genes assigned to the other clusters ranges between 20% and 40% (see Section 5 of the Supplementary Material). Interestingly, the promoter-proximal regions clustered to the U-shape methylation profile are dominated by CGIs. Of all common promoters assigned to the U-shape profiles, 95.6% are CGI associated. Not surprisingly, hyper-methylated promoters are only 35.7% CGI associated, however uniformly low-methylated promoters are 65.9% CGI associated. This suggests that promoters associated with totally unmethylated CGIs surrounded by hyper-methylated domains are transcriptionally active across cell lines.

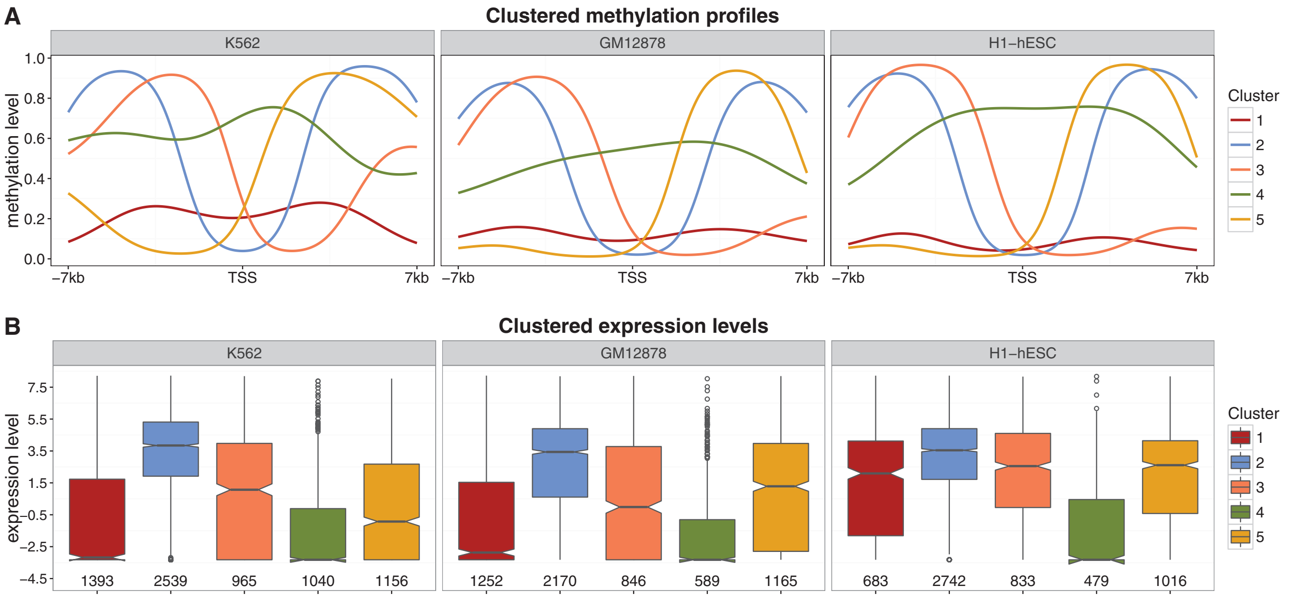

Fig. 5. Clustering DNA methylation profiles across promoter-proximal regions. (**A**) Five clustered methylation profiles over ±7 kb promoter region w.r.t. TSS in the direction of transcription for the three ENCODE cell lines (K562, GM12878 and H1-hESC). Each methylation profile is modeled using four RBFs. Comparing the clustered profiles it is evident that there are five prototypical methylation shapes across the cell lines. (**B**) Boxplots with the corresponding expression levels of the protein-coding genes assigned to each cluster for each of the three cell lines. The colors match with the clustered methylation profiles shown above. The numbers below each boxplot correspond to the total number of genes assigned to each cluster. *T*-test shows that the U-shape methylation profiles (Cluster 2, blue) correspond to significantly higher expression (*P*-value < 2.2e−16) compared to the expression of genes assigned to the remaining methylation profiles

Indeed, we find that 35% of the U-shape profile genes are associated with a curated set of housekeeping genes (Eisenberg and Levanon, 2013). In contrast, only a small fraction of genes assigned to hyper-methylated domains or uniformly low-methylated domains are housekeeping genes (1.4 and 17.7%, respectively). Finally, around 22% of the genes assigned to the S-shape and reverse S-shape profiles are associated with housekeeping genes.

6 Discussion

Alterations in DNA methylation are associated with regulatory roles and are involved in many diseases, most notably cancer (Baylin and Jones, 2011). Therefore, unraveling the function of DNA methylation and its relationship to transcription, is essential for understanding biological processes and developing biomarkers for disease diagnostics (Laird, 2003).

Our results demonstrate that representing methylation patterns by their average level is insufficient to understand the link between DNA methylation and expression, and one should consider the shape of the methylation profiles at the vicinity of the promoters. The contributions of this paper are twofold. First, we introduced a generic modeling approach to quantitate spatially distributed methylation profiles via the BPR model. The BPR features enabled us to build a powerful predictive model for gene expression in various cell lines which more than doubled the predictive accuracy of current methods based on average methylation levels.

Second, we have shown how the BPR features can be used in downstream analyses by clustering spatially similar methylation profiles. We revealed five distinct groups of methylation patterns across promoter regions that are well correlated with gene expression and are well reproducible across different cell lines. Some of these patterns recapitulate existing biological knowledge. The U-shape methylation profile, consisting of hypo-methylated CGIs followed by hyper-methylated CGI shores, has been identified in different studies, and is termed as ‘canyon’ (Jeong *et al.*, 2014) or ‘ravine’ (Edgar *et al.*, 2014). Our findings are in line with Edgar *et al.* (2014), where ravines are in general positively correlated with mRNA abundance. Since, the main difference of the U-shape methylation profile and the uniformly low-methylated profile is the CGI shore methylation, our results support the hypothesis that hyper-methylation on the edges of CGIs enhances transcriptional activity.

The existence of U-shape methylation profiles may help to explain observations that the methylation of gene body was sometimes positively correlated with transcript abundance (Lou *et al.*, 2014; Varley *et al.*, 2013). We hypothesize that these regions may correspond to U-shape methylation profiles, or a mixture of U-shape and S-shape methylation profiles. Another relevant study, showed that hyper-methylation of CGI shores on the mouse genome was associated with increased DNMT3A activity, which resulted in positive correlation with transcriptional activity; indicating that methylation outside of CGIs may be used for maintaining active chromatin states for specific genes (Wu *et al.*, 2010).

In this study, we focused on RRBS data, however, given the considerable robustness of the BPR model to low coverage, we expect that it may also be well suited for Whole Genome Bisulphite Sequencing data, which have the advantage of providing a more comprehensive coverage of CpG sites genome-wide. As an extension of this analysis, further work could include building a model to relate differential methylation profiles with differential gene expression levels, and evaluate the importance of profile changes in regulation of gene expression across different cell types. More generally, it is increasingly clear that transcriptional activity is regulated by a complex and still incompletely understood interaction network of molecular players, including DNA methylation, histone modifications and transcription factor binding. Several recent computational studies have highlighted the dependencies between these players (Benveniste *et al.*, 2014; Dong *et al.*, 2012). The BPR model provides an effective way of recapitulating DNA methylation patterns using higher order features, and may therefore play an important role in building more effective integrative models of high-throughput data.

Acknowledgements

We thank Duncan Sproul for valuable comments and discussion.

Funding

This study was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1 to C.A.K.) and the University of Edinburgh and the European Research Council through grant (MLCS306999 to GS).

Conflict of Interest: none declared.

References

Baylin,S.B. and Jones,P.A. (2011) A decade of exploring the cancer epigenome—biological and translational implications. *Nat. Rev. Cancer*, **11**, 726–734.

Benveniste,D. *et al*. (2014) Transcription factor binding predicts histone modifications in human cell lines. *Proc. Natl. Acad. Sci. USA*, **111**, 13367–13372.

Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*., **16**, 6–21.

Bock,C. *et al*. (2012) DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Mol. Cell*, **47**, 633–647.

Deaton,A. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev*., **25**, 1010–1022.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol*., **39**, 1–38.

Dong,X. *et al*. (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol*., **13**, R53.

Dunham,I., *et al*. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Edgar,R. *et al*. (2014) Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. *Epigenetics Chromatin*, **7**, 28.

Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet*., **29**, 569–574.

Feinberg,A.P. and Vogelstein,B. (1983) Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, **301**, 89–92.

Friedman,J.H. (1991) Multivariate adaptive regression splines. *Ann. Stat*., **19**, 1–67.

Hansen,K.D. *et al*. (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet*., **43**, 768–775.

Hansen,K.D. *et al*. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*., **13**, R83.

Hebestreit,K. *et al*. (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, **29**, 1647–1653.

Jeong,M. *et al*. (2014) Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat. Genet*., **46**, 17–23.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet*., **13**, 484–492.

Krueger,F. *et al*. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.

Laird,P.W. (2003) The power and the promise of DNA methylation markers. *Nat. Rev. Cancer*, **3**, 253–266.

Li,E. *et al*. (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Lou,S. *et al*. (2014) Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol*., **15**, 408.

Mayo,T.R. *et al*. (2015) M 3 D: A kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, **31**, 809–816.

McLachlan,G. and Peel,D. (2004). *Finite Mixture Models*. John Wiley & Sons, Hoboken.

Meissner,A. *et al*. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*., **33**, 5868–5877.

Min,I.M. *et al*. (2011) Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev*., **25**, 742–754.

Trapnell,C. *et al*. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*., **7**, 562–578.

Vanderkraats,N.D. *et al*. (2013) Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Res*., **41**, 6816–6827.

Varley,K.E. *et al*. (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res*., **23**, 555–567.

Wu,C.J. (1983) On the convergence properties of the EM algorithm. *Ann. Stat*., **11**, 95–103.

Wu,H. *et al*., (2010) Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science*, **329**, 444–448.