

SOFTWARE

Open Access



LuxHMM: DNA methylation analysis with genome segmentation via hidden Markov model

Maia H. Malonzo* and Harri Lähdesmäki

*Correspondence:
maia.malonzo@gmail.com

Department of Computer
Science, Aalto University,
00076 Espoo, Finland

Abstract

Background: DNA methylation plays an important role in studying the epigenetics of various biological processes including many diseases. Although differential methylation of individual cytosines can be informative, given that methylation of neighboring CpGs are typically correlated, analysis of differentially methylated regions is often of more interest.

Results: We have developed a probabilistic method and software, LuxHMM, that uses hidden Markov model (HMM) to segment the genome into regions and a Bayesian regression model, which allows handling of multiple covariates, to infer differential methylation of regions. Moreover, our model includes experimental parameters that describe the underlying biochemistry in bisulfite sequencing and model inference is done using either variational inference for efficient genome-scale analysis or Hamiltonian Monte Carlo (HMC).

Conclusions: Analyses of real and simulated bisulfite sequencing data demonstrate the competitive performance of LuxHMM compared with other published differential methylation analysis methods.

Keywords: Methylation, Bisulfite sequencing, Probabilistic, HMM

Background

DNA methylation is an important epigenetic modification associated with many biological processes including various diseases. In promoters, DNA methylation tends to repress gene expression whereas in intragenic locations they tend to upregulate expression [1]. Bisulfite sequencing, whether whole genome (WGBS) or reduced representation (RRBS) bisulfite sequencing, allows for interrogation of DNA methylation at the level of individual CpGs. Moreover, decreasing costs of sequencing have increased the use of these methods. DNA methylation are often studied by analyzing differentially methylated loci (DML) or regions (DMR). Although single differentially methylated CpGs are informative, often DMRs are of more interest [2]. Further, analyzing the combined methylation differences of CpGs within regions increase the statistical power of differential methylation detection.



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Given such interest in DMRs, several methods have been developed for identifying them (Table 1). RADMeth uses the beta-binomial regression method in handling complex experimental designs [3]. Beta-binomial regression is used to individually fit single cytosines and then measures the significance of differential methylation using the log-likelihood ratio test between the full and reduced models which generates p -values. To combine information from neighboring cytosines into regions it transforms p -values using the weighted Z-test which then determines which cytosines are combined into regions using an FDR threshold. A method called metilene first recursively segments the genome into regions using the circular binary segmentation algorithm which generates regions that maximizes the difference of CpG-wise mean methylation levels [4]. Then, it calculates p -values using a version of the Kolmogorov–Smirnov test which tests the significance of potential DMRs. HMM-DM uses hidden Markov model (HMM) to segment the genome into regions and Bayesian methods to infer model parameters. It then uses MCMC to compute the posterior probability of each state: hypermethylated, equally methylated or hypomethylated. To identify DMRs, it joins hypermethylated or hypomethylated CpGs into regions. In DMRcate, standard linear modelling is performed using limma which generates a signed statistic for measuring the difference between treatment effects per CpG site [5]. The square of this value is then applied to a Gaussian smoother. It then uses an approximation that generates a value for which a p -value is computed by comparison to a chi-square distribution. Individual sites below a given p -value threshold are selected and grouped into regions that are separated by, at most, a threshold number of nucleotides. DSS models the methylation counts by a beta-binomial distribution with an arcsine link function and fits the transformed methylation levels with a generalized least squares procedure from which it obtains estimates of the model coefficients at each CpG site [6]. Hypothesis testing is performed using Wald test on the coefficient estimates. After detection of statistically significant CpG sites, DSS merges nearby loci into regions.

LuxGLM [7] and LuxUS [8] use extended versions of generalized linear model (GLM) to analyze methylation data with complex experimental designs and incorporate estimation of experimental parameters that describe the underlying biochemistry in methylation sequencing data. LuxGLM uses matrix normal distribution to handle multiple methylation modifications. LuxUS uses a generalized linear mixed model (GLMM) to analyze cytosines within a genomic window simultaneously. To analyze the spatial correlation of cytosines it uses a random effect correlation structure. It also analyzes the

Table 1 Methods comparison

| Method | Methylation model | Algorithm for CpG correlation |
|----------|-------------------------|-------------------------------------|
| RADMeth | Beta-binomial | Weighted Z-test |
| Metilene | Kolmogorov–Smirnov | Circular binary segmentation |
| HMM-DM | Bayesian HMM | HMM |
| DMRcate | Linear Gaussian (limma) | Kernel smoothing |
| DSS | Beta-binomial | Smoothing via moving average |
| LuxUS | Bayesian GLMM | Random effect correlation structure |
| LuxHMM | Bayesian GLM | HMM |

variation of individual replicates using a replicate random effect. Features of previous methods as well as the proposed method, LuxHMM, are contrasted in Table 1.

Implementation

Bisulfite sequencing data consists of DNA where unmethylated cytosines are converted into uracil by bisulfite treatment and sequenced as thymine to differentiate it from methylated cytosine which are not converted and sequenced as cytosine.

A commonly used methylation level estimate is obtained by taking the ratio of methylated cytosine to the sum of methylated and unmethylated cytosine, $\mu = N_{BS,C}/N_{BS}$. To infer differential methylation, the methylation levels between groups are compared. Hypermethylation occurs when the methylation level for a comparison (or treatment) group is generally higher compared to a reference (or control) group, and hypomethylation when it is lower. We are interested in modeling methylation levels and differential methylation across T cytosines c_1, c_2, \dots, c_T . Differentially methylated regions are often of more interest than single cytosines due to their combined effect compared to the individual effect of a single cytosine. A methylated region C consists of consecutive CpGs c_t s that are hypermethylated, hypomethylated or have equal methylation (M_j), $C = \{c_t \mid c_t \in M_j\}$. A region is differentially methylated when it is either hypermethylated or hypomethylated.

Our method consists of two modules: (1) genome segmentation via HMM, and (2) estimation of methylation levels and inference of differential methylation using Bayesian GLM. In inference of differential methylation, significance of explanatory variable is measured by Bayes factors.

Genome segmentation via HMM

To extract regions from a sequence of cytosines, we use hidden Markov model (HMM). HMM is a statistical model that infers a sequence of hidden states from a sequence of observations. In this work, the hidden states x are the methylation states, specifically: (1) hypermethylation, (2) hypomethylation, and (3) equal methylation between two groups. For each cytosine, the observations y are the differences in the mean methylation levels between groups, $y = \bar{\mu}_1 - \bar{\mu}_2$, where $\bar{\mu}_1$ is the mean methylation level for one group and $\bar{\mu}_2$ for another.

HMM is parameterized by two distributions: the observation emission probabilities and the state transition probabilities (Fig. 1). The observation emission probabilities, $P(y_t|x_t)$, give the probability of observing y at cytosine position t given the underlying hidden state x_t , i.e. the probability of observing the differences in methylation levels between two groups given the underlying methylation states M_j (hypermethylation, hypomethylation or equal methylation). The state transition probabilities, $P(x_t|x_{t-1})$, give the probability of hidden state x_{t-1} moving to x_t in a sequence, i.e. the probability of moving from one methylation state to another (or remaining the same) between two consecutive CpGs.

For a given hidden state sequence $X = x_1, x_2, \dots, x_T$ and observation sequence $Y = y_1, y_2, \dots, y_T$, the observation sequence likelihood is

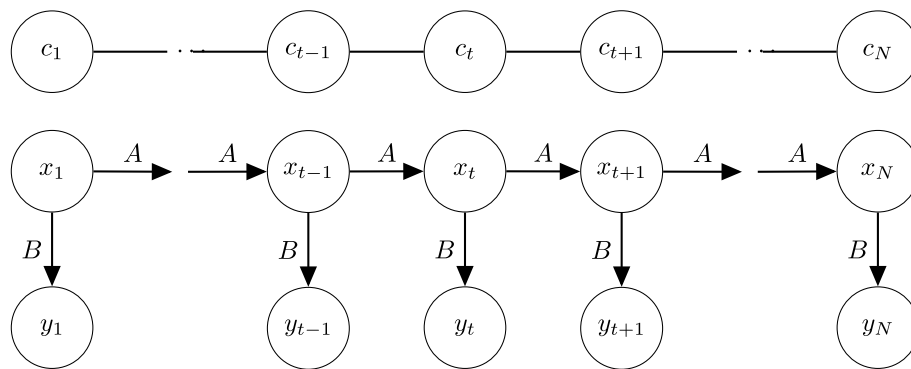


Fig. 1 Diagram of emission and transition probabilities. The top-most row (C) indicates the cytosine position, the second row (X) denotes the hidden methylation states and the bottom row (Y) represents the observed differences in methylation levels between groups. A denotes the state transition probabilities and B the observation emission probabilities

$$P(Y|X) = \prod_{t=1}^T P(y_t|x_t).$$

It is straightforward to compute the joint probability of a given sequence of methylation states and a sequence of observed methylation differences

$$P(Y, X) = P(Y|X)P(X) = \prod_{t=1}^T P(y_t|x_t) \prod_{t=1}^T P(x_t|x_{t-1}).$$

The total probability of the observed methylation differences can then be obtained by summing over the hidden states

$$P(Y) = \sum_X P(Y, X) = \sum_X P(Y|X)P(X).$$

With these definitions we can select the hidden state sequence that maximizes the observation likelihood. However, this is infeasible due to the high number of possible state sequences. Instead a dynamic programming algorithm like the Viterbi algorithm recursively computes $v_t(j)$ which denotes the probability of being in state j given the observations for the first t cytosines. For a given state x_j at cytosine position t , $v_t(j)$ is computed by

$$v_t(j) = \max_i v_{t-1}(i) a_{ij} b_j(x_t)$$

where $v_{t-1}(i)$ is the previous Viterbi path probability from the previous time step, a_{ij} is the transition probability from previous state x_i to current state x_j and $b_j(y_t)$ is the emission probability of the observation y_t given state j [9].

To learn the most likely transition, $\mathbf{A} = \{a_{ij}\}$, and emission, $\mathbf{B} = \{b_j(y_t)\}$, probabilities and initial state distribution $\pi_i = P(X_1 = i)$, the Baum-Welch algorithm, another dynamic programming algorithm, finds a (local) maximum of $\eta^* = \arg \max_{\eta} P(Y | \eta)$, where $\eta = (A, B, \pi)$, using the expectation-maximization (EM) algorithm [10].

In this work we use pomegranate, a Python package that implements probabilistic models, including HMMs [11]. The model is initialized with state and transition probabilities. We assume the emission distributions follow a Gaussian distribution $\mathcal{N}(\psi, \sigma)$, where ψ and σ are set to 0 and 0.08 (equal methylation), 0.3 and 0.06 (hypermethylation) and -0.3 and 0.06 (hypomethylation). The transition probabilities were optimized using the Baum-Welch algorithm using the initial values shown in Additional file 1: Section 1

To determine the most likely sequence of hidden states, i.e. the sequence of methylation states, we use the Viterbi algorithm implemented in the package. To learn the most likely emission and transition probabilities given the sequence of observations we use the Baum-Welch algorithm, also supported by pomegranate. After learning the hidden methylation states, adjacent cytosines with the same methylation state are combined into regions, as well as the total read counts $N_{ir}^{BS} = \sum_{k=1}^{W_{ir}} N_{irk}^{BS}$, where k is the k th CpG in C_{ir} and $W_{ir} = |C_{ir}|$ is the number of consecutive CpGs with the same methylation state in the i th sample and the r th region and, similarly for methylated read counts, $N_{ir}^{BS,C} = \sum_{k=1}^{W_{ir}} N_{irk}^{BS,C}$.

Estimation of methylation levels and differential methylation

We briefly review the underlying statistical model for the experimental parameters [7]. Experimental parameters that define the underlying biochemistry in bisulfite sequencing should be considered in estimation of methylation levels. Bisulfite conversion rate (BS_{eff}), sequencing error (seq_{err}) and incorrect bisulfite conversion rate (BS_{eff}^*) can significantly affect methylation estimates. Low BS_{eff} causes overestimation of methylation levels whereas high BS_{eff}^* results in underestimation. On the other hand, high seq_{err} can lead to either overestimation or underestimation.

BS_{eff} can be estimated by using the lambda phage genome. Since the lambda phage genome is unmethylated, BS_{eff} can be estimated by taking the ratio of all cytosine reads converted into thymine over the total number of reads. Similarly, BS_{eff}^* can be estimated with spike-ins of oligonucleotides where all the cytosines are methylated. On the other hand, seq_{err} can be estimated using Phred scores Q by converting them to base-calling error probabilities $P = 10^{-\frac{Q}{10}}$.

Given the above definitions, BS_{eff} , BS_{eff}^* and seq_{err} determine the conditional probability of a sequencing readout being “C”, given that the cytosine is methylated or unmethylated (Fig. 2). Specifically, since BS_{eff} is the probability of an unmethylated cytosine being converted into uracil, $1 - BS_{\text{eff}}$ is the probability of an unmethylated

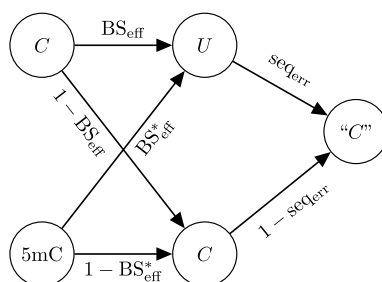


Fig. 2 Probability tree of observing “C” readout when the true methylation state is methylated or unmethylated

cytosine incorrectly not converted into uracil. If an unmethylated cytosine is correctly converted into uracil it still has seq_{err} probability of being incorrectly sequenced as “C”. Whereas, if it is incorrectly not converted to uracil and remains a cytosine, it has $1 - \text{seq}_{\text{err}}$ probability of being correctly sequenced as “C”. Put together, the conditional probability of sequencing “C” given the cytosine is unmethylated is

$$p_{\text{BS}}(\text{“C”}|\text{C}) = (1 - \text{BS}_{\text{eff}})(1 - \text{seq}_{\text{err}}) + \text{BS}_{\text{eff}}\text{seq}_{\text{err}}. \quad (1)$$

On the other hand, if a cytosine is methylated, the probability that it is correctly not converted to uracil is $1 - \text{BS}_{\text{eff}}^*$ and the probability that it is correctly sequenced as “C” is $1 - \text{seq}_{\text{err}}$. The probability that the unmethylated cytosine is incorrectly converted to uracil and incorrectly sequenced as “C” are, respectively, BS_{eff}^* and seq_{err} . Thus, the conditional probability of sequencing “C” given the cytosine is methylated is

$$p_{\text{BS}}(\text{“C”}|5\text{mC}) = (1 - \text{BS}_{\text{eff}}^*)(1 - \text{seq}_{\text{err}}) + \text{BS}_{\text{eff}}^*\text{seq}_{\text{err}}. \quad (2)$$

Thus far we have described individual cytosines. However, this description can be generalized to DNA regions. Let $\theta \in [0, 1]$ represent the unknown fraction (or probability) of methylated DNA. Following Eqs. 1 and 2, the probability of observing “C” readouts for a given region is $p_{\text{BS}}(\text{“C”}) = p_{\text{BS}}(\text{“C”}|5\text{mC})\theta + p_{\text{BS}}(\text{“C”}|\text{C})(1 - \theta)$. Finally, the total number of “C” readouts is binomially distributed,

$$N_{\text{BS,C}} \sim \text{Bin}(N_{\text{BS}}, p_{\text{BS}}(\text{“C”})), \quad (3)$$

where N_{BS} is the total number of reads. See Fig. 3 for the plate diagram of the model.

To incorporate complex experimental designs to the model, we simplify the method proposed in [8] by doing away with the spatial correlation component and use generalized linear regression,

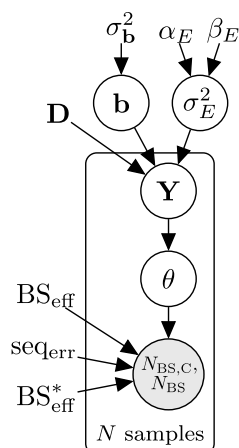


Fig. 3 Plate diagram of the LuxHMM model for analyzing experimental parameters and methylation levels. The circles represent latent (white) and observed (gray) variables and the unbordered nodes represent hyperparameters and constant values

$$\begin{aligned}\mathbf{b} &\sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{b}}^2 \mathbf{I}) \\ \sigma_E^2 &\sim \text{Gamma}(\alpha_E, \beta_E) \\ \mathbf{Y} &\sim \mathcal{N}(\mathbf{D}\mathbf{b}, \sigma_E^2),\end{aligned}$$

where $\mathbf{b} \in \mathbb{R}^{N_p}$ (where N_p is the number of covariates, possibly including the intercept) is the vector of regression coefficients, $\mathbf{D} \in \mathbb{R}^{N \times N_p}$ is the design matrix, and $\mathbf{Y} \in \mathbb{R}^N$. The values of the hyperparameters are $\sigma_B^2 = 15$, $\alpha_E = 5$, and $\beta_E = 5$, and were taken from [8]. We apply this model to regions instead of single CpGs to speed up computation. Finally, we use the sigmoid link function

$$\theta = \sigma(\mathbf{Y}).$$

The model is implemented using the probabilistic programming language Stan [12], and model inference is done using either Hamiltonian Monte Carlo (HMC) or automatic differentiation variational inference (ADVI) for faster estimation of the model parameters [13], both built-in features of Stan. Stan uses a locally adaptive version of dynamic HMC sampling. In variational inference (VI) the posterior $p(\phi|D)$ of all unknowns ϕ given observed data D is approximated with a simpler distribution $q(\phi; \rho)$, which is selected from a chosen family of distributions by minimizing divergence between $p(\phi|D)$ and $q(\phi; \rho)$.

To detect differential methylation w.r.t. any of the N_p covariates in \mathbf{D} , hypothesis testing was done using Bayes factors via the Savage-Dickey density ratio method as implemented in [7].

Results

To demonstrate the performance of LuxHMM and how well it performs compared to other methods, we analyze real and simulated BS-seq datasets. The first dataset is a simulated dataset based on real BS-seq data. The second is a simulated BS-seq dataset generated using a general experimental design. Lastly, we use a real BS-seq dataset with confounding covariates. We compare the performance of LuxHMM with RADMeth, metilene, HMM-DM, DMRcate and DSS.

Comparison of performance on simulated dataset based on real BS-seq data

To assess the accuracy of our method compared to other published methods we used a simulated dataset by [14]. Bisulfite sequencing data was obtained from real CpG islands which allowed variance and correlation to be incorporated into the simulated dataset. The dataset was derived from 12 individuals which were divided into 6 controls and 6 cases. The dataset was divided into two sets wherein 10,000 DMRs were incorporated into one set. Methylation counts were added to or subtracted from the case samples so that the methylation differences were 0.1, 0.2, 0.3 or 0.4.

In LuxHMM, either all regions or only candidate hypo- and hypermethylated regions, as classified by HMM, were used as input in determining DMRs. Parameter settings for competing methods are described in Additional file 1: Section 2.

The area under the receiver operating curve (AUROC) and the average precision (AP), to handle the imbalance in the dataset given that there are much more negative than positive samples, were computed (Table 2). For AP, the baseline is 0.11 which is the

fraction of the number of true positives over the total number tests. True positives are differentially methylated cytosines whereas negatives are non-differentially methylated cytosines. In all methods, cytosines which are not covered by the returned regions are given a score of zero. The highest AUROC and AP were generated by LuxHMM used with all regions. The higher recall suggests that the state assignment of HMM misses differentially methylated regions which are inaccurately classified as regions with equal methylation between two groups. This also demonstrates that LuxHMM more accurately detects DMRs compared to the other methods used. Another notable result is that DMRcate has a relatively high AUROC and a low AP. This could be caused by a high false positive rate which is masked in AUROC due to a high number of true negative samples. As true negative samples are excluded in the computation of AP, the high false positive rate results in a low AP.

Alternative emission probabilities

To test the sensitivity of the proposed model to different emission distribution parameters, we tested various parameter values on the [14] dataset using all regions. Table 3 shows that the model is not sensitive to different values of standard deviation but is sensitive to the means, with the highest AP when using means -0.3 and 0.3 .

We also tested using five hidden states with two hidden states each for the hypo- and hypermethylated regions (Table 4). The AUROC and AP are, respectively, 0.946 and 0.844, indicating that increasing the number of hidden states from three to five does not increase accuracy.

When not specifying the emission distributions and letting pomegranate instead estimate the emission distributions we obtain a higher AUROC and a lower AP (Table 5). We prioritize AP as it takes into account the imbalanced dataset. Genome segmentation was based on Fig. 4. For three hidden states, we used s_1 as candidate hypo- and hypermethylated states (with s_0 and s_2 as states with no difference between groups), whereas with four hidden states we used s_1 and s_2 as candidate hypo- and hypermethylated states (with s_0 and s_3 as states with no difference between groups). In computing AUROC and AP we used either all hidden states (including state with no difference between groups) or just candidate hypo- and hypermethylated states.

Comparing beta-values and M-values

We used the beta-value representation for methylation levels as they allow a more intuitive interpretation. However, the emission distributions used for the beta-values are normal distributions which are better suited with the support of M-values which is the set of real numbers. As such, we tested the method using M-values instead of beta-values for analyzing the dataset from [14] using as input candidate hypo- and hypermethylated regions. For the mean values of the emission distribution we used values that are roughly equivalent to a methylation difference of -0.3 and 0.3 to be comparable with the analysis using beta-values (Table 6). The highest AUROC and AP generated were obtained using means -1.2 (hypomethylated) and 1.2 (hypermethylated) and -1.7 and 1.7 , respectively. The AUROC was higher using M-values (0.942 vs. 0.935) but the AP was higher using beta-values (0.820 vs. 0.830). We prioritize

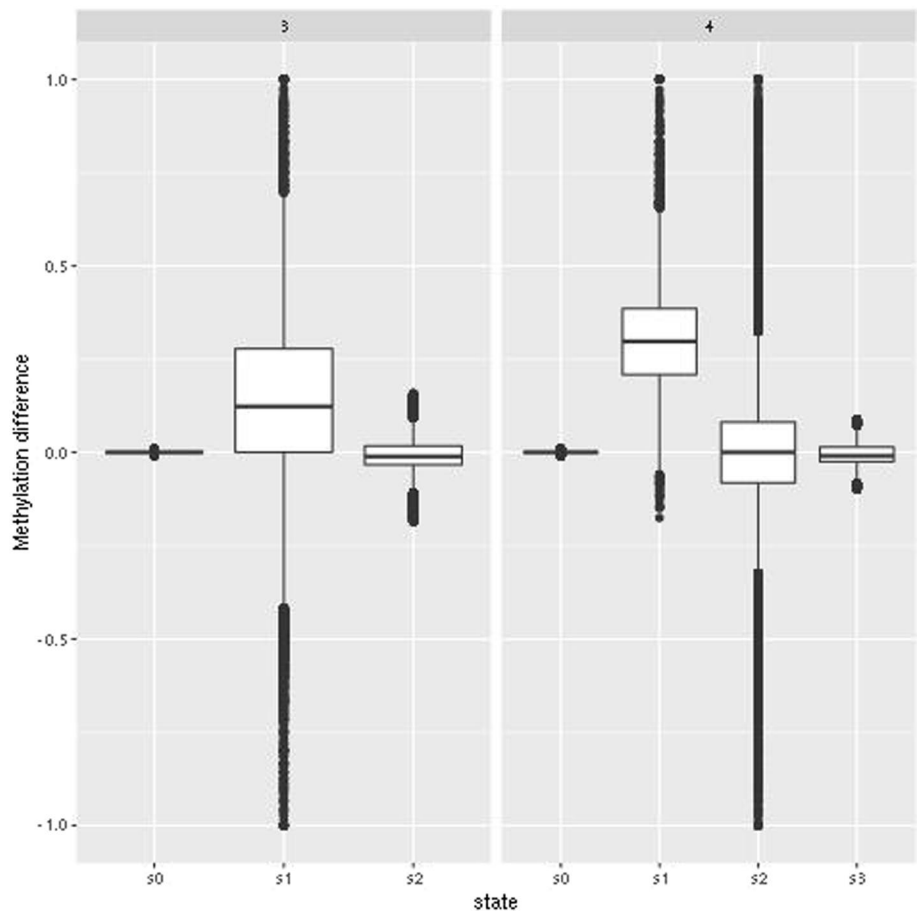


Fig. 4 Distribution of methylation differences with three and four hidden states when the emission distributions are not specified

Table 2 AUROC and AP for simulated dataset from [14]

| Method | AUROC | AP |
|---------------------|--------------|--------------|
| LuxHMM ¹ | 0.945 | 0.852 |
| LuxHMM ² | 0.935 | 0.830 |
| LuxUS | 0.900 | 0.601 |
| RADMeth | 0.831 | 0.644 |
| Metilene | 0.834 | 0.674 |
| HMM-DM | 0.626 | 0.315 |
| DMRcate | 0.621 | 0.182 |
| DSS | 0.857 | 0.712 |

Bold represent the highest values in each column

1 All regions

2 Hypo- and hypermethylated regions

Table 3 AUROC and AP for different emission distributions

| Equal mean | Std. dev. | Hypo mean | Std. dev. | Hyper mean | Std. dev. | AUROC | AP |
|------------|-----------|-----------|-----------|------------|-----------|-------|-------|
| 0 | 0.08 | − 0.1 | 0.06 | 0.1 | 0.06 | 0.930 | 0.730 |
| 0 | 0.06 | − 0.3 | 0.04 | 0.3 | 0.04 | 0.945 | 0.852 |
| 0 | 0.06 | − 0.3 | 0.05 | 0.3 | 0.05 | 0.945 | 0.852 |
| 0 | 0.08 | − 0.3 | 0.06 | 0.3 | 0.06 | 0.945 | 0.852 |
| 0 | 0.1 | − 0.3 | 0.07 | 0.3 | 0.07 | 0.945 | 0.852 |
| 0 | 0.1 | − 0.3 | 0.08 | 0.3 | 0.08 | 0.945 | 0.852 |
| 0 | 0.1 | − 0.3 | 0.1 | 0.3 | 0.1 | 0.945 | 0.852 |
| 0 | 0.08 | − 0.5 | 0.06 | 0.5 | 0.06 | 0.893 | 0.775 |
| 0 | 0.1 | − 0.5 | 0.07 | 0.5 | 0.07 | 0.892 | 0.774 |

Table 4 Emission parameters for a HMM model with five hidden states

| State | Mean | Std. dev. |
|--------------------|--------|-----------|
| Equal | 0 | 0.08 |
| Hypo ₁ | − 0.25 | 0.06 |
| Hypo ₂ | − 0.5 | 0.06 |
| Hyper ₁ | 0.25 | 0.06 |
| Hyper ₂ | 0.5 | 0.06 |

Table 5 AUROC and AP when not specifying state distributions

| Number of states | AUROC | AP |
|-------------------------------------|-------|-------|
| 3 ¹ | 0.939 | 0.784 |
| 4 ¹ | 0.930 | 0.761 |
| 3 ² | 0.959 | 0.764 |
| 4 ² | 0.947 | 0.772 |
| 1 All regions | | |
| 2 Hypo- and hypermethylated regions | | |

Table 6 AUROC and AP for different emission distributions using M-values

| Equal mean | Std. dev. | Hypo mean | Std. dev. | Hyper mean | Std. dev. | AUROC | AP |
|------------|-----------|-----------|-----------|------------|-----------|--------------|--------------|
| 0 | 0.5 | − 1.2 | 0.5 | 1.2 | 0.5 | 0.942 | 0.810 |
| 0 | 0.5 | − 1.7 | 0.5 | 1.7 | 0.5 | 0.936 | 0.820 |
| 0 | 0.5 | − 2.2 | 0.5 | 2.2 | 0.5 | 0.927 | 0.816 |

Bold represent the highest values in each column

the higher AP over AUROC as it controls for the imbalance in the dataset. This indicates that although the range of values of methylation difference using beta-values is $[-1, 1]$, the normal distributions we used for the emission probabilities is able to sufficiently approximate the distribution of methylation differences.

Table 7 Running times

| No. of CpGs | No. of DMRs | Method | Input | Time (minutes) |
|-------------|-------------|--------|-------------|----------------|
| 214,878 | 910 | HMC | All regions | 89 |
| 214,878 | 910 | ADVI | All regions | 34 |
| 214,878 | 910 | ADVI | Hypo/hyper | 4 |
| 107,439 | 450 | ADVI | Hypo/Hyper | 2 |
| 107,439 | 910 | ADVI | Hypo/hyper | 3 |

Table 8 Design matrix for simulated data

| Intercept | Case/control | Binary | Continuous |
|-----------|--------------|--------|------------|
| 1 | 0 | 0 | 0.3 |
| 1 | 0 | 1 | 0.5 |
| 1 | 0 | 0 | 0.7 |
| 1 | 1 | 1 | 0.3 |
| 1 | 1 | 0 | 0.5 |
| 1 | 1 | 1 | 0.7 |

Running time

We measured the time it takes to run the analysis using as input chromosome 1 from the dataset by [14] using a single CPU. For comparison we also used as input only the first half of chromosome 1. The running time for the HMM step was negligible hence we only show here the computational times for the Bayesian analysis. We also compared the running times when using all regions and when only using candidate hypo- and hypermethylated regions. As shown in Table 7, using ADVI for posterior inference significantly reduces running time compared to HMC. Also, when using all regions the running time is significantly increased in comparison to just using candidate hypo- and hypermethylated regions. As expected, the running times are proportional to the number of CpGs analyzed such that halving the number of CpG sites (and DMRs) approximately halves running time. The number of DMRs also affects running time by increasing it.

Comparison of performance on simulated dataset with confounding covariates

To test the performance of LuxHMM in datasets with general experimental design we simulated a dataset with multiple covariates: (1) binary case/control, (2) arbitrary binary, (3) arbitrary continuous. The design matrix **D** is shown in Table 8. This simulation was modified from [5].

To model the varying lengths of methylated regions, the length L of the regions in terms of number of CpGs was sampled from $L \sim \text{ceiling}(\text{gamma}(\text{shape} = 4, \text{rate} = 0.2))$. The genomic coordinates were taken from the hg19 build. To model the varying differences in methylation levels, the covariate coefficients **b** were sampled from $\mathbf{b} \sim \mathcal{N}(\mu = 0, \sigma^2 = 5)$. For non-differentially methylated regions, the coefficient corresponding to the covariate of interest was set to zero. Conversely, for differentially methylated regions, the coefficient corresponding to the covariate of interest b was set so that $b < -3$ or $b > 3$ to ensure significant differential methylation. Finally, $\theta = \sigma(\mathbf{Y})$ where

Table 9 AUROC and AP for simulated dataset with confounding covariates

| Method | AUROC | AP |
|---------------------|--------------|--------------|
| LuxHMM ¹ | 0.823 | 0.536 |
| LuxHMM ² | 0.756 | 0.549 |
| LuxUS | 0.679 | 0.321 |
| RADMeth | 0.644 | 0.246 |
| metilene | 0.714 | 0.348 |
| HMM-DM | 0.616 | 0.180 |
| DMRcate | 0.658 | 0.065 |
| DSS | 0.672 | 0.339 |

Bold represent the highest values in each column

¹ All regions

² Hypo- and hypermethylated regions

$\mathbf{Y} \sim \mathcal{N}(\mathbf{D}\mathbf{b}, \sigma_E^2)$ where $\sigma_E^2 \sim \text{gamma}(\text{shape} = 0.5, \text{scale} = 1)$. Around 1700 DMRs were added to the genome.

In LuxHMM, either all regions or only candidate hypo- and hypermethylated regions, as classified by HMM, were used as input in determining DMRs. Parameter settings for competing methods are described in Additional file 1: Section 3.

AUROC and AP, to handle the imbalance in the dataset given that there are much more negative than positive samples, were computed (Table 9). For AP, the baseline is 0.0014. LuxHMM using all regions generated the highest AUROC and LuxHMM using just candidate hypo- and hypermethylated regions generated the highest AP. This indicates that, like in Section 3.1, using LuxHMM with all regions has a higher recall whereas using LuxHMM with just candidate hypo- and hypermethylated regions has a higher precision. This also shows that LuxHMM is able to more accurately detect DMRs from a dataset with confounding covariates.

Comparison of performance on real BS-seq data with confounding covariates

To test the performance of LuxHMM on real BS-seq data with multiple covariates we evaluated the different statistical methods in terms of gene set enrichment using the webtool GREAT [15] on the dataset with GEO accession number GSE47966 as originally performed by [16]. The dataset consists of samples taken from mice brain tissue (WGBS). Three samples consisted of neuron cells and three consisted of non-neuron cells. In addition, the samples were divided into male and female mice and different ages (6 week and 12 month old females, and 7 week old males). DMRs between neurons and non-neurons were identified using the different methods and then gene ontology (GO) enrichment were performed to test the ability of the various methods to identify biologically relevant regions. The top 25 and 60 enriched GO terms based on binomial ranking were taken and the percentage of GO terms related to the neural system were determined. Gene set enrichment analysis were performed with mouse phenotype annotations.

In LuxHMM, candidate hypo- and hypermethylated regions, as determined by HMM, were used as input in determining differentially methylated regions. HMC was used to sample from the posterior distribution with four chains, 1000 iterations for warmup for each chain and a total of 1000 iterations for sampling. In addition, as in [16], for the

Table 10 Enriched GO terms related to the neural system

| Method | Top 25 (%) | Top 60 (%) |
|---------------------|------------|------------|
| LuxHMM ¹ | 92 | 83 |
| LuxUS | 88 | 85 |
| RADMeth | 88 | 80 |
| Metilene | 20 | 32 |
| HMM-DM | 96 | 93 |
| DMRcate | 84 | 68 |
| DSS | 88 | 87 |

Bold represent the highest values in each column

¹ Hypo- and hypermethylated regions

regions, a threshold of > 25 CpGs was used. To make a comparable assessment, the top 10,000 to 15,000 DMRs from all methods were used as input to GREAT. Parameter settings for competing methods are described in Additional file 1: Section 4.

As shown in Table 10, HMM-DM generated the highest percentages of enriched GO terms related to the neural system in both the top 25 and top 60 enriched GO terms. In the top 25 enriched GO terms, LuxHMM generated the second highest number of enriched GO terms related to the neural system and in the top 60 LuxHMM was fourth highest after DSS and LuxUS (Additional file 2). This shows that LuxHMM performs comparatively well in finding biologically relevant regions relative to other methods tested.

Conclusions

We propose the tool LuxHMM for detecting differentially methylated regions. This tool uses HMM to segment the genome into regions with hypomethylation, hypermethylation and equal methylation between two groups and Bayesian regression for evaluating differential methylation. Further, model inference is done using either variational inference for efficient genome-scale analysis or HMC.

We show using simulated and real BS-seq data with general experimental designs that LuxHMM outperforms other published methods in detecting differentially methylated regions from simulated datasets and performs comparatively well in a real dataset.

Abbreviations

ADVI: Automatic differentiation variational inference
HMC: Hamiltonian Monte Carlo
BS-seq: Bisulfite sequencing

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-023-05174-7>.

Additional file 1: Description of data: (i) Initial state transition probabilities, (ii) Parameter settings for competing methods on simulated dataset based on real BS-seq data, (iii) Parameter settings for competing methods on simulated dataset with confounding covariates, and (iv) Parameter settings for competing methods on real BS-seq data with confounding covariates

Additional file 2: Enriched GO terms using Mouse Phenotype

Acknowledgements

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

Author Contributions

MM and HL developed the method. MM implemented the method. MM and HL wrote the manuscript. Both authors read and approved the final version of the manuscript.

Funding

This work was supported by the Ella and Georg Ehrnrooth Foundation and the Academy of Finland (grant number 314445). The funding body played no role in the design of the study, the collection, analysis, interpretation of data, or in writing the manuscript.

Availability of data and materials

LuxHMM is open source and freely available from <https://github.com/malonzm1/LuxHMM>. Project name: LuxHMM. Project home page: <https://github.com/malonzm1/LuxHMM>. Operating system(s): Linux. Programming language: Python, Stan. Other requirements: CmdStan (tested on version 2.29.0), Python (tested on version 3.8.12), pomegranate (tested on version 0.14.8), pystan (tested on version 3.1.1), Numpy (tested on version 1.22.2), Scipy (tested on version 1.8.0), cmdstanpy (tested on version 1.0.1). LuxHMM is freely available at <https://github.com/malonzm1/LuxHMM> along with documentation. License: GNU GPL Any restrictions to use by non-academics: Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publication

Not applicable.

Competing interest

The authors declare that they have no financial or non-financial competing interests.

Received: 5 November 2022 Accepted: 6 February 2023

Published online: 22 February 2023

References

- Jeschke J, Evelyne C, Francois F. Dna methylome profiling beyond promoters-taking an epigenetic snapshot of the breast tumor microenvironment. *FEBS J*. 2015;282:1801–14.
- Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, Zhou X. Statistical methods for detecting differentially methylated loci and regions. *Front Genet*. 2014;5:324.
- Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multi-factor whole-genome bisulfite sequencing experiments. *BMC Bioinf*. 2014;15:1–8.
- Juhling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. Metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res*. 2016;26:256–62.
- Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord RV, Clark SJ, Molloy PL. De novo identification of differentially methylated regions in the human genome. *Epigenet Chromatin*. 2015;8:1–16.
- Park Y, Wu H. Differential methylation analysis for bs-seq data under general experimental design. *Bioinformatics*. 2016;32:1446–53.
- Äijö T, Yue X, Rao A, Lähdesmäki H. Luxglm: a probabilistic covariate model for quantification of dna methylation modifications with complex experimental designs. *Bioinformatics*. 2016;32:511–9.
- Halla-Aho V, Lähdesmäki H. Luxus: Dna methylation analysis using generalized linear mixed model with spatial correlation. *Bioinformatics*. 2020;36:4535–43.
- Jurafsky D, Martin JH. Speech and language processing. 2nd edn. Upper Saddle River, NJ: Prentice-Hall, Inc.; 2009.
- Bilmes JA. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Berkeley, CA: International Computer Science Institute, Berkely, CA; 1998.
- Schreiber J. Pomegranate: fast and flexible probabilistic modeling in python. *J Mach Learn Res*. 2018;18:1–6.
- Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. Stan: A probabilistic programming language. *J Stat Softw*. 2017;76:1.
- Kucukelbir A, Ranganath R, Gelman A, Blei D. Automatic variational inference in stan. In: Cortes C, Lee DD, Sugiyama M, editors. *Advances in neural information processing systems 28 (NIPS 2015)*, neural information processing systems. 2015;pp. 568–576.
- Klein HU, Hebestreit K. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. *Brief Bioinf*. 2016;17:796–807.
- McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. Great improves functional interpretation of cis-regulatory regions. *Nature Biotechnol*. 2010;28:495–501.
- Wen Y, Chen F, Zhang Q, Zhuang Y, Li Z. Detection of differentially methylated regions in whole genome bisulfite sequencing data using local getis-ord statistics. *Bioinformatics*. 2016;32:3396–404.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.