LuxHS: DNA methylation analysis with spatially varying correlation structure

Viivi Halla-aho
Department of Computer Science
Aalto University
FI-00076 Aalto, Finland
viivi.halla-aho@aalto.fi

Harri Lähdesmäki
Department of Computer Science
Aalto University
FI-00076 Aalto, Finland
harri.lahdesmaki@aalto.fi

January 21, 2020

Abstract

Bisulfite sequencing (BS-seq) is a popular method for measuring DNA methylation in basepair-resolution. Many BS-seq data analysis tools utilize the assumption of spatial correlation among the neighboring cytosines' methylation states. While being a fair assumption, most existing methods leave out the possibility of deviation from the spatial correlation pattern. Our approach builds on a method which combines a generalized linear mixed model (GLMM) with a likelihood that is specific for BS-seq data and that incorporates a spatial correlation for methylation levels. We propose a novel technique using a sparsity promoting prior to enable cytosines deviating from the spatial correlation pattern. The method is tested with both simulated and real BS-seq data and compared to other differential methylation analysis tools.

***K*eywords** DNA methylation · Bayesian analysis · Spatial correlation

1 Introduction

DNA methylation is an epigenetic modification of the DNA where a methyl group is attached to a cytosine of the DNA. This phenomenon is essential for normal function of eukaryotic cells, and abnormal DNA methylation levels have been linked to diseases and cancer. DNA methylation is known to be a spatially correlated phenomena. In some cases, however, one or more cytosines in a local neighbourhood can deviate from the spatial correlation pattern due to e.g. transcription factor binding [4].

Many of the tools for differential methylation analysis assume spatial correlation without allowing cytosines to deviate from a common spatial correlation pattern. This inflexibility can lead us to not detecting all the possibly differentially methylated cytosines and could muddle the evidence for the non-deviating cytosines as well. For example RADMeth [3], which uses beta-binomial regression and weighted Z test and M^{3}D [8] where maximum mean discrepancies over the regions are used for p-value calculation do not support finding deviating cytosines. One of the tools that could take such deviation into account is BiSeq [7] which has a hierarchical procedure, where defined CpG clusters are first tested by taking the spatial correlation into account and then trimming the found differentially methylated regions (DMRs) by removing the not differentially methylated cytosines from the regions. Even though spatial correlation is assumed in the first testing phase and preprocessing of the data includes smoothing, the second step allows for controlling location-wise false discovery rate (FDR). Also, BiSeq tool divides a DMR into smaller regions if the sign of the methylation difference changes.

In [5] we proposed a novel method LuxUS, that assumes spatial correlation for cytosines in a genomic window of interest. However, the method does not support detecting deviating cytosines and it calculates one Bayes factor for the whole genomic window. Here we present a different formulation of the spatial correlation that enables the analysis of deviating cytosines by introducing weight variables d_i for each cytosine i in the genomic window. The weight variable will tell whether the corresponding cytosine follows the general spatial correlation pattern or not. Horseshoe priors [2] are often used to enhance sparsity of the coefficients in generalized linear models, where the number of covariates in the

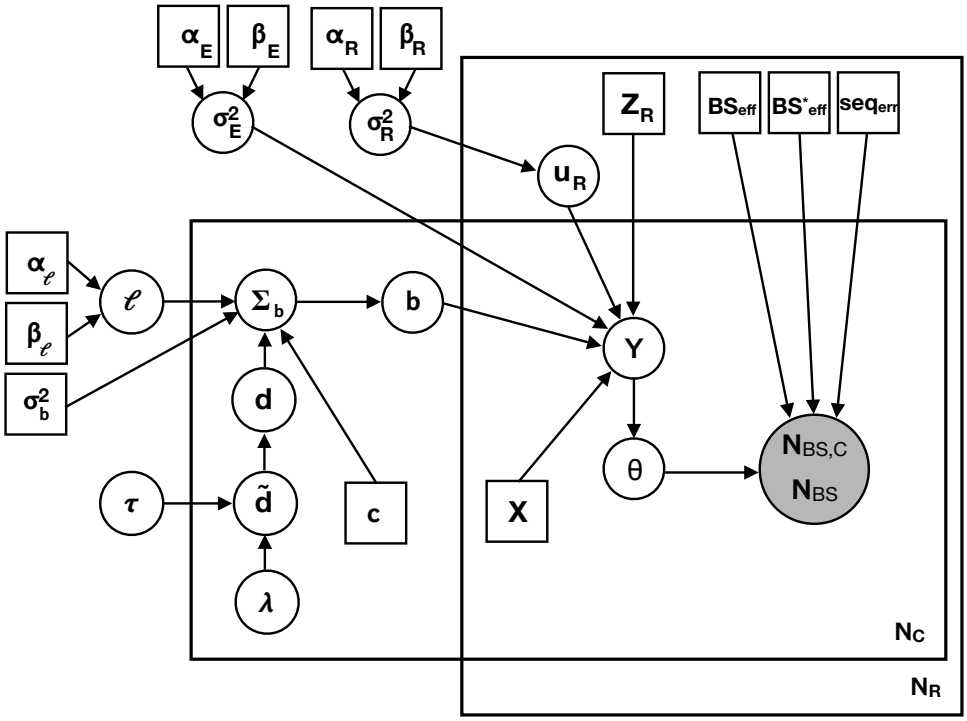

Figure 1: Plate diagram of the LuxHS model. Rectangles represent input data such as design matrices or fixed hyperparameters. White and grey circles represent latent and observed variables respectively.

model is very high and it is assumed that many of them have little effect on the predicted variable. Here we utilize horseshoe prior in the definition of the weight variables d_i. The statistical testing, i.e. calculation of Bayes factors, is done for each cytosine separately.

2 Methods

In this section the LuxHS model and analysis workflow is described. The model consists of a data generating process and a generalized linear mixed model, which models the methylation proportions by taking into account covariates, replicate effects and spatial correlation. After this the fitting of the model parameters and the method for testing for differential methylation is explained.

The first step of workflow in LuxHS analysis is to divide the data set of interest into genomic windows, which are then analysed one at a time with possibility of parallelisation. A simple preanalysis method for this purpose was proposed by Halla-aho and Lähdesmäki [5]. The cytosines are divided into genomic windows based on their genomic distance and a maximum number of cytosines in a window while filtering out cytosines with low coverage. Genomic windows with too low average coverage can also be filtered out. For windows with high enough coverage, an F-test is performed to quantify the significance for the variable of interest. The F-test p-value threshold is set to a moderate value to refrain from filtering out too many prospective genomic windows. The windows that passed the F-test phase are further processed into LuxHS input format. The model parameters are fitted, after which statistical testing is performed. Finally, the found differentially methylated cytosines can be combined into DMRs for which a follow-up analysis, such as gene-enrichment analysis, can be performed.

2.1 Model

Plate diagram of the model is presented in Fig. 1. The analysis of an experiment with N_R samples is performed for a genomic window with N_C cytosines at a time. The total sequencing read counts for each cytosine are stored in $\mathbf{N}_{\text{BS,tot}}$ out of which $\mathbf{N}_{\text{BS,C}}$ (both vectors of length $N_C \cdot N_R$) were methylated. If the experimental parameters for each sample are available, they can be stored in vectors $\mathbf{BS}_{\text{eff}}$, $\mathbf{BS}^*_{\text{eff}}$ and $\mathbf{seq}_{\text{err}}$, each of length N_R, which correspond to bisulfite conversion efficiency, incorrect bisulfite conversion efficiency and sequencing error. If the experimental parameters

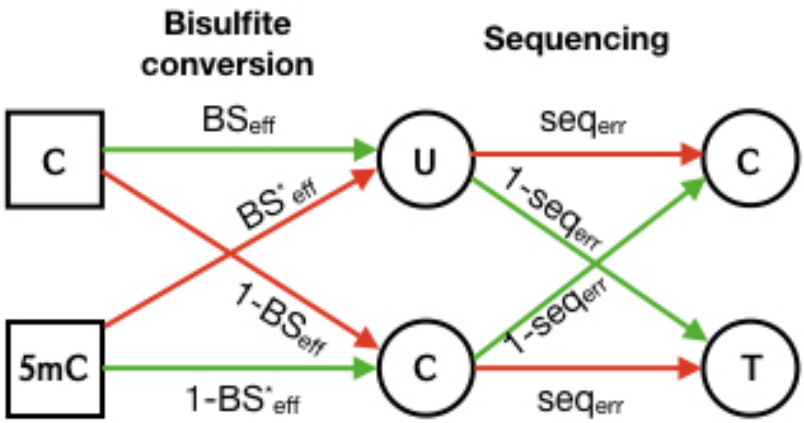

Figure 2: The probability tree for observing a C or T in bisulfite sequencing data when the true methylation state is methylated or unmethylated.

are not known, they can be set to correspond to a perfect experiment with no sequencing error and perfect bisulfite conversion efficiency.

The methylated cytosine count $N_{\text{BS,C},i}$ for observation i, $i = 1, ..., N_R \cdot N_C$, follows binomial distribution

$$N_{\text{BS,C},i} \sim \text{Binomial}(N_{\text{BS,tot},i}, p_{\text{BS,C},i}) \tag{1}$$

with success probability, e.g. probability of observing a C in bisulfite sequencing experiment, $p_{\text{BS,C},i}$, which is calculated as

$$\begin{aligned} p_{\text{BS,C},i} = \quad & \theta_i((1 - \text{seq}_{\text{err},i})(1 - \text{BS}_{\text{eff},i}) + \text{seq}_{err}\text{BS}_{\text{eff},i}) \\ & +(1 - \theta_i)((1 - \text{seq}_{\text{err},i})(1 - \text{BS}^*_{\text{eff},i}) + \text{seq}_{err}\text{BS}^*_{\text{eff},i}), \end{aligned}$$

where θ_i is the methylation proportion for the observation i, $i = 1, ..., N_C \cdot N_R$. $\text{seq}_{\text{err},i}$, $\text{BS}_{\text{eff},i}$ and $\text{BS}^*_{\text{eff},i}$ are the experimental parameters for the replicate corresponding to index i. The equation follows the probability tree in Fig. 2. This is the same data generating process as in LuxGLM [1] for non-methylated and methylated cytosines. Methylation proportions are estimated using the generalized linear mixed model of the form

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Z}_R\mathbf{u}_R + \mathbf{e}, \tag{2}$$

where term $\mathbf{Xb}$ is fixed effect, $\mathbf{Z}_R\mathbf{u}_R$ is replicate random effect and $\mathbf{e}$ is noise term with distribution $\mathbf{e} \sim N(0, \sigma^2_E\mathbf{I})$ and prior $\sigma^2_E \sim \text{Gamma}(\alpha_E, \beta_E)$. The number of covariates in the fixed effect term is N_P. Each of the cytosines has its own set of fixed effect coefficient vector $\mathbf{b}_j$ of length N_P, $j = 1, ..., N_C$, and thus $\mathbf{b} = [\mathbf{b}^T_1, \ldots, \mathbf{b}^T_{N_C}]^T$ has length $N_C \cdot N_P$. The design matrix $\mathbf{X}$ size is $(N_C \cdot N_R) \times (N_C \cdot N_P)$ and it has the individual cytosine design matrices as block matrices in the diagonal. The fixed effect coefficients have prior distribution $\mathbf{b} \sim N(0, \mathbf{\Sigma_b})$, where $\mathbf{\Sigma_b}$ is a covariance matrix. Matrix $\mathbf{Z}_R$ is the random effect design matrix of size $(N_C \cdot N_R) \times N_R$ and the vector $\mathbf{u}_R$ of length N_R contains the effects for each replicate. The effects have a normal prior distribution $\mathbf{u}_R \sim N(0, \sigma^2_R\mathbf{I})$, where $\sigma^2_R \sim \text{Gamma}(\alpha_R, \beta_R)$ is the variance term for the replicate random effect.

The spatial correlation structure is brought to the model through the fixed effect coefficients' covariance matrix $\mathbf{\Sigma_b}$. Using indexing notation $b_{j,k}$, $j = 1, ..., N_C$, $k = 1, ..., N_P$, to distinguish coefficients for each cytosine and covariate, $\mathbf{\Sigma_b}$ can be expressed as

$$\mathbf{\Sigma_b} = \begin{pmatrix} \sigma^2_b & \text{cov}(b_{1,1}, b_{1,2}) & \cdots & \text{cov}(b_{1,1}, b_{N_C,N_P}) \\ \text{cov}(b_{1,2}, b_{1,1}) & \sigma^2_b & \cdots & \text{cov}(b_{1,2}, b_{N_C,N_P}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(b_{N_C,N_P}, b_{1,1}) & \text{cov}(b_{N_C,N_P}, b_{1,2}) & \cdots & \sigma^2_b \end{pmatrix}, \tag{3}$$

where the covariance terms are

$$\text{cov}(b_{j,k}, b_{j',k'}) = \begin{cases} \sigma^2_b \cdot \exp\left(\frac{-|c_j - c_{j'}|}{\ell^2}\right) \cdot d_j \cdot d_{j'} & , \text{if } k = k' \\ 0 & , \text{if } k \neq k', \end{cases}$$

which gives the coefficients of different covariates zero covariance to fulfill linear model requirements. In the computation of the covariance terms, the cytosine locations c_j and $c_{j'}$ and the lengthscale parameter ℓ with prior $\ell \sim \text{Gamma}(\alpha_\ell, \beta_\ell)$ are used. The coefficient variance σ_b^2 is set to the value 15. Weight variables d_j and $d_{j'}$ tell whether the corresponding cytosine follows the correlation pattern along with its neighboring cytosines. The weight variables can have values ranging from 0 to 1, and they have the ability to scale down the covariance terms $\text{cov}(b_{j,k}, b_{j',k'})$.

The correlation weight variable d_j for cytosine $j = 1, ..., N_C$ is calculated through transformation

$$d_j = 1 - f(\tilde{d}_j), \tag{4}$$

where transformation function $f(x)$ is a generalized logistic function

$$f(x) = A + \frac{K - A}{(C + Q \cdot \exp(-B \cdot x))^{\frac{1}{\nu}}}, \tag{5}$$

where $A = 0$, $K = 1$, $C = 1$, $Q = 10$, $B = 5$ and $\nu = 0.5$. This transformation ensures that the resulting d_j have values from range $[0, 1]$. The auxiliary variable $\tilde{d}_j$ has a horseshoe prior with the modification of the normal priors for $\tilde{d}_j$ being restricted to the positive side, defined as

$$\tilde{d}_j \sim N^+(0, \tau^2 \cdot \lambda_j^2), \tag{6}$$

where the global shrinkage parameter τ and local shrinkage parameters λ_j have positive Cauchy hyperpriors $\tau \sim C^+(0, 1)$, and $\lambda_j \sim C^+(0, 1)$. The level of sparsity of vector $\tilde{\mathbf{d}} = [\tilde{d}_1, \ldots, \tilde{d}_{N_C}]^T$ containing $\tilde{d}_j$, $j = 1, ..., N_C$, can be controlled with the choice of hyperprior for τ.

Finally, the methylation proportions θ_i in Eq. 2 are calculated with the sigmoid function

$$\theta_i = \frac{1}{1 + \exp(-Y_i)}. \tag{7}$$

2.2 Fitting the model parameters with Stan and testing differential methylation

The model is implemented with probabilistic programming language Stan, and the Stan program is used for sampling from the posterior distribution. Stan offers both Hamiltonian Monte Carlo (HMC) and automatic differentation variational inference (ADVI) approaches for obtaining posterior samples and either one can be used for LuxHS. As variational inference approaches are often faster than Markov chain Monte Carlo (MCMC) methods such as HMC, they are a potential alternative to MCMC in computationally heavy tasks.

After obtaining samples for the model parameters, Bayes factors can be calculated for each cytosine to describe the evidence for two alternative models. The testing is done cytosine-wise, which enables deviating Bayes factor values inside a genomic window. There are two versions of the differential methylation test, with the type 1 test having a base model $M_0 : b_{j,k} = 0$ and an alternative model $M_1 : b_{j,k} \neq 0$, subscript j corresponding to the cytosines $j = 1, ..., N_C$ and subscript k corresponding to the covariate of interest. The type 2 test has a base model $M_0 : b_{j,k} - b_{j,k'} = 0$ and an alternative model $M_1 : b_{j,k} - b_{j,k'} \neq 0$, subscripts k and k' corresponding to the covariates of interest. Corresponding Bayes factors (BF) are used for the testing. As exact Bayes factors are intractable, Savage-Dickey estimates of the BFs are used instead. S-D estimate for the type 1 test is

$$BF \approx \frac{p(b_{i,j} = 0|M_1)}{p(b_{i,j} = 0|M_1, \mathcal{D})}, \tag{8}$$

where $\mathcal{D}$ is the data. The numerator is calculated using the normal prior for $\mathbf{b} \sim N(\mathbf{0}, \Sigma_{\boldsymbol{b}})$ and the denominator is estimated from the obtained samples using kernel density estimation. The type 2 test S-D estimate is formed similarly.

3 Results

In this section we present the results for real and simulated BS-seq data sets. We first analyze whole genome bisulfite sequencing (WGBS-seq) data from [6] and demonstrate that LuxHS can identify differentially methylated cytosines as well as individual cytosines whose methylation state deviate from the general spatial correlation pattern. With simulated data (for which we know the ground truth) we quantitatively evaluate LuxHS performance and compare with other state-of-the-art methods.

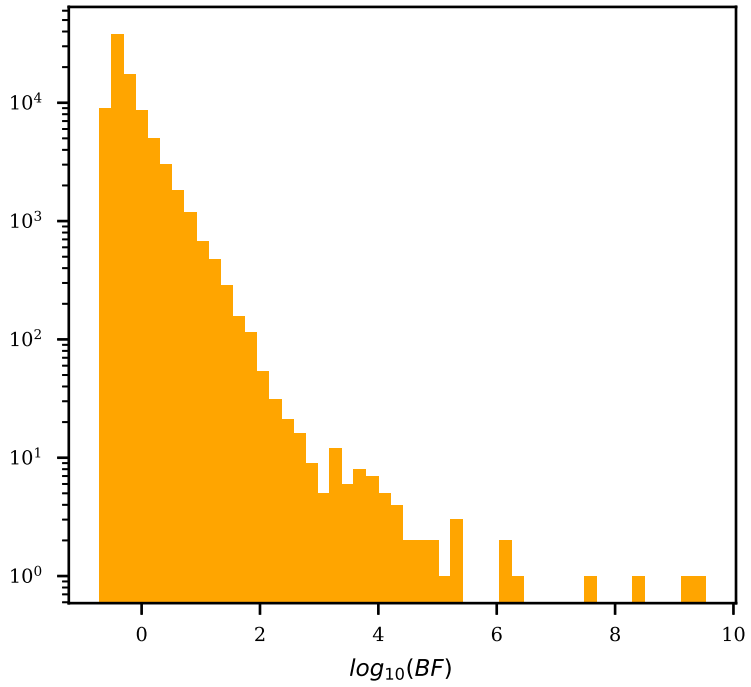

Figure 3: Histogram of the $\log(\text{BF})$ values for the colon cancer data set. The y-axis of the histogram is in log-scale.

3.1 Real bisulfite sequencing data

The colon cancer data set by Hansen [6] was used for testing LuxHS. The data set consists of six paired colon cancer and healthy colon tissue samples. The preanalysis step was run on data from chromosome 22 with the same settings as for LuxUS [5], and it resulted with 4728 genomic windows that passed the coverage and F-test criteria. Those genomic windows covered 86189 cytosines in total. For these windows, LuxHS analysis was performed. The LuxHS BF value distribution consisting of all 86189 cytosines is shown in Fig. 3. The histogram demonstrates that large majority of the Bayes factors had value smaller than 10 (or $\log(BF) \leq 1$), but there are also a few cytosines with high BF values indicating differential methylation. In total 5334 cytosines had BF> 3. To filter the results even further, a threshold for the minimum average difference between the case and control sample methylation states can be applied. In comparison, LuxUS analysis resulted in 593 windows (covering 10324 cytosines) with BF> 3, more detailed description of the results can be found from [5].

The number of cytosines for which the weight variables d were below 0.5 was 464. The Figure 4 demonstrates the differences between LuxUS and LuxHS results for a genomic window chr22:27014415-27015343. LuxUS gives one BF value (1.480) for the whole window, which suggests that there is no statistically significant differential methylation in the region. In contrary, LuxHS gives a Bayes factor for every cytosine separately while at the same time achieving two important goals: utilizing spatial correlation across the whole window of interest, and simultaneously detecting individual cytosines that deviate from this correlation pattern. Consequently, LuxHS is able to adapt to changes in the data swiftly. In the lowest panel of Fig. 4 it can be seen how LuxHS finds the cytosines for which the methylation states especially for the case samples are lower than in general in the window, and gives those cytosines lower weight parameter d values.

3.2 Simulated data

The data simulation was done using the LuxHS model, using variances $\sigma_E^2 = 1$, $\sigma_B^2 = 0.25$ and $\sigma_R^2 = 0.69$. The experimental design of the simulations included an intercept term and a case-control binary covariate. The used coefficient mean μ_B values were $[-1.4, 1]$, $[-1.4, 1.8]$,$[-1.4, 2.3]$ and $[-1.4, 2.8]$, corresponding to methylation state differences between the case and control groups $\Delta\theta$ values 0.2, 0.4, 0.5 and 0.6 respectively. The data is generated for type 1 tests. The number of total reads $N_{\text{BS,tot}}$ for the methylated counts generation and the number of replicates N_R

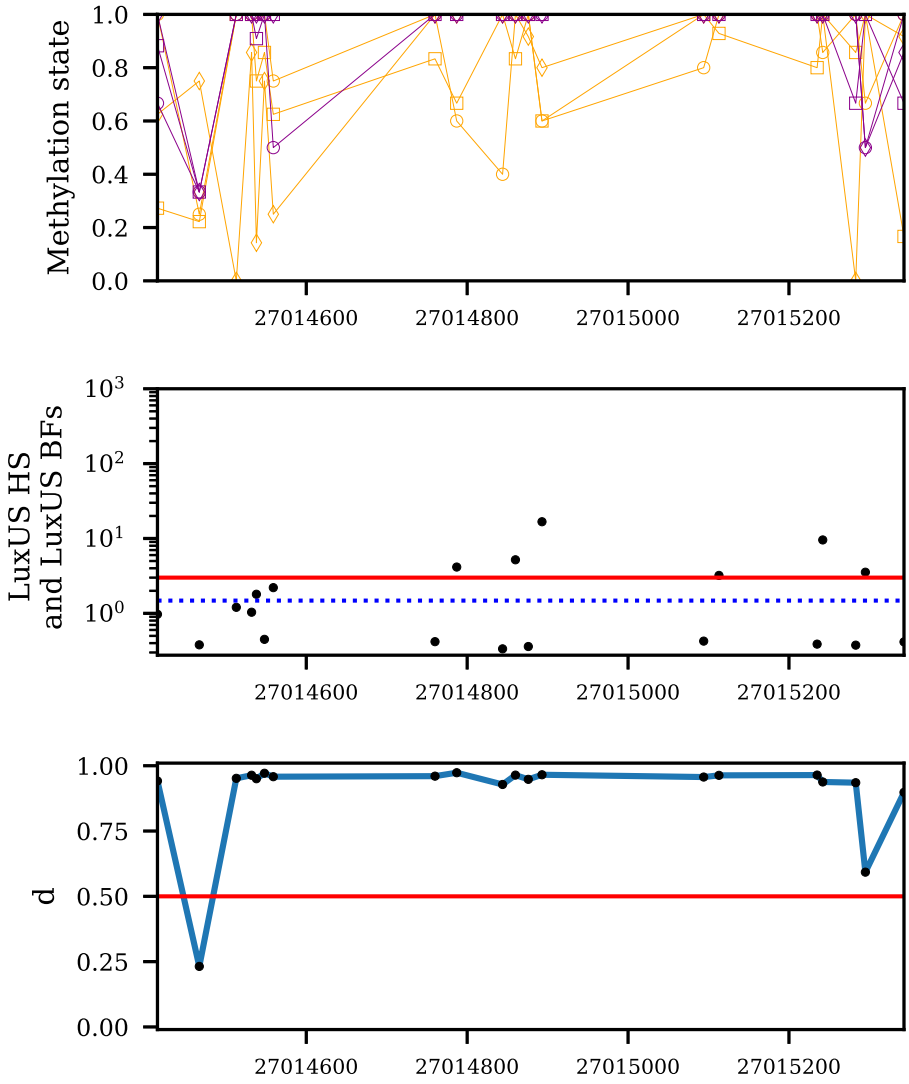

Figure 4: Results for a genomic region chr22:27014415-27015343. The top panel shows the methylation state data (as fractions $N_{\text{BS,tot}}/N_{\text{BS,C}}$) for the cytosines included in this genomic region. The cases have been plotted with purple and controls with orange, each replicate pair with a marker of its own. In the middle panel the LuxUS BF for the same region is plotted with the dashed blue line. The red line shows the threshold of BF value 3. The black dots are the LuxHS Bayes factors for each cytosine. The lowest panel shows the posterior mean of the samples for d, red line is plotted at value $d = 0.5$.

both had values 6, 12 and 24. For each combination of μ_B, $N_{\text{BS,tot}}$ and N_R we generated 100 data sets (each containing a genomic window of width 1000bp with 10 cytosines at randomly chosen locations) with and without differential methylation. The deviating cytosines had both opposite differential methylation status and deviating methylation state. We simulated data sets with 0, 1 and 2 deviating cytosines per data set. In this section we will refer to the number of deviating cytosines as N_D.

LuxHS model was compared to four other models and tools: LuxUS [5], LuxUS applied separately to every cytosine (cytosine random effect removed from the model), RADMeth [3] and BiSeq [7]. Also, LuxHS models estimated with HMC and ADVI approaches are compared. BiSeq and RADMeth were ran with default settings. We decided not to present the BiSeq results, as BiSeq did not perform very well with the simulated data. This is perhaps due to the small size of the simulated genomic regions. The comparison was done with Receiver Operating Characteristic (ROC) curve

Table 1: AUROC values for the simulated data set with one deviating cytosine ($N_D = 1$) with best value for each simulation setting in bold.

		$\mu_B = [-1.4, 1]$					$\mu_B = [-1.4, 2.3]$				
N_R	$N_{\text{BS,tot}}$	LuxHS HMC	LuxHS ADVI	LuxUS sep	LuxUS	RADMeth (NaN values)	LuxHS HMC	LuxHS ADVI	LuxUS sep	LuxUS	RADMeth (NaN values)
6	**6**	0.528	0.532	0.488	**0.587**	0.565 (29)	0.821	0.809	0.747	**0.836**	0.808 (46)
12	**6**	0.622	0.62	0.592	**0.654**	0.648 (0)	**0.894**	0.874	0.856	0.872	0.839 (31)
24	**6**	0.682	0.674	0.666	**0.732**	0.676 (20)	**0.959**	0.947	0.947	0.893	0.887 (30)
6	**12**	0.539	0.543	0.522	0.565	**0.569** (0)	0.835	0.811	0.783	**0.836**	0.796 (10)
12	**12**	0.612	0.601	0.588	**0.66**	0.614 (10)	**0.917**	0.904	0.899	0.883	0.871 (10)
24	**12**	0.708	0.704	0.699	**0.714**	0.698 (0)	**0.975**	0.967	0.967	0.896	0.899 (40)
6	**24**	0.59	0.584	0.569	**0.618**	0.59 (10)	**0.828**	0.812	0.792	0.812	0.791 (10)
12	**24**	0.664	0.656	0.641	**0.688**	0.651 (30)	**0.906**	0.894	0.894	0.852	0.839 (10)
24	**24**	0.75	0.747	0.74	**0.767**	0.727 (10)	**0.974**	0.969	0.97	0.891	0.884 (30)

Table 2: AUROC values for the simulated data set with two deviating cytosines ($N_D = 2$) with best value for each simulation setting in bold.

		$\mu_B = [-1.4, 1]$					$\mu_B = [-1.4, 2.3]$				
N_R	$N_{\text{BS,tot}}$	LuxHS HMC	LuxHS ADVI	LuxUS sep	LuxUS	RADMeth (NaN values)	LuxHS HMC	LuxHS ADVI	LuxUS sep	LuxUS	RADMeth (NaN values)
6	**6**	0.554	0.544	0.537	**0.568**	**0.568**(36)	**0.759**	0.73	0.742	0.712	0.685 (53)
12	**6**	0.618	0.607	0.61	**0.622**	0.599 (20)	**0.858**	0.823	0.845	0.75	0.742 (20)
24	**6**	**0.702**	0.687	0.697	0.679	0.678 (30)	**0.952**	0.934	0.95	0.782	0.797 (30)
6	**12**	0.563	0.538	0.556	**0.579**	0.564 (1)	**0.796**	0.769	0.779	0.722	0.721 (30)
12	**12**	0.599	0.585	0.6	**0.606**	0.586 (20)	**0.897**	0.88	0.896	0.757	0.751 (10)
24	**12**	0.686	0.68	**0.692**	0.635	0.639 (20)	0.956	0.945	**0.958**	0.783	0.802 (10)
6	**24**	0.557	0.555	0.553	**0.565**	0.538 (30)	**0.835**	0.808	0.825	0.738	0.729 (10)
12	**24**	0.648	0.641	**0.651**	0.622	0.588 (30)	0.905	0.889	**0.906**	0.75	0.774 (10)
24	**24**	0.696	0.695	**0.702**	0.658	0.639 (30)	**0.965**	0.957	**0.965**	0.785	0.787 (10)

statistics for all method. RADMeth runs resulted in a few NaN p-values, which were removed from the AUROC and TPR calculation. Area Under ROC curve (AUROC) value tables for $N_D = 1$ and $N_D = 2$ in Table 1-2 show that when the magnitude of differential methylation $\Delta\theta$ is smaller, LuxUS performs the best. When $\Delta\theta$ is higher, LuxHS and LuxUS for each cytosine separately have the highest AUROC values.

Based on the AUROC values, HMC version of LuxHS performs consistently slightly better than ADVI. The strength of ADVI is its computational efficiency. The mean runtime (over the 200 generated genomic windows) of the HMC version ranged from 40 to 971 seconds (for $\mu_B = [-1.4, 1]$, $N_{\text{BS,tot}} = [6, 12, 24]$ and $N_R = [6, 12, 24]$), while for ADVI the range is $5 - 47$ seconds. The computations were run on a computation cluster.

The accuracy of estimating whether a cytosine is deviating or non-deviating from the common spatial correlation pattern was assessed using the estimated weight variable values d_j. The posterior means of all weight variables were computed, and AUROC and TPRs were determined. The results in Table 3 show, that overall LuxHS can determine the deviance status accurately, but it seems not to be able to find all of the deviating cytosines. This indicates, that LuxHS rather gives too high d values than too low.

To investigate how LuxHS behaves when there are no deviating cytosines, such data sets were simulated and LuxHS analysis was performed along with the other methods it was compared to earlier. The data was simulated with $\Delta\theta = 0.5$. Out of the compared methods, LuxUS had the best AUROC values (see Table 4). LuxHS showed relatively good performance, demonstrating that the added flexibility of modeling cytosines that can deviate from the general spatial correlation pattern does not significantly decrease the performance of differential methylation analysis in the case of all cytosines following the same correlation pattern. Recall that for the cases where one or more of the cytosines deviate from the spatial correlation pattern LuxHS can reach state-of-the-art performance (see Tables 1-2). Moreover, LuxHS does not impose small values of weight variables d_j where it is not appropriate. There were no d_j values smaller than 0.5 for any of the generated genomic windows in any of the simulation settings.

Table 3: AUROC and true positive rates (TPR) for detecting the deviating cytosines in simulated data sets for LuxHS (HMC). AUROC was calculated using the posterior means for each d_j. For the TPR calculation the j^{th} cytosine is considered deviating if the posterior mean of d_j is smaller than a threshold value. The results for two weight value thresholds 0.5 and 0.75 are shown in separate columns.

	AUROC			TPR (0.5)			TPR (0.75)		
N_R	$N_{\text{BS,tot}}$								
	6	12	24	6	12	24	6	12	24
$N_D = 1, \mu_B = [-1.4, 1]$									
6	0.788	0.832	0.858	0.095	0.12	0.065	0.37	0.295	0.23
12	0.877	0.882	0.905	0.09	0.105	0.07	0.255	0.29	0.295
24	0.935	0.935	0.938	0.1	0.035	0.055	0.33	0.31	0.25
$N_D = 1, \mu_B = [-1.4, 2.3]$									
6	0.774	0.853	0.888	0.14	0.09	0.125	0.315	0.36	0.34
12	0.885	0.930	0.939	0.145	0.13	0.18	0.38	0.345	0.46
24	0.965	0.967	0.983	0.175	0.17	0.19	0.455	0.45	0.535
$N_D = 2, \mu_B = [-1.4, 1]$									
6	0.751	0.775	0.763	0.138	0.093	0.073	0.435	0.35	0.218
12	0.791	0.842	0.829	0.095	0.09	0.07	0.305	0.3075	0.268
24	0.869	0.846	0.864	0.075	0.078	0.065	0.275	0.29	0.278
$N_D = 2, \mu_B = [-1.4, 2.3]$									
6	0.741	0.786	0.804	0.103	0.11	0.118	0.36	0.318	0.358
12	0.833	0.863	0.869	0.153	0.145	0.11	0.423	0.408	0.313
24	0.893	0.899	0.937	0.15	0.16	0.183	0.458	0.458	0.5

Table 4: AUROC values for the simulated data set with zero deviating cytosines. $\mu_B = [-1.4, 2.3]$ was used for the simulations. The best AUROC for each simulation setting is shown bolded.

N_R	$N_{\text{BS,tot}}$	LuxHS HMC	LuxHS ADVI	LuxUS sep	LuxUS	RADMeth (NaN values)
6	**6**	0.865	0.875	0.755	**0.943**	0.896 (3)
12	**6**	0.925	0.93	0.863	**0.972**	0.945 (0)
24	**6**	0.965	0.965	0.935	**0.994**	0.977 (30)
6	**12**	0.874	0.879	0.794	**0.941**	0.885 (21)
12	**12**	0.964	0.962	0.915	**0.992**	0.972 (40)
24	**12**	0.967	0.961	0.948	**0.99**	0.973 (10)
6	**24**	0.848	0.846	0.787	**0.906**	0.869 (10)
12	**24**	0.925	0.92	0.889	**0.967**	0.936 (40)
24	**24**	0.982	0.978	0.971	**0.998**	0.986 (20)

4 Discussion

The analysis of real and simulated BS-seq data shows that LuxHS model can detect loci where the methylation state deviates from the surrounding cytosines. The tests with the simulated data show that the way LuxHS calculates Bayes factors separately for each cytosine can improve the accuracy when compared to LuxUS or other state-of-the-art methods, especially if the proportion of deviating cytosines is high.

The proportion of deviating cytosines that can be found in a genomic window could be further tweaked through the choice of hyperprior for global horseshoe prior τ. For example, the recommendations in [9] could be used if the default prior does not match the user's beliefs about the number of deviating cytosines.

The covariance structure with possibility of breaking the correlation pattern might also be advantageous in other bioinformatic modeling purposes, where a spatial correlation pattern with possibility of deviation is needed. The spatial correlation structure proposed in here can be easily applied in a general or generalized linear model setting. Another application could be time series analysis, where consecutive time points are often correlated, but some of the time points may deviate from the expected correlation pattern e.g. due to an outlier value.

5 Conclusion

In this work we propose a novel method for differential methylation analysis, LuxHS. The tool supports detecting cytosines, which do not follow the same methylation pattern as its neighboring cytosines. This could happen because of

e.g. transcription factor binding. The results with simulated and real BS-seq data show, that LuxHS is able to detect such cytosines and that this feature increases the accuracy of differential methylation analysis, especially when the number of deviating cytosines or the amount of differential methylation is higher. The tool and usage instructions are available in GitHub repository in `https://github.com/hallav/LuxUS-HS`.

Funding

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

Acknowledgements

The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project.

References

[1] Äijö, T., Yue, X., Rao, A., Lähdesmäki, H.: Luxglm: a probabilistic covariate model for quantification of dna methylation modifications with complex experimental designs. Bioinformatics **32**(17), i511–i519 (2016)

[2] Carvalho, C.M., Polson, N.G., Scott, J.G.: Handling sparsity via the horseshoe. In: Artificial Intelligence and Statistics. pp. 73–80 (2009)

[3] Dolzhenko, E., Smith, A.D.: Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC bioinformatics **15**(1), 215 (2014)

[4] Domcke, S., Bardet, A.F., Ginno, P.A., Hartl, D., Burger, L., Schübeler, D.: Competition between dna methylation and transcription factors determines binding of nrf1. Nature **528**(7583), 575 (2015)

[5] Halla-aho, V., Lähdesmäki, H.: Luxus: Detecting differential dna methylation using generalized linear mixed model with spatial correlation structure. bioRxiv p. 536722 (2019)

[6] Hansen, K.D.: bsseqdata: example whole genome bisulfite data for the bsseq package (2016)

[7] Hebestreit, K., Dugas, M., Klein, H.U.: Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics **29**(13), 1647–1653 (2013)

[8] Mayo, T.R., Schweikert, G., Sanguinetti, G.: M3D: a kernel-based test for spatially correlated changes in methylation profiles. Bioinformatics **31**(6), 809–816 (11 2014)

[9] Piironen, J., Vehtari, A.: On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. arXiv preprint arXiv:1610.05559 (2016)