LuxUS: Detecting differential DNA methylation using generalized linear mixed model with spatial correlation structure

Viivi Halla-aho
Department of Computer Science
Aalto University School of Science
Espoo, Finland
viivi.halla-aho@aalto.fi

Harri Lähdesmäki
Department of Computer Science
Aalto University School of Science
Espoo, Finland
harri.lahdesmaki@aalto.fi

January 31, 2019

Abstract

DNA methylation is an important epigenetic modification, which has multiple functions. DNA methylation and its connections to diseases have been extensively studied in recent years. It is known that DNA methylation levels of neighboring cytosines are correlated and that differential DNA methylation typically occurs rather as regions instead of individual cytosine level.
We make use of the correlation between neighboring cytosines through a generalized linear mixed model to facilitate analysis of differential methylation where the experimental design can be taken into account and where both binary and continuous covariates can be included in the analysis. We show with simulation experiments that utilizing the spatial correlation we gain more power to the statistical testing of differential DNA methylation. Results with real bisulfite sequencing data set show that LuxUS is able to detect biologically significant differentially methylated cytosines. The tool is available at `https://github.com/hallav/LuxUS`.

1 Introduction

DNA methylation is a widely studied epigenetic modification, which is involved in gene regulation. Aberrant methylation states have been shown to be connected with diseases and cancer. One popular method for quantifying DNA methylation level is bisulfite sequencing (BS-seq) and its variants. Along with the development of measuring techniques, several tools have been proposed for analysis of differential methylation from BS-seq data. These methods aim to model the underlying methylation proportions and perform statistical tests to assess the statistical significance of the effect of a covariate of interest on the methylation level.

Spatial correlation of cytosines' methylation states is a widely known phenomenon and methylation state is generally thought to vary rather as regions instead of individual cytosine level. Consequently, many DNA methylation analysis tools have attempted to capture the phenomenon with different model structures and computational techniques. For example popular tool RadMeth by Dolzhenko and Smith (2014) gives as a result differentially methylated regions (DMRs) by combining neighboring differentially methylated cytosines based on their p-values, which have been first calculated for individual loci. Similarly, Wen *et al*. (2016) first use beta-binomial regression model to calculate p-values for individual CpG sites and then combine them using Getis-Ord statistic. BSmooth tool by Hansen *et al*. (2012) is based on local-likelihood smoothing, which can compensate for low-coverage data but models transformed count data using Gaussian approximation. Recently published dmrseq tool (Korthauer *et al*., 2018) combined smoothing and covariance structure with weighted least squares regression. Mayo *et al*. (2014) proposed a kernel based method where statistical testing of putative DMRs is done through maximum mean discrepancy values. Rackham *et al*. (2017) take a different approach with their tool ABBA, which uses a latent Gaussian model.

The LuxGLM model was proposed by Äijo *et al*. (2016) and it enabled testing for differential methylation while taking into account different methylcytosine species and experimental design through generalized linear model part. LuxGLM also takes into account the experimental parameters, such as bisulfite conversion efficiency. In (Äijo *et al*., 2016) the tool was shown to be accurate and perform on par or even better than other recent methods in detecting differential methylation. Here we propose a method where the methylation proportions are modeled with a generalized linear mixed model and simultaneously we can take into account both binary and continuous variables and random effects in the analysis. The spatial correlation is included in the model through the covariance structure of the cytosine random effect in the generalized linear mixed model.

2 Methods

The observation model and the generalized linear model parts of LuxGLM were used as a basis for this new method *Lux*GLM *U*sing *S*patial correlation (LuxUS). Here we present the model which only handles unmethylated and 5mC states, but the model could be further expanded to take into account oxidized DNA methylation species. In this work we use Bayesian approach and probabilistic programming language Stan for implementing the model and estimating the model parameters. Hypothesis testing for the differential methylation analysis is done using Bayes factors.

2.1 A generalized linear mixed effect model with spatial correlation

The plate diagram for the LuxUS model is presented in Fig. 1. From bisulfite sequencing data we can retrieve for each cytosine in each sample the number of total reads overlapping the corresponding cytosine $\mathbf{N}_{\text{BS}}$ out of which $\mathbf{N}_{\text{BS,C}}$ observations were cytosines. Recall that observing a C read in a BS-seq experiment means that the cytosine was (most likely) in a methylated state, whereas non-methylated cytosines are (most likely) observed as thymine. We model a genomic window of N_C cytosines at a time and the genomic coordinates of the cytosines are stored in vector $\boldsymbol{c}$ of length N_C. The cytosines in the genomic window have to come from the same chromosome, as otherwise it would not be possible to calculate linear genomic distance between the cytosines.

There are in total N_R replicates and there are N_P covariates considered in an experiment. The design matrix $\mathbf{X}$ has size $(N_C \cdot N_R) \times N_P$ and contains the experimental design. As stated above, there are in total N_R replicates and there are N_P covariates considered in the experiment. The $N_R \times N_p$ design matrix $\mathbf{D}$ for one cytosine is used as a block matrix in the diagonal of $\mathbf{X}$ which is the design matrix for the whole window

$$\mathbf{X} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D} \end{pmatrix}. \tag{1}$$

Generalized linear mixed model formulation is used to write the model. In mixed linear models we can have both fixed and random effects. Here we assume that the effects of the covariates in X are fixed, but each of the replicates and

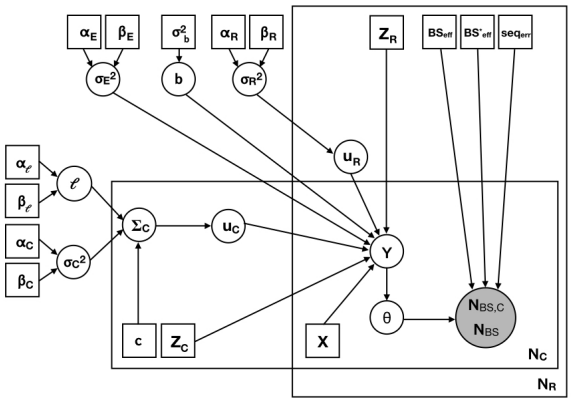

Figure 1: Plate diagram of the LuxUS model. The grey and white circles represent observed and latent variables respectively. Rectangles represent fixed hyperparameters, design matrices or other type of input data.

cytosines have their own random intercept terms. These random terms are expressed as vectors

$$\mathbf{u}_R \sim N(\mathbf{0}, \sigma_R^2 \mathbf{I}), \tag{2}$$

and

$$\mathbf{u}_C \sim N(\mathbf{0}, \mathbf{\Sigma}_C), \tag{3}$$

which are vectors of length N_R and N_C, respectively. Random term $\mathbf{u}_R$ has a diagonal covariance matrix with σ_R^2 as the diagonal term and $\mathbf{u}_C$ has a covariance matrix $\mathbf{\Sigma}_C$ which includes the spatial correlation of the model. The covariance matrix $\mathbf{\Sigma}_C$ is defined as

$$\mathbf{\Sigma}_C = \begin{pmatrix} \sigma_C^2 & \mathrm{cov}(u_{C_1}, u_{C_2}) & \cdots & \mathrm{cov}(u_{C_1}, u_{C_{N_C}}) \\ \mathrm{cov}(u_{C_1}, u_{C_2}) & \sigma_C^2 & \cdots & \mathrm{cov}(u_{C_2}, u_{C_{N_C}}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(u_{C_1}, u_{C_{N_C}}) & \mathrm{cov}(u_{C_2}, u_{C_{N_C}}) & \cdots & \sigma_C^2 \end{pmatrix}, \tag{4}$$

where the covariance terms between the elements of $\mathbf{u}_C$ are defined as

$$\mathrm{cov}(u_{C_i}, u_{C_j}) = \sigma_C^2 \cdot \exp\left(\frac{-|c_i - c_j|}{\ell^2}\right), \tag{5}$$

where $i = 1, \ldots, N_C$, $j = 1, \ldots, N_C$, and where ℓ is the length-scale parameter, which has a prior

$$\ell \sim \mathrm{Gamma}(\alpha_\ell, \beta_\ell). \tag{6}$$

The hyperparameters α_ℓ and β_ℓ are set to 38 and 1 respectively. With the selection of the hyperprior parameters we can adjust the covariance term in Eq. (5). By setting $\alpha_\ell = 38$ and $\beta_\ell = 1$ the shape of the covariance term in Eq. (5) as function of distance $|c_i - c_j|$ is similar to the methylation correlation plots as function of genomic distance between cytosines presented in (Eckhardt *et al*., 2006; Song *et al*., 2017).

The variance parameters σ_C^2 and σ_R^2 have gamma priors

$$\sigma_C^2 \sim \mathrm{Gamma}(\alpha_C, \beta_C), \tag{7}$$

and

$$\sigma_R^2 \sim \mathrm{Gamma}(\alpha_R, \beta_R), \tag{8}$$

where α_R, β_R, α_C and β_C are set to make the prior distribution match our prior information about the cytosine and replicate random effects.

We also define random effects design matrices $\mathbf{Z}_C$ of size $(N_C \cdot N_R) \times N_C$ and $\mathbf{Z}_R$ of size $(N_R \cdot N_C) \times N_R$ which indicate from which cytosine and replicate each observation is coming from respectively.

Then vector $\mathbf{Y}$ of length $N_R \cdot N_C$ can be expressed as a sum of fixed and random effects

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_R \mathbf{u}_R + \mathbf{Z}_C \mathbf{u}_C + \mathbf{e}, \tag{9}$$

where $\mathbf{e}$ is a residual noise term with distribution

$$\mathbf{e} \sim N(0, \sigma_E^2 \mathbf{I}). \tag{10}$$

The prior for σ_E^2 is

$$\sigma_E^2 \sim \mathrm{Gamma}(\alpha_E, \beta_E), \tag{11}$$

where α_E and β_E are the shape and rate parameters respectively.

The fixed effect coefficients vector $\mathbf{b}$ has a normal prior

$$\mathbf{b} \sim N(\mathbf{0}, \sigma_b^2 \mathbf{I}), \tag{12}$$

where the prior variance σ_b^2 should be set to high enough value to enable large enough variation for the fixed effect coefficients.

The distribution of $\mathbf{Y}$ can be expressed as

$$\mathbf{Y} \sim N(\mathbf{X}\mathbf{b}, \mathbf{\Sigma}_Y), \tag{13}$$

where the elements of the covariance matrix $\mathbf{\Sigma}_Y$ are defined as

$$\mathbf{\Sigma}_{Y_{i,j}} = \begin{cases} \sigma_C^2 \exp\left(\frac{-|c_i - c_j|}{\ell^2}\right) & \text{, if } Z_{R_i} \neq Z_{R_j} \\ \sigma_R^2 + \sigma_C^2 \exp\left(\frac{-|c_i - c_j|}{\ell^2}\right) & \text{, if } Z_{R_i} = Z_{R_j} \text{ and } Z_{C_i} \neq Z_{C_j} \\ \sigma_R^2 + \sigma_C^2 + \sigma_E^2 & \text{, if } i = j, \end{cases}$$

where $i = 1, ..., N_R \cdot N_C$ and $j = 1, ..., N_R \cdot N_C$ and Z_{R_i} and Z_{C_i} denote the ith rows of random effect design matrices $\mathbf{Z}_R$ and $\mathbf{Z}_C$ respectively.

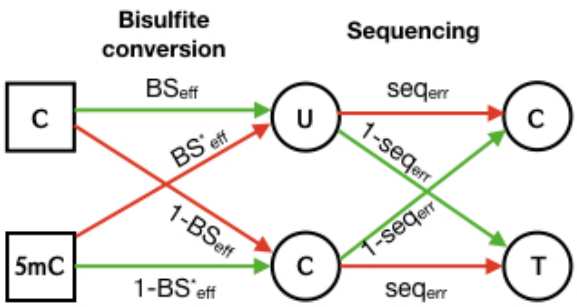

Figure 2: The probability tree for observing a C or T in sequencing data when the true methylation state is methylated or unmethylated.

2.2 An observation model for bisulfite sequencing data

The generalized linear mixed model connects the effects of the covariates to the methylation proportions vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{N_R \cdot N_C})^T$, which is retrieved from $\mathbf{Y}$ through sigmoid link function

$$\boldsymbol{\theta} = \frac{1}{1 + \exp(-\mathbf{Y})}. \tag{14}$$

The methylation proportions vector $\boldsymbol{\theta}$ can then be used for calculating the probabilities for observing a cytosine in a bisulfite sequencing experiment, $\mathbf{p}_{\text{BS,C}}$. Finally, the probability of observing $\mathbf{N}_{\text{BS,C}}$ many cytosine reads out of total read count $\mathbf{N}_{\text{BS}}$ in a BS-seq experiment is modelled with binomial distribution $\mathbf{p}_{\text{BS,C}}$

$$\mathbf{N}_{\text{BS,C}} \sim \text{Binomial}(\mathbf{N}_{\text{BS}}, \mathbf{p}_{\text{BS,C}}), \tag{15}$$

where probabilities $p_{\text{BS,C},i}$, $i = 1, ..., N_C \cdot N_R$ for each observation are calculated as

$$\begin{aligned} p_{\text{BS,C},i} = \theta_i((1 - \text{bs}_{\text{Eff},i})(1 - \text{seq}_{\text{Err},i}) \\ + \text{bs}_{\text{Eff},i} \cdot \text{seq}_{\text{Err},i}) \\ + (1 - \theta_i)((1 - \text{bs}^*_{\text{Eff},i})(1 - \text{seq}_{\text{Err},i}) \\ + \text{bs}^*_{\text{Eff},i} \cdot \text{seq}_{\text{Err},i}), \end{aligned} \tag{16}$$

following the probability tree diagram in Fig. 2. The $\text{seq}_{\text{Err},i}$, $\text{bs}_{\text{Eff},i}$ and $\text{bs}^*_{\text{Eff},i}$ are experimental parameters, which describe the probabilities of sequencing error, bisulfite conversion and incorrect bisulfite conversion, respectively, for the sample corresponding to the ith observation. It is often thought that samples with low bisulfite conversion efficiency should be excluded from analyses, but as the experimental parameters can be taken into account with LuxUS, one does not have to leave out such samples from the analysis as long as the conversion efficiencies can be reliably estimated. As for the samples with low conversion efficiencies the experiments are often redone to reach acceptable experiment quality, including the experimental parameters in the model reduces the need for redoing experiments. The experimental parameters must be estimated for each sample e.g. from the lambda phage genome spike-in and given to the model along with other input data.

2.3 Model estimation

To make inference based on the model, we have to estimate the unknown parameters. The variables to be estimated are $\mathbf{b}$, ℓ, σ_R^2, σ_C^2, σ_E^2, $\mathbf{u}_R$, $\mathbf{u}_C$ and $\mathbf{Y}$, from which $\boldsymbol{\theta}$ can be calculated.

The model was implemented with probabilistic programming language Stan (Carpenter *et al.*, 2017) and the model parameters are estimated using the Hamiltonian Monte Carlo (HMC) sampling, implemented in Stan. The PyStan version (Stan Development Team, 2017) of the software was used for the HMC sampling. Stan has also built-in automatic differentiation variational inference (ADVI) feature (Kucukelbir *et al.*, 2015), which was also tested for estimating the model parameters. As variational inference approaches are often faster than Markov chain Monte Carlo (MCMC) methods, it could be more favourable model estimation method especially when analysing large reduced representation BS-seq (RRBS-seq) or whole genome BS-seq (WGBS-seq) data sets. For running ADVI we utilize CmdStan (Stan Development Team, 2018), which is the command line interface to Stan.

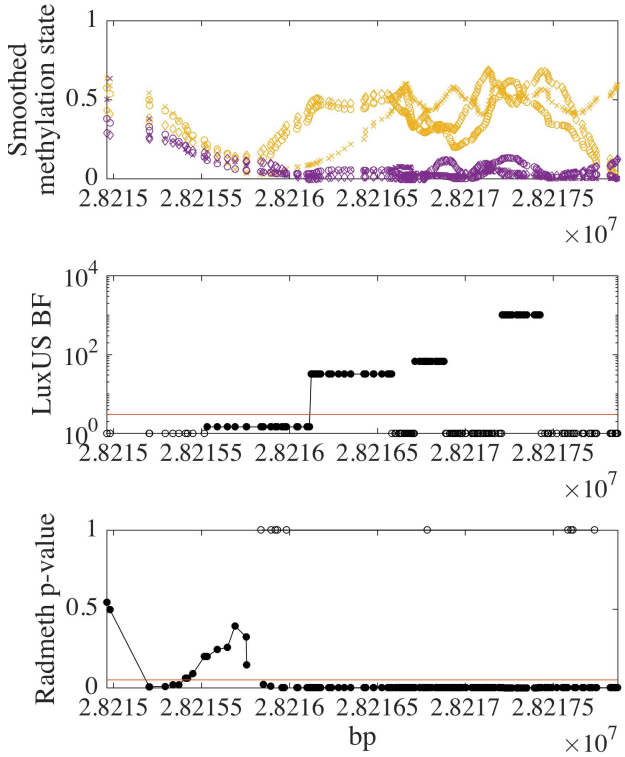

Figure 3: The top panel shows the smoothed methylation states for a chosen genomic region in chromosome 21 (from 28214962 bp to 28217867 bp) for the six samples in the Hansen data. The colon cancer sample methylation states have been plotted in orange and and the normal colon samples in purple. The experiment was paired, and the cancer and colon samples from the same person have been plotted with the same marker symbols. The middle panel shows the LuxUS Bayes factor values with log scaling. The Bayes factor values have been truncated to 1000. The cytosines which were filtered from further analysis in the preanalysis phase have been plotted with value 0 with empty markers. The horizontal red line corresponds to the Bayes factor value 3. The bottom panel shows the RadMeth p-values for the same genomic region. The horizontal red line corresponds to the adjusted p-value of 0.05. The cytosines for which RadMeth produced no p-value are plotted with p-value of 1 with empty markers.

2.4 Testing for differential methylation

After estimating the model parameters as described above, test for differential methylation can be performed. With the models described above it is possible to perform two types of tests. The first test has null hypothesis $H_0 : \mathbf{b}_i = 0$ and alternative hypothesis $H_1 : \mathbf{b}_i \neq 0$, i.e. the statistical significance of the covariate i is tested. The second test has null hypothesis $H_0 : \mathbf{b}_i - \mathbf{b}_j = 0$ and alternative hypothesis $H_1 : \mathbf{b}_i - \mathbf{b}_j \neq 0$, i.e. the difference between the effects of covariates i and j is tested for statistical significance. The testing is done using Bayes factors. As calculating the exact value for Bayes factors is often infeasible, instead a Savage-Dickey density ratio estimate of the Bayes factor is used. For the first test the Savage-Dickey estimate is

$$BF \approx \frac{p(\mathbf{b}_i = 0|H_1)}{p(\mathbf{b}_i = 0|H_1, \mathcal{D})}, \tag{17}$$

where we denote data with $\mathcal{D}$. Similarly for the second test

$$BF \approx \frac{p(\mathbf{b}_i - \mathbf{b}_j = 0|H_1)}{p(\mathbf{b}_i - \mathbf{b}_j = 0|H_1, \mathcal{D})}. \tag{18}$$

The numerator of the Savage-Dickey density ratio is analytically solvable as we use normal prior for $\mathbf{b}$. As for the denominator, it can be approximated from the posterior samples for $\mathbf{b}_i$ by evaluating a kernel density estimate of the posterior at origin. The Gaussian kernel density estimation function from statistical functions module from the SciPy

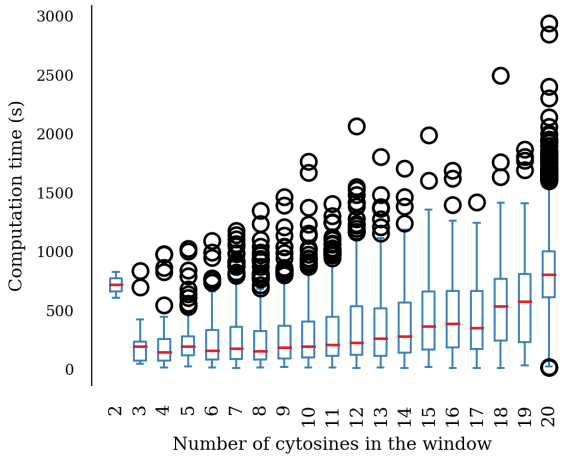

Figure 4: Boxplots of the computation times for the genomic windows for the Hansen data. One boxplot represents all the windows with the same number of cytosines in the genomic window.

package is used to calculate the kernel density estimates. The Bayes factor is calculated for the whole window being analyzed, as the linear model coefficients $\mathbf{b}$ are shared for all cytosines.

2.5 Preanalysis for filtering the genomic windows

As stated above, our model calculates the Bayes factor for a genomic window at a time. To prepare a BS-seq experiment data set for analysis, we have provided a script which also includes a simple preanalysis filtering method. The genomic windows, for which the analysis is performed, can be determined using a fixed window size (in terms of number of nucleotides) or number of cytosines in a window. The prefiltering step is computationally efficient and can be used to filter away regions which e.g. do not contain sufficient number of reads or are unlikely to exhibit differential DNA methylation. A coverage limit can be used to filter out cytosines with low coverages. As estimating the model for Bayes factor calculation can be computationally burdensome, it might be desirable to filter out the unpromising windows using the preanalysis method. We have implemented a simple F-test for testing the significance of a variable of interest using the logit transformed sample means (calculated over all the cytosines in the window) of the methylation states. The mean methylation states are truncated to range $[0.00001, 0.9999]$ to enable the logit transformation. The p-value limit for accepting a window for further analysis with the LuxUS model can be set as desired by the user. Each genomic window can then be analysed separately, enabling parallellisation.

3 Results

3.1 Analysis of colon cancer WGBS-seq data

The model was tested using a WGBS-seq data set by Hansen (2018), consisting of matched human colon and colon cancer samples. The processed sequencing data was provided for two chromosomes, 21 and 22. First, the genomic windows for LuxUS analysis were determined using the preanalysis method and the data for the resulting windows was stored as Stan input files. The p-value cutoff for the preanalysis F-test for the significance of the cancer covariate was set to 0.1. The p-value cutoff value was chosen so that it would not be too conservative and filter out too many potentially differentially methylated cytosines. For a cytosine to be added in a genomic window it had to have at least coverage of one in at least one sample from both case and control groups. Additionally, the mean coverage over the genomic window had to be at least 5 at least in one sample from both case and control groups for the genomic window to be accepted for analysis. The maximum number of cytosines in a genomic window was set to 20 and the maximum length of the genomic window was set to 2000 basepairs. This resulted in 9945 genomic windows that passed the preanalysis phase. The hyperparameters for the random effect and noice variances were set to $\alpha_R = 2$, $\beta_R = 2$, $\alpha_C = 2$, $\beta_C = 2$, $\alpha_E = 2$ and $\beta_E = 2$. The variance parameter for the coefficients $\mathbf{b}$, σ_b^2, was set to 15. The bisulfite conversion efficiencies $\mathrm{BS}_{\mathrm{eff}}$ were set to the mean of the conversion rates over the two flowcells that were used per sample, presented in Hansen *et al*. (2011). As the rest of the experimental parameters for this data set were not known, the parameters were set to

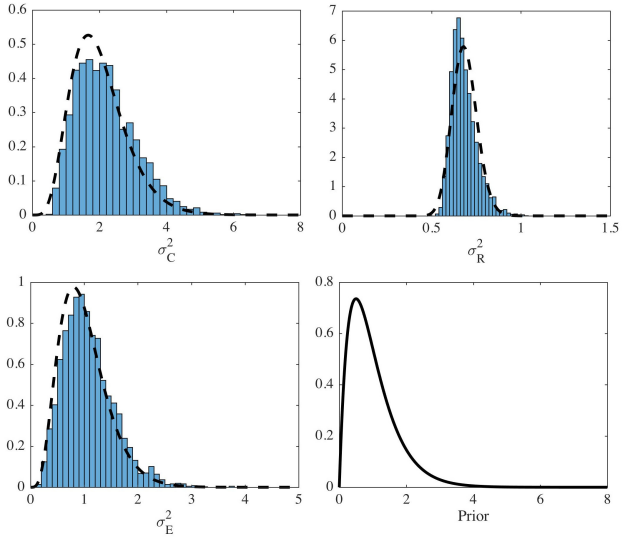

Figure 5: Histograms of the sample means for the parameters σ_C^2, σ_R^2 and σ_E^2 for the Hansen data and the probability density function of the prior Gamma(2,2). The probability density functions for the estimated gamma distributions for the variance parameters have been plotted with dashed line on top of the histograms.

$\text{BS}^*_{\text{eff}} = 0$ and $\text{seq}_{\text{err}} = 0$ for each sample, which correspond to perfect incorrect bisulfite conversion efficiency and zero sequencing error. Then Stan was run with HMC sampling to fit the LuxUS model parameters and Bayes factors were calculated. As a result 1422 genomic windows were discovered with Bayes factor ≥ 3. These genomic windows covered in total 21127 cytosines.

In comparison, we performed RadMeth analysis on the same data set and discovered 137142 cytosines with adjusted p-value ≤ 0.05. The overlap of the significant cytosines from LuxUS and RadMeth approaches is 15008 cytosines. So about three-fourths of the cytosines found by LuxUS are also found by RadMeth. The results for a genomic region in chromosome 21 are shown in Fig. 3, which demonstrates that both LuxUS and RadMeth produce similar results, but LuxUS is more conservative. For example, near the end of the genomic region in Fig. 3 the difference in average methylation state between the cancer and normal colon cells decreases, which RadMeth fails to detect and produces p-values smaller than 0.05. Whereas, the said region was filtered from the analysis by LuxUS already in the preanalysis phase. The boxplots of the computation times (on one standard computing node in a cluster) for each number of cytosines in a genomic window for LuxUS are shown in Fig. 4. Running RadMeth for the whole dataset took 31 minutes and 33 seconds.

To investigate the biological significance of the found differentially methylated cytosines, we used the GREAT tool (McLean *et al*., 2010) for gene set enrichment analysis (online version 3.0.0 with the default parameter settings). The lists of significantly differentially methylated cytosines found by LuxUS and RadMeth were given to the tool as inputs and the unfiltered list of CpG sites for the chromosomes 21 and 22 was used as background regions file. The results from both methods showed enrichment in cancer related terms in Disease Ontology and Molecular Signatures Database (MSigDB) Perturbation ontology, which indicates that the methods were able to discover truly differentially methylated cytosines. The lists of the enriched terms for LuxUS and RadMeth can be found from Supplementary Fig S1-S5.

The analysis of the colon cancer WGBS-seq data was also used to set certain parameters for the simulation experiments to realistic values. Gamma(2,2) prior was used for the variance parameters σ_C^2, σ_R^2 and σ_E^2 when estimating the model for the Hansen data. The means of the posterior samples for these parameters across all windows were stored and their distributions can be seen in Fig. 5. Gamma densities were fitted to the sample means using gamfit function from Matlab. For this purpose only the genomic windows with 10 or more cytosines and with minimum mean sample coverage of 5 were considered. The resulting shape and scale parameters for the gamma distributions were approximately $(6, 1/3)$, $(98, 1/143)$ and $(5, 1/5)$, respectively, for σ_C^2, σ_R^2 and σ_E^2.

The average number of cytosines in a genomic window for the Hansen data was also explored. The windows of 2000 bp were first formed with only filtering the cytosines so that there must be at least one sample present from both case and control groups for a cytosine to be included in a genomic window. This resulted in 16210 windows in chromosome 21 and 16502 windows in chromosome 22, and the distribution of number of cytosines in these windows is shown in Supplementary Fig. S6. The mean of the number of cytosines in the 2000 basepair windows is 22.8 and the mode is 12.

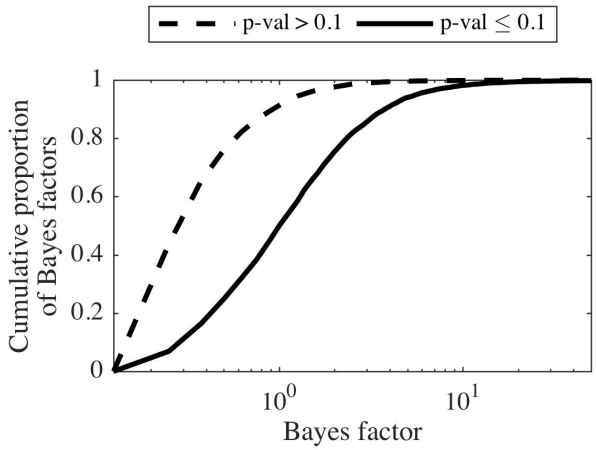

Figure 6: Cumulative proportion of Bayes factors as a function of Bayes factor value for the Hansen data set. Logarithmic scale is used for the x-axis. In this figure the cumulative proportions are shown up to Bayes factor value 50. The dashed line represents the Bayes factors for the randomly picked 10000 genomic windows that had p-value > 0.1 and the solid line represents the genomic windows with p-value ≤ 0.1 in the preanalysis test. The figure shows that the proportion of small Bayes factors is greater for the cytosines with greater p-values in the preanalysis phase.

Based on this result we decided to use cytosine frequency of 10 cytosines in 1000 basepair window for the simulation experiments.

To validate the preanalysis step with the simple F-test, we compared the Bayes factor values for the 9945 genomic windows that were accepted by the preanalysis step to the Bayes factor values for 10000 randomly chosen genomic windows which had preanalysis F-test p-value > 0.1. The same coverage requirements were applied to the both sets of genomic windows. In Fig. 6 the cumulative proportion of Bayes factors is plotted as the function of Bayes factor value, which has been plotted in logarithmic scale. The proportion of small Bayes factors is considerably higher amongst the genomic windows with F-test p-value> 0.1 than for the genomic windows which passed the preanalysis phase. We conclude that using the preanalysis step lead to discarding only a few genomic windows with high Bayes factor values.

3.2 Performance comparisons on simulated BS-seq data

Simulated BS-seq data was generated for performance comparison purposes. The simulations were made using the LuxUS model. First, 100 sets of experimental parameters and cytosine locations were generated. The number of cytosines was 10 and their locations were randomly chosen from range from 0 to 1000. Two BS-seq data sets were then produced for each of these 100 sets: one with differential methylation and one with no differential methylation. The used values for the fixed effect coefficient means, $\boldsymbol{\mu}_b = (\mu_{B,0}, \mu_{B,1})$, for the cases with differential methylation were $(-1.4, 1)$, $(-1.4, 2.3)$, $(1.4, -1)$ and $(1.4, -2.3)$. These values correspond to approximate values of 0.2, 0.5, 0.2 and 0.5 of $\Delta\theta$, the difference in methylation proportions between the case and control groups, respectively. The variance for the fixed effect coefficients, σ_b^2, was set to 0.25 for the simulations. The variances for the cytosine and replicate random effects and the noise term were set to the means of the gamma distributions which were fitted to the posterior samples from the colon cancer WGBS-seq data analysis. The values used for number of reads, i.e. coverage, and number of replicates were 6, 12 and 24. The resulting simulated data sets were saved as Stan input files. The priors for σ_C^2, σ_R^2 and σ_E^2 were set using the results from the real data analysis and σ_b^2 is set to 15 for the model estimation. The detailed description of the generation of the simulated data can be found from Supplementary file Section S1.

The model parameters were then estimated using the ADVI and HMC methods in Stan. The HMC sampler was run with four chains and the number of posterior samples for each chain was 1000, out of which half were discarded as burn-in. When running ADVI we took 2000 output samples from the approximate posterior distribution to match the number of samples retrieved from HMC sampler. The number of gradient samples was 10 and number of expected lower bound (ELBO) samples was 200. After estimating the model parameters, the Bayes factors for the type 1 test were calculated. For comparison, we also ran the analysis for each cytosine separately to see the difference with the model with added spatial correlation. For this purpose the cytosine random effect was removed from the model, as there was only one cytosine being analysed at a time.

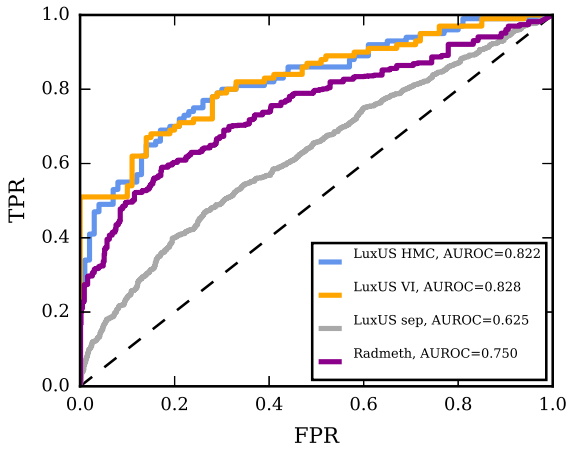

Figure 7: Receiver operating characteristics curves for LuxUS with HMC and VI model estimation, LuxUS separately for each cytosine and RadMeth. The dashed black line shows the expected ROC curve for random guessing. Results for simulated data with $\mu_b = (1.4, -1)$, 12 reads and 12 replicates were used for the figure.

To compare our method to other published tools, we ran the analysis also with RadMeth and dmrseq. For this purpose the simulated data is also saved into formats that are accepted by RadMeth and dmrseq. RadMeth was run with default parameters. dmrseq was run first with default parameters and then with higher maximum number of permutations and lower cutoff value for the single CpG coefficient, but the AUROC values for the method remained very low (possibly for some technical incompatibility reasons) and were thus left out from the result table.

The area under receiver operating characteristic curve (AUROC) values for the different approaches are presented in Table 1. The ROC curve plot for the case of 12 reads and 12 replicates is presented in Fig 7. The LuxUS model, estimated with either HMC or VI, gives the highest AUROC values. The difference between the results from running LuxUS for a genomic window and for each cytosine separately is considerable. For this simulation setting, using variational inference for estimating the model parameters leads into better results than using Hamiltonian Monte Carlo sampling, while also being significantly faster as can be seen from Table 2. The AUROC value tables and ROC curves for the simulation settings with different different μ_b can be found from Supplementary Tab S1-S3 and Supplementary Fig S7-S9 respectively.

In Fig. 8 the boxplots of the posterior means of $\boldsymbol{\theta}$ are plotted, combining the means for all cytosines and replicates for a simulation scenario where there is no differential methylation. The true methylation state without noise terms was 0.8 and the number of replicates was 6. The number of reads was varied to see whether increasing the number of reads would increase the precision of the estimates. The boxplots are plotted for the LuxUS model, estimated with both HMC and ADVI methods, and for the separate cytosine analysis, where model estimation was done with HMC sampling. The boxplots are wider for the separate cytosine analysis than for the whole window analysis, indicating higher variance

Table 1: AUROC values for LuxUS with HMC and VI, LuxUS for separate cytosines and RadMeth for simulated data set with $\boldsymbol{\mu}_b = (1.4, -1)$, corresponding to $\Delta\theta \approx 0.2$. The highest AUROC value is bolded for each simulation scenario.

Reads	Replicates	LuxUS HMC	LuxUS VI	LuxUS sep	RadMeth
6	6	0.734	**0.755**	0.522	0.642
6	12	0.814	**0.830**	0.606	0.706
6	24	0.923	**0.934**	0.687	0.822
12	6	0.687	**0.705**	0.545	0.621
12	12	0.822	**0.828**	0.625	0.750
12	24	0.927	**0.941**	0.721	0.868
24	6	0.667	**0.682**	0.541	0.622
24	12	0.840	**0.850**	0.637	0.758
24	24	0.933	**0.947**	0.762	0.892

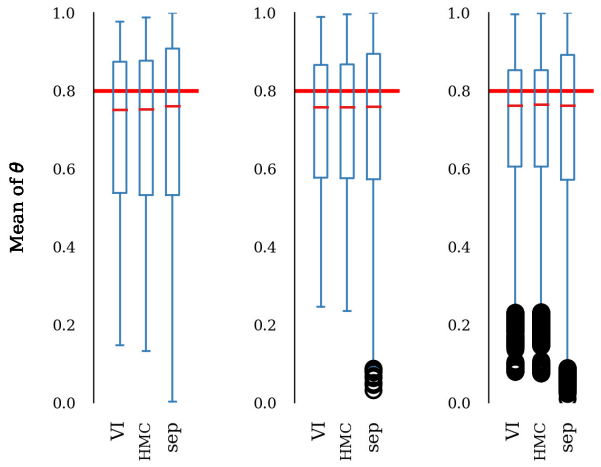

Figure 8: Boxplots of the estimated θ posterior means for the 200 simulated datasets, where number of replicates was 6 and $\boldsymbol{\mu}_b = (1.4, -1)$ was used for the simulations. The number of reads was 6, 12 and 24 respectively in the figures from left to right. The thick red lines show the true methylation level.

in the estimates. There is no clear difference between the HMC nd ADVI approaches for the whole window analysis. Increasing the number of reads from 6 to 12 and to 24 decreases the variation of the estimates, but the shift from 12 to 24 reads seems to increase the number of outliers, too.

To validate the concept of adding spatial correlation into the model, we simulated data with different numbers of cytosines in a window of 1000 basepairs and calculated the AUROC values for each case. The μ_b was set to $(1.4, -1)$ and the number of reads was 12 and number of replicates was 12. The results of this experiment can be seen in Fig. 9, which shows that analysing multiple cytosines at a time gives more statistical power. The turning point for this simulation setting seems to be between 6 and 8 cytosines, after which adding more cytosines does not seem to add more power at least in this specific case. As the number of cytosines in the analysis increases, so does the number of parameters in the model and the size of the covariance matrices. This increases computation time during model estimation and can also lead to estimation inaccuracies or convergence issues during sampling. In Fig. 4 the computation times for different number of cytosines being analysed at a time are demonstrated for the Hansen data set. The results shown in Fig. 9 indicate that increasing the number of cytosines in the analysis after some point does not give more power to the analysis, and it is not advisable in the sense of computational efficiency either.

4 Discussion

The LuxUS method was tested on both simulated and real data. The results for the simulated data show that including spatial correlation to the model gives clear advantage when compared to the analysis with only one cytosine being

Table 2: Comparison of mean computation times in seconds for estimating the LuxUS model with HMC and ADVI for a single genomic window, using simulated data set with $\boldsymbol{\mu}_b = (1.4, -1)$.

Reads	Replicates	HMC	ADVI
6	6	84.136	3.489
6	12	503.055	13.178
6	24	2918.109	55.816
12	6	83.792	3.193
12	12	448.537	11.249
12	24	2671.042	53.291
24	6	78.075	2.961
24	12	400.321	10.306
24	24	2504.799	48.830

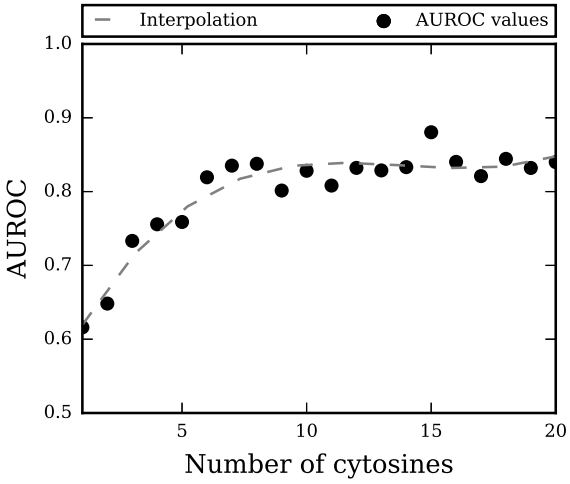

Figure 9: The AUROC value as function of number of cytosines in a 1000 bp window for simulated data and its interpolation with third degree polynomial. Each AUROC value is estimated using 400 data sets, including 200 data sets with differential methylation and 200 data sets with no differential methylation.

analysed at a time. Comparison with RadMeth method showed that our method performed notably better with each simulation setting, when the data was simulated from our model with moderate noise added. On real bisulfite sequencing data analysis LuxUS was more conservative than RadMeth, as the number of differentially methylated cytosines found by LuxUS was considerably lower than for RadMeth. This of course depends on the chosen significance levels and Bayes factor cutoff-values. There was a clear overlap between the differentially methylated cytosines retrieved from these two methods.

Based on the results on the simulated data sets, the optimal number of cytosines in a window might be $5 - 10$, as this enhances the statistical power of the statistical testing, while the model estimation is still computationally reasonable. The simulation experiment results show that the variance of the methylation state estimates is higher if we estimate the model for each cytosine separately. The variance is smaller if the whole genomic window is analysed at the same time. The means of the samples for the methylation states $\boldsymbol{\theta}$ are about as close to the true value in both cases. Based on the AUROC values calculated for the simulated data, both the HMC and VI methods for estimating the model seem to perform equally well. The HMC approach provides better model inference tools for cases where closer investigation on the fitted model is desirable, whereas VI is considerably faster.

Our method enables inference on the estimated model, which is not possible with some other tools such as RadMeth, which does not provide a summary of the fitted models. The posterior samples for the linear model coefficients and variance parameters can be explored using the summary and plotting utilities provided in Stan.

Two questions concerning our model are that does the correlation structure match reality well enough and how the priors for the variance parameters should be set. We have provided a set of default parameters for the tool, but the parameter values can be adjusted by the user to match the problem at hand and to the available computational resources. Similarly, the choice of the parameters used in the preanalysis step and the cutoff-value for the Bayes factors affect the final results. As we use the Savage-Dickey Bayes factor estimate, it may be advisable to empirically calibrate Bayes factor cutoffs for significance e.g. by using some known differentially methylated loci. It is possible to extend our model to incorporate the oxidized methylation species into the analysis as done previously in LuxGLM.

5 Conclusion

We have proposed a new tool for detecting differential methylation, which uses the spatial correlation of neighboring cytosines to enhance the accuracy of detecting differential methylation. The presented results show that our tool is able to quantify differential methylation from both simulated and real BS-seq data. Comparison with RadMeth tool on simulated data shows that our model is more accurate. The provided preanalysis step can be used to reduce the number of genomic windows for which the Bayesian analysis is done, and the computations can be parallellised for computational efficiency. Opting for variational inference in the model estimation step reduces the needed computation time even further, without having to compromise the accuracy of the fitted model. The tool is available at `https://github.com/hallav/LuxUS`.

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work has been supported by the Academy of Finland (project numbers: 292660 and 314445).

References

Carpenter, B., Gelman, A., Hoffman,M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A. (2017) Stan: A probabilistic programming language, *Journal of Statistical Software*, 76(1).

Dolzhenko, E., Smith, A. D. (2014) Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments, *BMC Bioinformatics*, 15:215.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., Beck, S. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12), 1378-85.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A. and Feinberg, A. P. (2011) Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8), 768-75.

Hansen, K. D., Langmead, B., Irizarry, R. A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13:R83.

Hansen, K. D. (2018). bsseqData: Example whole genome bisulfite data for the bsseq package. R package version 0.20.0.

Korthauer, K., Chakraborty, S., Benjamini, Y., Irizarry, R. A. (2018) Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics*, kxy007.

Kucukelbir, A., Ranganath, R., Gelman, A., Blei, D. (2015) Automatic Variational Inference in Stan. *Advances in Neural Information Processing Systems 28*.

Mayo, T. R., Schweikert, G., Sanguinetti, G. (2014) M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, 31(6), 809-16.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5), 495-501.

Rackham, O. J., Langley, S. R., Oates, T., Vradi, E., Harmston, N., Srivastava, P. K., Behmoaras, J., Dellaportas, P., Bottolo, L., Petretto, E. (2017) A Bayesian approach for analysis of whole-Genome bisulfite sequencing data identifies disease-associated changes in DNA methylation. *Genetics*, 205(4), 1443-1458.

Song, Y., Ren, H., Lei, J. (2017) Collaborations between CpG sites in DNA methylation. *International Journal of Modern Physics B*, Vol. 31, No. 20, 1750243.

Stan Development Team (2018) CmdStan: the command-line interface to Stan, Version 2.18.0. http://mc-stan.org

Stan Development Team (2017) PyStan: the Python interface to Stan, Version 2.16.0.0. http://mc-stan.org

Wen, Y., Chen, F., Zhang, Q., Zhuang, Y., Li, Z. (2016) Detection of differentially methylated regions in whole genome bisulfite sequencing data using local Getis-Ord statistics. *Bioinformatics*, 32(22), 3396-3404.

Äijö, T., Yue, X., Rao, A., Lähdesmäki, H (2016) LuxGLM: a probabilistic covariate model for quantification of DNA methylation modifications with complex experimental designs. *Bioinformatics*, 32(17), i511-i519.