Bioinformatics, 32, 2016, i511–i519
doi: 10.1093/bioinformatics/btw468
ECCB 2016

OXFORD

LuxGLM: a probabilistic covariate model for quantification of DNA methylation modifications with complex experimental designs

Tarmo Äijö,[1,2,*] Xiaojing Yue,[3] Anjana Rao[3,4,5] and Harri Lähdesmäki[2,*]

[1]Center for Computational Biology, Simons Foundation, New York, NY 10010, USA, [2]Department of Computer Science, Aalto University School of Science, Aalto FI-00076, Finland, [3]La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA, [4]Department of Pharmacology and Moores Cancer Center, University of California, La Jolla, CA 92037, USA and [5]Sanford Consortium for Regenerative Medicine, La Jolla, CA 92037, USA

*To whom correspondence should be addressed.

Abstract

Motivation: 5-methylcytosine (5mC) is a widely studied epigenetic modification of DNA. The ten-eleven translocation (TET) dioxygenases oxidize 5mC into oxidized methylcytosines (oxi-mCs): 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). DNA methylation modifications have multiple functions. For example, 5mC is shown to be associated with diseases and oxi-mC species are reported to have a role in active DNA demethylation through 5mC oxidation and DNA repair, among others, but the detailed mechanisms are poorly understood. Bisulphite sequencing and its various derivatives can be used to gain information about all methylation modifications at single nucleotide resolution. Analysis of bisulphite based sequencing data is complicated due to the convoluted read-outs and experiment-specific variation in biochemistry. Moreover, statistical analysis is often complicated by various confounding effects. How to analyse 5mC and oxi-mC data sets with arbitrary and complex experimental designs is an open and important problem.

Results: We propose the first method to quantify oxi-mC species with arbitrary covariate structures from bisulphite based sequencing data. Our probabilistic modeling framework combines a previously proposed hierarchical generative model for oxi-mC-seq data and a general linear model component to account for confounding effects. We show that our method provides accurate methylation level estimates and accurate detection of differential methylation when compared with existing methods. Analysis of novel and published data gave insights into to the demethylation of the forkhead box P3 (*Foxp3)* locus during the induced T regulatory cell differentiation. We also demonstrate how our covariate model accurately predicts methylation levels of the *Foxp3* locus. Collectively, LuxGLM method improves the analysis of DNA methylation modifications, particularly for oxi-mC species.

Availability and Implementation: An implementation of the proposed method is available under MIT license at https://github.org/tare/LuxGLM/

Contact: taijo@simonsfoundation.org or harri.lahdesmaki@aalto.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

5-methylcytosine (5mC) is a widely studied epigenetic modification of DNA, which controls mammalian development, X-chromosome inactivation, gene imprinting and genomic instability (Smith and Meissner, 2013). DNA methylation research was revolutionized by the discovery that the members of the ten-eleven translocation (TET) protein family oxidise 5mC sequentially into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)

(He *et al.*, 2011; Ito *et al.*, 2011; Tahiliani *et al.*, 2009). These oxidized methylcytosines (oxi-mC) have been proposed to play a role in active DNA demethylation through 5mC oxidation and DNA repair, and in chromatin regulation (Pastor *et al.*, 2013). 5mC and all the oxi-mC species are of great interest due to the alleged role of DNA methylation in diseases, such as different cancers (Baylin, 2005), Alzheimer (De Jager *et al.*, 2014), asthma (Rastogi *et al.*, 2013), autism (Nardone *et al.*, 2014) and type 2 diabetes (Dayeh *et al.*, 2014). However, studies of primary human clinical samples are complicated by many factors; for instance, greater biological variation compared with more controlled molecular biology studies, possible confounding factors and case-control matching.

Bisulphite sequencing (BS-seq) has become the gold standard technique for profiling methylation at single nucleotide resolution (Lister *et al.*, 2009, 2013; Rein *et al.*, 1998). In BS-seq, genomic DNA is treated with sodium bisulphite, which will rapidly deaminate unmodified cytosine (and 5fC and 5caC) to uracil, while deamination of 5mC and 5hmC are much slower (Frommer *et al.*, 1992). Next, after PCR amplification, uracil and cytosine are read as thymine and cytosine, respectively. Importantly, 5fC and 5caC will have the same read-out as unmodified cytosine and, similarly, 5hmC and 5mC share the same read-out in BS-seq (Huang *et al.*, 2010). This observation drove the development of various modified bisulphite sequencing protocols (reviewed in Plongthongkum *et al.*, 2014). For instance, oxidative bisulphite sequencing (oxBS-seq) (Booth *et al.*, 2012) and Tet-assisted bisulphite sequencing (TAB-seq) (Yu *et al.*, 2012) were developed for distinguishing 5hmC from 5mC. Both methods, oxBS-seq and TAB-seq, are based on oxidation; 5hmC is oxidised into 5fC by $KRuO_4$ in oxBS-seq, whereas in TAB-seq 5mC is oxidised into 5caC by recombinant mouse *Tet1*. To gain information on 5fC, 5fC chemical modification-assisted bisulphite sequencing (fCAB-seq) (Lu *et al.*, 2013) and reduced bisulphite sequencing (redBS-seq) (Booth *et al.*, 2014) have been proposed. Chemical modification-assisted bisulphite sequencing (CAB-seq) together with BS-seq allows the quantification of 5caC by protecting 5caC from deamination by sodium bisulphite with 1-ethyl-3-[3-dimethylaminopropyl]-carbodiimide hydrochloride (Lu *et al.*, 2013). CpG methyltransferase (M.SssI) assisted bisulphite sequencing (MAB-seq) when combined with BS-seq distinguishes 5fC/5caC from C (Wu *et al.*, 2014). A summary of the read-outs of the described bisulphite sequencing approaches is listed in Figure 1A.

In order to estimate proportions of multiple methylation modifications, one has to deconvolute and integrate data from multiple bisulphite based measurements (Fig. 1A) which often have biases due to imperfect experimental steps (Plongthongkum *et al.*, 2014). Many computational methods have been developed for analysing the standard bisulphite sequencing data (here we will describe only the most relevant methodologies, for a more comprehensive list of different methods see Äijö *et al.*, 2016). Methods based on beta-binomial models have been proposed allowing modeling of sampling and biological variation. For instance, MOABS uses a hierarchical beta-binomial model with an empirical Bayesian approach (Sun *et al.*, 2014). To assess differential methylation, MOABS uses credible methylation difference metric for summarizing statistical and biological significance (Sun *et al.*, 2014). Another method, RADMeth, takes into account covariates under the beta-binomial model using a generalised linear model approach with the logit link function (Dolzhenko and Smith, 2014). RADMeth detects differential methylation by using the log-likelihood ratio test and the evidence for differential methylation across neighbouring cytosines is shared using the Stouffer-Liptak weighted Z test. Recently, the MACAU method was proposed, which combines a binomial mixed model with a sampling-based inference algorithm to model various genetic relatedness/population structures (Lea *et al.*, 2015). MACAU uses Wald test statistics on the posterior samples to call whether a covariate has an effect on methylation (Lea *et al.*, 2015).

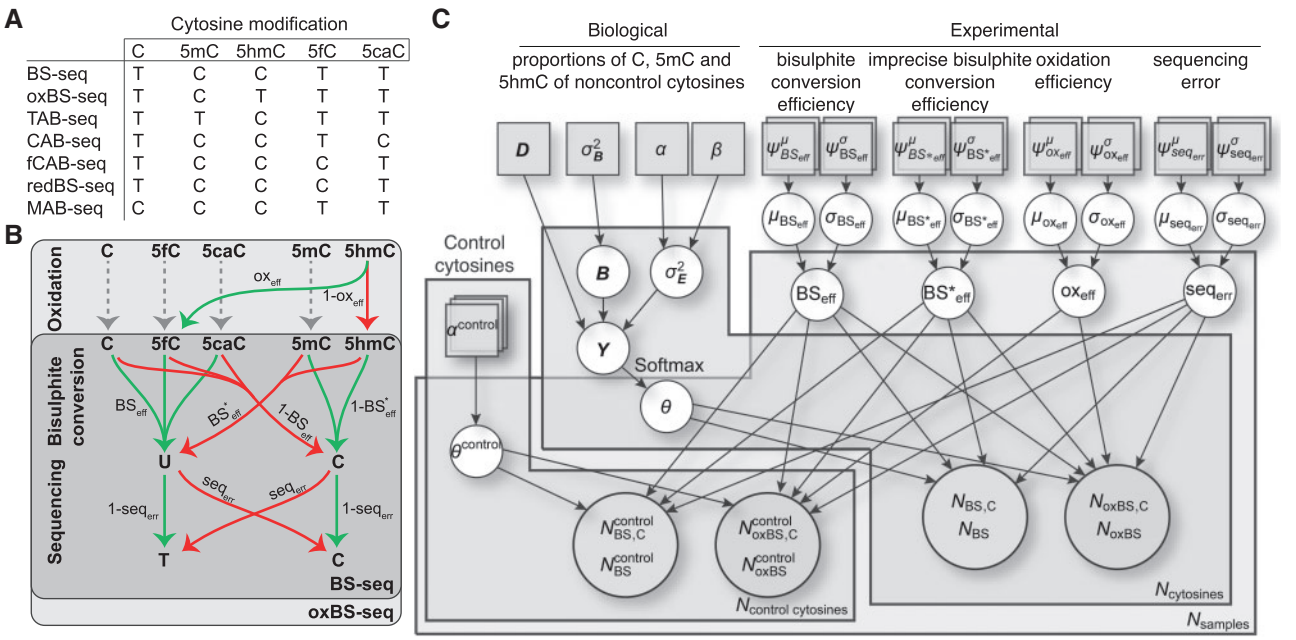

Fig. 1. (**A**) The conversion chart of C, 5mC, 5hmC, 5fC and 5caC in BS-seq, oxBS-seq, TAB-seq, CAB-seq, fCAB-seq, redBS-seq and MAB-seq experiments. (**B**) The experimental steps of BS- and oxBS-seq experiments are represented in terms of experimental parameters. Green and red arrows depict successful and unsuccessful steps, respectively. (**C**) The proposed hierarchical model for modeling methylation modification proportions for BS-seq and oxBS-seq data and parts of the original Lux model represented in the plate notation. The grey and white circles are used to represent observed variables and latent variables, respectively. The grey squares represent fixed hyperparameters. The components, which model the experimental parameters and control cytosines are the same as in the Lux model (Äijö *et al.*, 2016)

Neither MOABS, RADMeth nor MACAU support model-based consideration of experimental parameters or deconvolution of multiple oxi-mC data types. MLML partially solves the latter limitation by integrating BS-seq and oxBS-seq or BS-seq and TAB-seq data to provide consistent C, 5mC, and 5hmC estimates but it ignores other oxi-mC species and imperfect experimental parameters and does not support analysis of replicates or differential methylation (Qu *et al.*, 2013). These limitations motivated us to develop the Lux methodology, which integrates different combinations of oxi-mC measurements (BS-seq, oxBS-seq, TAB-seq, redBS-seq, fCAB-seq, CAB-seq and MAB-seq) while taking into account the relevant experimental parameters (Äijö *et al.*, 2016). Lux is expressed as a hierarchical generative model based on the experimental steps involved in the measurement protocols. Specifically, Lux uses two hierarchical Dirichlet components to model methylation modification proportions and their variation between biological replicates. The first Dirichlet component is used together with a gamma distributed random variable to model the pseudo-count parameter of the second Dirichlet component, which models the actual DNA methylation modification proportions.

Here, we propose LuxGLM which uses a logistic matrix normal distribution and a general linear model (GLM) to model the hierarchical structure across replicates, conditions, and confounding factors. This extension has several important implications. For example, first, it allows for modeling covariates through the GLM and, second, it enables modeling covariation between the methylation modification proportions through the logistic matrix normal distribution. The consideration of covariates decreases false positives and increases true positives when confounding effects are present as we demonstrate on synthetic data. Analysis of novel and published BS-seq and oxBS-seq data gave insights into the demethylation of the forkhead box P3 (*Foxp3*) locus during induced T regulatory cell differentiation. We used the inferred model to produce a testable prediction on the methylation of the *Foxp3* locus and validated the prediction using independent data. LuxGLM can also be applied in a reduced setting (only BS-seq data is available and by ignoring imperfect experimental parameters), which corresponds to the standard BS-seq based DNA methylation analysis. A comparison with the state-of-the-art methods on this reduced setting illustrates that LuxGLM achieves similar or better performance. In contrast to previous methods, LuxGLM generalises for an integrative analysis of multiple oxi-mC data types and, provided proper spike-in control cytosines are included in experimental protocols, LuxGLM also supports model-based analysis of experiment-specific variation in biochemistry. To summarize, LuxGLM improves the analysis of methylomes from complex experimental designs and provides a comprehensive tool for accurate quantification of all DNA methylation modifications.

2 Methods

2.1 Probabilistic generative model for DNA methylation quantification

In this section, we will first briefly review the Lux model (Äijö *et al.*, 2016) before introducing the covariate-aware LuxGLM model. The novel idea of Lux was to develop a generative model for oxi-mC-seq data. Lux models the effects of the experimental steps (bisulphite conversion, oxidation, chemical labeling, protection steps, etc.) through their efficiencies on the sequencing read-outs (Fig. 1B and C). For instance, the conditional probabilities of getting 'C' as a read-out from BS-seq and oxBS-seq experiments given the methylation status of a cytosine is C (or 5fC, 5caC), 5mC or 5hmC are

$$\begin{aligned}
p_{\mathrm{BS}}(\text{“C”}|\mathrm{C}) &= (1-\mathrm{BS}_{\mathrm{eff}})(1-\mathrm{seq}_{\mathrm{err}}) + \mathrm{BS}_{\mathrm{eff}}\mathrm{seq}_{\mathrm{err}},\\
p_{\mathrm{BS}}(\text{“C”}|5\mathrm{mC}) &= (1-\mathrm{BS}^{*}_{\mathrm{eff}})(1-\mathrm{seq}_{\mathrm{err}}) + \mathrm{BS}^{*}_{\mathrm{eff}}\mathrm{seq}_{\mathrm{err}},\\
p_{\mathrm{BS}}(\text{“C”}|5\mathrm{hmC}) &= (1-\mathrm{BS}^{*}_{\mathrm{eff}})(1-\mathrm{seq}_{\mathrm{err}}) + \mathrm{BS}^{*}_{\mathrm{eff}}\mathrm{seq}_{\mathrm{err}},\\
p_{\mathrm{oxBS}}(\text{“C”}|\mathrm{C}) &= (1-\mathrm{BS}_{\mathrm{eff}})(1-\mathrm{seq}_{\mathrm{err}}) + \mathrm{BS}_{\mathrm{eff}}\mathrm{seq}_{\mathrm{err}},\\
p_{\mathrm{oxBS}}(\text{“C”}|5\mathrm{mC}) &= (1-\mathrm{BS}^{*}_{\mathrm{eff}})(1-\mathrm{seq}_{\mathrm{err}}) + \mathrm{BS}^{*}_{\mathrm{eff}}\mathrm{seq}_{\mathrm{err}},\\
p_{\mathrm{oxBS}}(\text{“C”}|5\mathrm{hmC}) &= \mathrm{ox}_{\mathrm{eff}}[(1-\mathrm{BS}_{\mathrm{eff}})(1-\mathrm{seq}_{\mathrm{err}}) + \mathrm{BS}_{\mathrm{eff}}\mathrm{seq}_{\mathrm{err}}]\\
&\quad + (1-\mathrm{ox}_{\mathrm{eff}})[(1-\mathrm{BS}^{*}_{\mathrm{eff}})(1-\mathrm{seq}_{\mathrm{err}})\\
&\quad + \mathrm{BS}^{*}_{\mathrm{eff}}\mathrm{seq}_{\mathrm{err}}],
\end{aligned} \tag{1}$$

where $\mathrm{BS}_{\mathrm{eff}}$, $\mathrm{BS}^{*}_{\mathrm{eff}}$, $\mathrm{ox}_{\mathrm{eff}}$ and $\mathrm{seq}_{\mathrm{err}}$ are bisulphite conversion efficiency, inaccurate bisulphite conversion efficiency, oxidation efficiency, and sequencing error, respectively (Fig. 1B). In practice, we are interested in estimating the underlying unobserved methylation modification proportions $\theta = (p(\mathrm{C}), p(5\mathrm{mC}), p(5\mathrm{hmC}))$ $(\sum \theta = 1)$ in a sample or among replicates [Note that with BS-seq and oxBS-seq data one can only quantify $p(\mathrm{C})$, $p(5\mathrm{mC})$ and $p(5\mathrm{hmC})$ and that $p(\mathrm{C}) \equiv p(\mathrm{C}) + p(5\mathrm{fC}) + p(5\mathrm{caC})$ but the model generalises to 5fC and 5caC too.]. The (unconditional) binomial parameters, $p_{\mathrm{BS}}(\text{“C”})$ and $p_{\mathrm{oxBS}}(\text{“C”})$, for the sequencing data generation are obtained by applying the total probability theorem

$$\begin{aligned}
p_{\mathrm{BS}}(\text{“C”}) &= p(\mathrm{C})p_{\mathrm{BS}}(\text{“C”}|\mathrm{C}) + p(5\mathrm{mC})p_{\mathrm{BS}}(\text{“C”}|5\mathrm{mC})\\
&\quad + p(5\mathrm{hmC})p_{\mathrm{BS}}(\text{“C”}|5\mathrm{hmC})\\
p_{\mathrm{oxBS}}(\text{“C”}) &= p(\mathrm{C})p_{\mathrm{oxBS}}(\text{“C”}|\mathrm{C}) + p(5\mathrm{mC})p_{\mathrm{oxBS}}(\text{“C”}|5\mathrm{mC})\\
&\quad + p(5\mathrm{hmC})p_{\mathrm{oxBS}}(\text{“C”}|5\mathrm{hmC}).
\end{aligned} \tag{2}$$

The 'C' read-out counts are then distributed as $N_{\mathrm{BS,C}} \sim B(N_{\mathrm{BS}}, p_{\mathrm{BS}}(\text{“C”}))$ and $N_{\mathrm{oxBS,C}} \sim B(N_{\mathrm{oxBS}}, p_{\mathrm{oxBS}}(\text{“C”}))$ for BS-seq and oxBS-seq data, respectively, where $B(\cdot,\cdot)$ denotes the binomial distribution (Fig. 1C) (similarly for control cytosine). Details of the priors for experimental parameters are summarized in Supplementary Equations (S2–S5).

The goal is to estimate the cytosine and condition specific θ parameters simultaneously with the sample specific experimental parameters ($\mathrm{BS}_{\mathrm{eff}}$, $\mathrm{BS}^{*}_{\mathrm{eff}}$, $\mathrm{ox}_{\mathrm{eff}}$ and $\mathrm{seq}_{\mathrm{err}}$) (sample and cytosine indices are omitted) (Fig. 1C). Previously, we have demonstrated that the experimental parameters can be estimated from spike-in control cytosines by providing prior knowledge, $\alpha^{\mathrm{control}}$, on the methylation levels of control cytosines, $\theta^{\mathrm{control}}$ (Äijö *et al.*, 2016).

2.2 Matrix normal distribution

The matrix normal distribution is a generalisation of the multivariate normal distribution to matrix-valued random variables

$$\mathbf{X} \sim \mathcal{MN}(\mathbf{M}, \mathbf{U}, \mathbf{V}), \tag{3}$$

where $\mathbf{M} \in \mathbb{R}^{N\times M}$ is the location matrix and $\mathbf{U} \in \mathbb{R}^{N\times N}_{\mathrm{pos-def}}$ and $\mathbf{V} \in \mathbb{R}^{M\times M}_{\mathrm{pos-def}}$ are scale matrices. The first and second moments of $\mathbf{X}$ are

$$\begin{aligned}
\mathrm{E}[\mathbf{X}] &= \mathbf{M},\\
\mathrm{E}\left[(\mathbf{X}-\mathbf{M})(\mathbf{X}-\mathbf{M})^{\mathrm{T}}\right] &= \mathbf{U}\,\mathrm{tr}(\mathbf{V}),\\
\mathrm{E}\left[(\mathbf{X}-\mathbf{M})^{\mathrm{T}}(\mathbf{X}-\mathbf{M})\right] &= \mathbf{V}\,\mathrm{tr}(\mathbf{U}),
\end{aligned} \tag{4}$$

where $\mathrm{tr}(\cdot)$ is the matrix trace. The matrix-valued random variable $\mathbf{X} \in \mathbb{R}^{N\times M}$ in Equation (3) can be stated equivalently as a vector-valued random variable $\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{NM}$

$$\mathrm{vec}\,(\mathbf{X}) \sim \mathcal{N}(\mathrm{vec}\,(\mathbf{M}), \mathbf{V} \otimes \mathbf{U}), \tag{5}$$

where $\mathrm{vec}\,(\cdot)$ is the vectorization operator and $\otimes$ is the Kronecker product (Gupta and Nagar, 1999).

2.3 Logistic normal distribution

Let us consider that $\mathbf{x} \in \mathbb{R}^M$ is distributed according to the multivariate normal distribution

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma), \tag{6}$$

where $\mu \in \mathbb{R}^M$ is the location parameter and $\Sigma \in \mathbb{R}^{M\times M}_{\mathrm{pos-def}}$ is the covariance matrix. Next, let us apply the softmax transformation on $\mathbf{x}$

$$\theta = \mathrm{Softmax}\,(x) = \left(\frac{\exp(x_1)}{\Sigma_{k=1}^{M}\exp(x_k)}, \ldots, \frac{\exp(x_M)}{\Sigma_{k=1}^{M}\exp(x_k)}\right)^{\mathrm{T}}. \tag{7}$$

Then, the following two statements are true: (1) $\theta \in [0, 1]^M$ and (2) $\Sigma_{k=1}^{M}\theta_i = 1$. In other words, $\theta \in \mathcal{S}^M$ is a M-dimensional simplex-valued random variable.

2.4 General linear model

The GLM has the following form

$$\mathbf{Y} = \mathbf{DB} + \mathbf{E}, \tag{8}$$

where $\mathbf{Y}$ is a matrix containing multivariate measurements, $\mathbf{D}$ is a design matrix, $\mathbf{B}$ is a parameter matrix, and $\mathbf{E}$ is a noise matrix. Bayesian inference of the model in Equation (8) requires that we specify modeling assumptions and set the priors. Often, it is assumed that $\mathbf{E}$ and $\mathbf{B}$ have the following prior distributions

$$\mathbf{E}|\mathbf{U}_\mathbf{E}, \mathbf{V}_\mathbf{E} \sim \mathcal{MN}(\mathbf{0}, \mathbf{U}_\mathbf{E}, \mathbf{V}_\mathbf{E}) \tag{9}$$

and

$$\mathbf{B}|\mathbf{M}_\mathbf{B}, \mathbf{U}_\mathbf{B}, \mathbf{V}_\mathbf{B} \sim \mathcal{MN}(\mathbf{M}_\mathbf{B}, \mathbf{U}_\mathbf{B}, \mathbf{V}_\mathbf{B}). \tag{10}$$

Under these assumptions

$$\begin{aligned}\mathrm{vec}(\mathbf{Y})|\mathbf{D}, \mathbf{M}_\mathbf{B}, \mathbf{U}_\mathbf{B}, \mathbf{V}_\mathbf{B}, \mathbf{U}_\mathbf{E}, \mathbf{V}_\mathbf{E} \sim \mathcal{N}((\mathbf{I}\otimes\mathbf{D})\mathrm{vec}\,(\mathbf{M}_\mathbf{B}),\\ (\mathbf{I}\otimes\mathbf{D})(\mathbf{V}_\mathbf{B}\otimes\mathbf{U}_\mathbf{B})(\mathbf{I}\otimes\mathbf{D})^{\mathrm{T}} + \mathbf{V}_\mathbf{E}\otimes\mathbf{U}_\mathbf{E}),\end{aligned} \tag{11}$$

where we have used the property $\mathrm{vec}\,(\mathbf{DB}) = (\mathbf{I}\otimes\mathbf{D})\mathrm{vec}\,(\mathbf{B})$. Finally, one can also specify hyperpriors for the hyperparameters $\mathbf{M}_\mathbf{B}$, $\mathbf{U}_\mathbf{B}$, $\mathbf{V}_\mathbf{B}$, $\mathbf{U}_\mathbf{E}$, and $\mathbf{V}_\mathbf{E}$.

2.5 Probabilistic generative covariate model for DNA methylation quantification

Here, we describe our covariate approach to model oxi-mC data. The statistical models of LuxGLM and Lux have common components in that given the methylation modification proportions θ (and $\theta^{\mathrm{control}}$) the 'C' read-out counts as well as the experimental parameters are handled similarly (Fig. 1C). Lux's hierarchical Dirichlet component allows model-based analysis of replicates but it does not support covariates. LuxGLM generalises the model for methylation modification proportions θ by incorporating the GLM to account for complex covariate structures.

In contrast to Lux, we model M methylation modification proportions θ_i ($\sum\theta_i = 1$) over N samples using the GLM given in Equation (8) that supports covariates and the softmax transformation defined in Equation (7) which is needed for mapping real-valued vectors into simplex-valued vectors θ. Let us assume that there are P covariates, hence $\mathbf{Y} \in \mathbb{R}^{N\times M}$, $\mathbf{D} \in \mathbb{R}^{N\times P}$, and $\mathbf{B} \in \mathbb{R}^{P\times M}$. Naturally, the design matrix $\mathbf{D}$ relates N samples to P covariates. Finally, the parameter $\theta_i \in \mathcal{S}^M$ is given by $\mathrm{Softmax}\,(\mathrm{row}_i(\mathbf{Y}))$, where $\mathrm{row}_i(\mathbf{Y})$ is the ith row of $\mathbf{Y}$.

Now we will describe how we choose the priors. First, we assume that $\mathbf{E}$ is distributed as follows

$$\mathrm{vec}\,(\mathbf{E}) \sim \mathcal{N}(\mathrm{vec}\,(\mathbf{0}), \sigma^2_\mathbf{E}(\mathbf{I}\otimes\mathbf{I})). \tag{12}$$

Second, we assume an inverse gamma prior on $\sigma^2_\mathbf{E}$

$$\sigma^2_\mathbf{E} \sim \Gamma^{-1}(\alpha, \beta), \tag{13}$$

where $\alpha = \beta = 1$. Third, we assume that $\mathbf{B}$ is distributed as follows

$$\mathrm{vec}\,(\mathbf{B}) \sim \mathcal{N}(\mathrm{vec}\,(0), \sigma^2_\mathbf{B}(\mathbf{I}\otimes\mathbf{I})), \tag{14}$$

where $\sigma^2_\mathbf{B} = 5$. Consequently, $\mathbf{Y}$ is distributed as follows

$$\mathrm{vec}(\mathbf{Y})|\mathbf{D}, \sigma^2_\mathbf{B}, \sigma^2_\mathbf{E} \sim \mathcal{N}(\mathrm{vec}\,(\mathbf{0}), \sigma^2_\mathbf{B}(\mathbf{I}\otimes\mathbf{D})(\mathbf{I}\otimes\mathbf{I})(\mathbf{I}\otimes\mathbf{D})^{\mathrm{T}} + \sigma^2_\mathbf{E}(\mathbf{I}\otimes\mathbf{I})). \tag{15}$$

To illustrate the selected $\mathbf{B}$ prior [Equation (14)], we visualise the softmax transformed independent and identically distributed 3-dimensional normal random variable with zero mean and variance 5 (Supplementary Fig. S1). The prior on $\mathbf{B}$ will have an effect on the sensitivity of the methylation level estimates θ_i. That is, if the value of the variance parameter $\sigma^2_\mathbf{B}$ in Equation (16) is small, then the prior of $\mathrm{vec}\,(\mathbf{Y})$ is concentrated more around zero or, equivalently, θ_i is concentrated around $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Consequently, it will take more observations (deeper sequencing depth or more replicates) to update the posterior distribution away from the prior. The graphical model implemented in LuxGLM is shown in Figure 1C.

2.6 Model inference

Bayesian inference of the aforedescribed model is analytically intractable. Therefore, to estimate the posterior distributions, we use Hamiltonian Monte Carlo (HMC) as implemented in Stan's No-U-Turn sampler (NUTS) to obtain posterior samples (Carpenter *et al.*, in press). Convergence of the chains was monitored using the Gelman-Rubin statistic ($\widehat{R} < 1.1$). Importantly, prior and data contributions on the parameters are fully summarised in the parameter posterior distributions allowing us to estimate different statistical measures on the distributions, such as mean and credible intervals.

2.7 Detection of differential methylation

Similar to our earlier work Äijö *et al.* (2016), we use the Savage-Dickey density ratio to quantify evidence of differential methylation. Specifically, we consider two schemes of testing differential methylation: (i) testing whether two covariates differ and (ii) testing whether a covariate has an effect (differs from zero).

Let us start with case (i) by assuming that we are interested in assessing differential methylation between two conditions C_1 and C_2. Without loss of generality, let C_1 and C_2 be represented in the design matrix $\mathbf{D}$ by the ith and jth covariates, respectively. Then, $\mathrm{row}_i(\mathbf{B})$ and $\mathrm{row}_j(\mathbf{B})$ correspond to the ith and jth covariates, respectively. To assess the difference in methylation between the conditions C_1 and C_2, we study the difference of the random variables $\mathrm{row}_i(\mathbf{B}) - \mathrm{row}_j(\mathbf{B}) \equiv C_1 - C_2$. Next, we will formulate two hypotheses: (i) the null hypothesis $H_0 : C_1 - C_2 = \mathbf{0}$ (no differential methylation) and (ii) the alternative hypothesis $H_1 : C_1 - C_2 \neq \mathbf{0}$ (differential methylation). To approximate the Bayes factor (BF) between the models representing the hypotheses H_0 and H_1, we use the Savage-Dickey density ratio

$$\mathrm{BF} \approx \frac{p(C_1 - C_2 = \mathbf{0}|H_1)}{p(C_1 - C_2 = \mathbf{0}|H_1, \mathcal{D})}, \tag{16}$$

where $\mathcal{D}$ denotes data. The numerator is analytically solvable under the assumption of the normal prior defined in Equation (14), whereas, the denominator is estimated using the standard kernel density estimation with the normal kernel on the samples from the posterior obtained through HMC sampling (Äijö *et al.*, 2016). Notable, the value of σ_B^2 in Eq. (14) has an effect on the approximation of the BFs (16) through the numerator. That is, the variance σ_B^2 commensurate with the value of the numerator; thus, the approximated BFs tend to decrease as σ_B^2 increases (albeit σ_B^2also has an effect on the denominator through the posterior).

Next, we will describe how differential methylation of case (ii) is quantified. Let us assume that the condition C_1 is represented in the design matrix $\mathbf{D}$ by the ith covariate. Consequently, $\mathrm{row}_i(\mathbf{B}) \equiv C_1$ corresponds to the ith covariate. Similarly to the previous case, we formulate two hypotheses: (i) the null hypothesis $H_0 : C_1 = \mathbf{0}$ (no differential methylation) and (ii) the alternative hypothesis $H_1 : C_1 \neq \mathbf{0}$ (differential methylation). The Savage-Dickey density ratio for approximating the BF in this case is

$$\mathrm{BF} \approx \frac{p(C_1 = \mathbf{0}|H_1)}{p(C_1 = \mathbf{0}|H_1, \mathcal{D})}. \tag{17}$$

As in the previous case, the numerator can be solved analytically under the assumption of the normal prior, whereas, the denominator is estimated from the posterior samples as described earlier.

3 Results

In our previous study, we demonstrated the identifiability of the experimental parameters of the statistical model of Lux from spike-in controls (Äijö *et al.*, 2016). Hence, we focus on validating the novel part of the model here, that is, the simplex model component implemented with the matrix normal distribution together with the GLM and softmax transformation. Briefly, we will confirm LuxGLM's ability to estimate methylation levels, detect differential methylation and incorporate covariates into the analysis. Additionally, we will compare LuxGLM with MLML (Qu *et al.*, 2013) and Lux (Äijö *et al.*, 2016) in the context of oxi-mC quantification and with RADMeth (Sun *et al.*, 2014) and MACAU (Lea *et al.*, 2015) in the context of the standard BS-seq data analysis. Finally, we will analyse novel and published targeted BS-seq and oxBS-seq data on T cells. Although in this study we focus on BS-seq and oxBS-seq data, the integrative analysis described here can be easily generalised for other oxi-mC data types (Äijö *et al.*, 2016).

3.1 LuxGLM accurately estimates methylation levels

To demonstrate that the proposed covariate-aware LuxGLM model can identify methylation levels, we carried out a simulation experiment for comparing the performances of LuxGLM, Lux and MLML in estimating methylation levels from BS-seq and oxBS-seq data with different sequencing depths. We consider three cases commonly observed in experiments: (i) a hypomethylated cytosine (Supplementary Fig. S2; top row), (ii) a hypermethylated cytosine (Supplementary Fig. S2; middle row) and (iii) an actively demethylated cytosine (Supplementary Fig. S2; bottom row). Additionally, we assume realistic values for the experimental parameters in the simulation of the data (Supplementary Fig. S2). The box plots of the LuxGLM, Lux and MLML estimates on data with different sequencing depths are shown in Supplementary Figure S2. In 48% (26/54) of the considered cases LuxGLM produced the most accurate estimate (the median of the estimates is closest to the true value), whereas in 37% (20/54) and 15% (8/54) of the cases Lux and MLML produced the most accurate estimate, respectively. The GLM part of the LuxGLM model together with its priors appear to make the LuxGLM more sensitive for methylation level estimation than the original Lux model. Importantly, the LuxGLM and Lux estimates lack the small biases of MLML estimates due to the explicit modeling of imperfect experimental parameters. LuxGLM and Lux also generally result in smaller variance in the methylation level estimates than MLML. Collectively, the results show that LuxGLM achieves similar or better performance on methylation level estimation than Lux and MLML when the experimental data is not confounded with covariates.

3.2 Detecting differential methylation from BS-seq data

We next validate the described Savage-Dickey approach described in Equation (16) and compare it with state-of-the-art methods, RADMeth and MACAU, for detecting differential methylation. To compare LuxGLM with RADMeth and MACAU in a fair manner, we considered only BS-seq data and assumed perfect experimental steps ($\mathrm{BS}_{\mathrm{eff}} = 1$, $\mathrm{BS}^*_{\mathrm{eff}} = 0$ and $\mathrm{seq}_{\mathrm{err}} = 0$). To include co-varying effects, we generated synthetic data in two batches ('pure' and 'garbled') from differentially methylated and similarly methylated cytosines between two conditions [see Supplementary Equations (S6–S8) and Supplementary Fig. S3]. Specifically, we generated data sets of 400 differentially and 400 similarly methylated cytosines with different numbers of replicates and sequencing depths.

The generated data sets were analysed with LuxGLM, RADMeth, and MACAU using the true design matrices followed by ranking the cytosines according to the BFs (LuxGLM) and p-values (RADMeth, MACAU) and deriving the receiver operating characteristics curves. Overall, LuxGLM, RADMeth and MACAU perform similarly in distinguishing differential methylation; however, LuxGLM produced slightly greater area under receiver operating characteristics (AUROC) curve values in most of the cases (Supplementary Table S1). As expected, the performances of all the three methods improve commensurate with the number of replicates and sequencing depth (Supplementary Table S1). Notably, when there are only few replicates available, detection performances can be improved by increasing the sequencing depth. In general, however, it is more beneficial in terms of detecting differential methylation to increase the number of replicates than sequencing depth. The similar performance of the methods in this settings is not surprising due to the high similarity of the statistical models of RADMeth and MACAU and the reduced version of LuxGLM (only BS-seq data and perfect experimental parameters). The only major difference between the approaches is the inference; for instance, RADMeth uses maximum likelihood principle, whereas LuxGLM uses Bayesian reasoning. However, the full version of LuxGLM has two important advancements over RADMeth and MACAU: (i) LuxGLM generalises for the integrative and simultaneous analysis of multiple oxi-mC data types and (ii) LuxGLM supports a model-based analysis of experimental parameters.

Next, we considered a scenario motivated by clinical studies in which subjects are monitored over time using BS-seq. To model trajectories of two conditions, we incorporate two additional covariates to the previous model [see Supplementary Equations (S9 and S10)]. We generate synthetic data sets composed of differentially and similarly methylated cytosines with different number of replicates and sequencing depths using the aforementioned model. Then, we studied how well LuxGLM, RADMeth, and MACAU detect differentially methylated cytosines from similarly methylated cytosines. As in the previous example, we assume that the design matrices are

Table 1. The AUROC values obtained using LuxGLM, RADMeth and MACAU on the data generated using the model in Supplementary Equation (S9) are listed

	Number of replicates								
	6			10			20		
Number of reads	LuxGLM	RADMeth	MACAU	LuxGLM	RADMeth	MACAU	LuxGLM	RADMeth	MACAU
6	**0.674**	0.621	0.654	**0.843**	0.746	0.818	**0.976**	0.900	0.967
12	**0.744**	0.633	0.713	**0.884**	0.772	0.878	**0.985**	0.913	**0.985**
24	**0.760**	0.642	0.722	**0.900**	0.774	0.890	**0.993**	0.927	**0.993**

The cases of 6, 12 and 24 reads are considered. Moreover, the cases of 6 (3 'pure' and 3 'garbled'), 10 (5 'pure' and 5 'garbled') or 20 (10 'pure' and 10 'garbled') replicates are considered. The values are calculated from 2000 simulated cytosines (1000 differentially methylated and 1000 similarly methylated cytosines) in each case. The greater AUROC for each case is in boldface.

known. Since RADMeth supports only binary covariates, we discretised the time covariates in the RADMeth analyses using the following criterion

$$d(t_{r_i}) = \begin{cases} 0, & \text{if } t_{r_i} \leq 0.5 \\ 1, & \text{otherwise.} \end{cases} \quad (18)$$

To quantify the performances of the methods, we calculated the receiver operating characteristics statistics as above (Table 1). Overall, the obtained AUROC values are lower in this case due to the additional confounding covariates. Additionally, LuxGLM and MACAU produce measurably better AUROC values than RADMeth in all nine cases. Presumably, this is due to two reasons: (i) LuxGLM and MACAU supports continuous covariates and (ii) the uncertainty introduced by additional covariates is better accounted for with the fully probabilistic approach.

3.3 Importance of including covariates in methylation analysis

Next, we demonstrate the importance of covariate modeling in estimating methylation levels and in detecting differential methylation for oxi-mC species. To do this, we consider similar scenarios as above: (i) a batch effect (termed as 'garbled') concealing a true difference between the 'pure' conditions (Fig. 2A) and (ii) a batch effect ('garbled') causing a false difference (Fig. 3A). First, we generate synthetic data with 20 replicates (10 "pure" and 10 "garbled" samples) per condition using the model [see Supplementary Equations (S12–S14)]. Then, we analyse the generated data with LuxGLM and MLML; the covariate structure is used in the LuxGLM analysis but ignored in the MLML analysis as it does not support covariates. To see how well our covariate model extracts information from the garbled samples, we visualise the posterior samples of θ corresponding to the 'pure' samples (Fig. 2B and Fig. 3B). For comparison, we include the MLML covariate-ignorant point estimates to the ternary plots. The results show that LuxGLM is able to integrate information from 'pure' and 'garbled' samples through the covariate model. Moreover, the comparison with MLML estimates shows that our covariate model produces more accurate estimates in both of the considered scenarios. As a matter of fact, LuxGLM (and Lux) also fails to estimate methylation accurately when the covariates are not considered (data not shown).

To quantify the benefits of covariate modeling, we compared the results obtained with ('full' model) or without ('reduced' model) the batch covariates ('pure'/'garbled'). As expected, the differential methylation detection is improved (BF is greater) when the covariate information is included in the analysis (Fig. 2C); In 78, 84 and 88% of the cases with 6, 10 and 20 replicates the BF is greater, respectively. Similarly, the non-differential methylation detection is improved when the covariate information is included in the analysis (Fig. 3C); In 49, 55 and 84% of the cases with 6, 10, 20 replicates the BF is decreased, respectively. The demonstrated advantage of covariate modeling is important because it enables to utilise replicates more efficiently across covariates.

Finally, we checked the effects of the increased (Fig. 2C) and decreased (Fig. 3C) BFs in discriminating differential methylation. To do this, we first pooled the differentially and similarly methylated cytosines, and then, we analysed the cytosines for differential methylation. Next, we derived the receiver operating characteristics curves to quantify the performances of the 'full' and 'reduced'

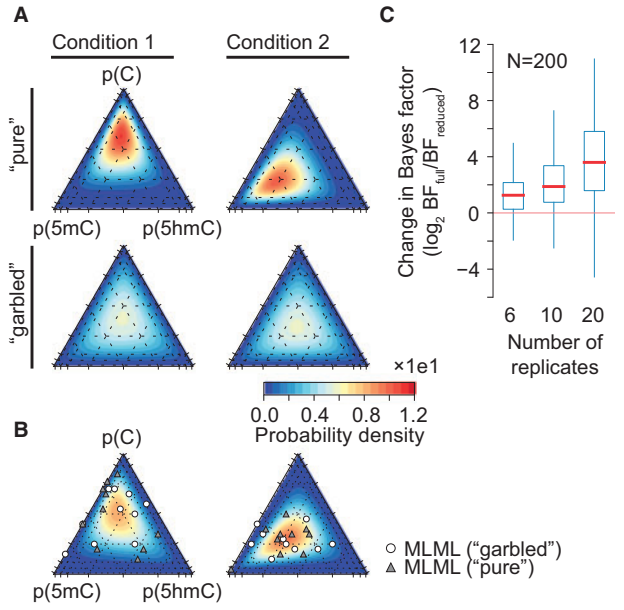

Fig. 2. (**A**) Ternary plot representations of the two considered different conditions (columns) and the corresponding two batches ('pure' on top row; 'garbled' on bottom row). (**B**) The ternary plot shows the condition specific posterior distributions obtained using LuxGLM. The samples of θ corresponding to the 'pure' samples are used. The estimates of condition 1 and 2 are on left and right, respectively. The white dots and gray triangles are the MLML estimates for the 'garbled' and 'pure' samples, respectively. The analysis is done with 20 (10 'pure' and 10 'garbled') replicates per condition. (**C**) The BFs obtained using the full or reduced model are compared. The full model has covariates for the condition and batch, whereas the reduced model has only a covariate for the condition. The data in the box plots are the changes of the BFs ($\log_2$). The analysis is done either with 6 (3 'pure' and 3 'garbled'), 10 (5 'pure' and 5 'garbled') or 20 (10 'pure' and 10 'garbled') replicates per condition. The box plots are derived from 200 random simulations

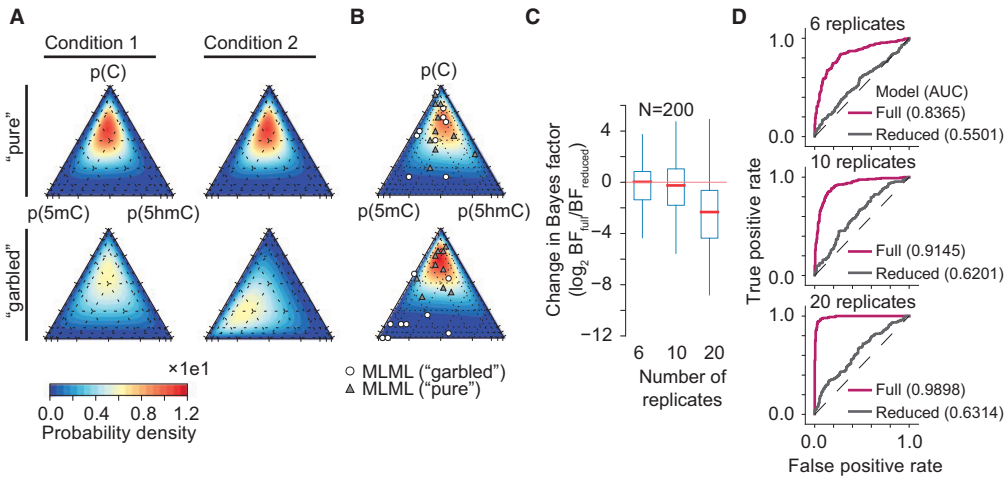

Fig. 3. (**A**) Ternary plot representations of the two considered similar conditions (columns) and the corresponding two batches ('pure' on top row; 'garbled'on bottom row). (**B**) The ternary plot shows the condition specific posterior distributions obtained using LuxGLM. The samples of θ corresponding to the 'pure' samples are used. The estimates of condition 1 and 2 are on top and bottom row, respectively. The white dots and gray triangles are the MLML estimates for the 'garbled' and 'pure' samples, respectively. The analysis is done with 20 (10 'pure' and 10 'garbled') replicates per condition. (**C**) The BFs obtained using the full or reduced model are compared. The full model has covariates for the condition and batch, whereas the reduced model has only a covariate for the condition. The data in the box plots are the changes of the BFs ($\log_2$). The analysis is done either with 6 (3 'pure' and 3 'garbled'), 10 (5 'pure' and 5 'garbled') or 20 (10 'pure' and 10 'garbled') replicates per condition. The box plots are derived from 200 random simulations. (**D**) A receiver operating characteristics analysis of discriminative abilities of the full and reduced models. Differentially ($N=200$) and similarly methylated ($N=200$) cytosines are generated as in Figures 2A and 3A, respectively. The cases of 6 (3 'pure' and 3 'garbled'), 10 (5 'pure' and 5 'garbled') and 20 (10 'pure' and 10 'garbled') replicates are considered. The cytosines are ordered based on the BFs and the receiver operating characteristics curves are derived. The areas under the curves are listed in the parentheses

models in detecting differential methylation (Fig. 3D). Indeed, the 'full' model provided measurably better performance then the 'reduced' model with 6, 10, 20 replicates; for instance, with 10 replicates AUROC values were 0.915 and 0.620 with the 'full' and 'reduced' models, respectively. Therefore, the covariate model of LuxGLM has practical importance as it improves the detection of methylation levels and differential methylation.

3.4 Quantifying the effects of time and vitamin C on demethylation of the *Foxp3* CNS1 locus

To investigate the applicability of LuxGLM on real biological data, we analysed a subset of a recently published longitudinal data set and novel data containing targeted BS-seq and oxBS-seq data (Yue *et al.*, 2016). The data were measured from induced T regulatory cells (iTregs) [generated *in vitro* from naïve $CD4^+$ T cells with TGF-β (Chen *et al.*, 2003)] under different conditions and the sequencing was targeted to loci within the *Foxp3* gene (Yue *et al.*, 2016). iTregs (*in vivo* generated) develop outside of the thymus from naïve $CD4^+$ T cells and they have an important role in immune tolerance by suppressing T cell proliferation and autoimmune diseases (Sakaguchi *et al.*, 2008). The *Foxp3* protein-coding gene has been shown to be essential for the development and function of regulatory T cells (Ramsdell and Ziegler, 2014). Among other mechanisms, the function of *Foxp3* has been reported to be regulated through three conserved non-coding sequence (CNS1, CNS2 and CNS3) loci in the *Foxp3* gene (Ramsdell and Ziegler, 2014). For instance, regulatory T cell lineage stability is regulated by DNA methylation at the *Foxp3* CNS2 locus (Zheng *et al.*, 2010).

Here we analysed the BS-seq and oxi-mC-seq data collected during *in vitro* iTreg differentiation with and without vitamin C (VitC) at multiple time points and with varying number of replicates. The sequencing libraries had spike-in control C, 5mC and 5hmC cytosines allowing to confirm that the experiments were successful. Additionally, these spike-in controls enabled us to estimate the values of the experimental parameters; the values of the experimental parameters are in the expected ranges; BS_{eff} varies between 0.988 and 0.999, BS^*_{eff} between 0.020 and 0.023, ox_{eff} between 0.733 and 0.937 and seq_{err} between 0.001 and 0.002 (Supplementary Table S2). Presumably, the observed variation in ox_{eff} has an observable effect on the oxBS-seq read-outs (incomplete oxidation increases 'C' read-outs) emphasizing the importance of including the experimental parameters to the analysis.

Next, we aimed to quantify the effects of the presence of VitC and time on demethylation of the CpG nucleotides (assumed to be independent from each other) in the *Foxp3* CNS1 locus. To do this, we used the presence/absence of VitC and time as covariates [see Supplementary Equation (S15)]. Note that the 'Basal/TGF-β' term also takes into account, in addition to TGF-β, the methylation state in the naïve $CD4^+$ T cells. To quantify the effects of VitC and time, we studied the posterior distributions of **B** for all the four CpG cytosines within the *Foxp3* CNS1 locus (Fig. 4A, Supplementary Fig. S4). The posterior distributions and the calculated (BF>10) for the individual covariates suggest that the cytosines chrX:7159186, chrX:7159222 and chrX:7159235 are methylated in the TGF-β stimulated cells at early time points. Moreover, our analysis suggests (BF>10) a vital role for VitC in demethylation of these cytosines; the estimated parameters slightly increase the proportion of C, decrease the proportion of 5mC, and in three cases of four slightly increase the proportion of 5hmC. Finally, our model proposes a significant role (BF>10) for the time in demethylation of all the four considered cytosines (proportion of C and 5hmC increases and decreases, respectively).

We investigated more closely the methylation status of the cytosine chrX:7159069 and tested whether LuxGLM can be used to predict methylation proportions for unseen experimental conditions. First, we represented the considered conditions (i.e. 16, 24, 32, 40, 48, 56, 64, 72 h with TGF-β+VitC) using our GLM [see Supplementary Equation (S16) for the model]. We estimated the posterior model parameters from BS-seq and oxBS-seq data

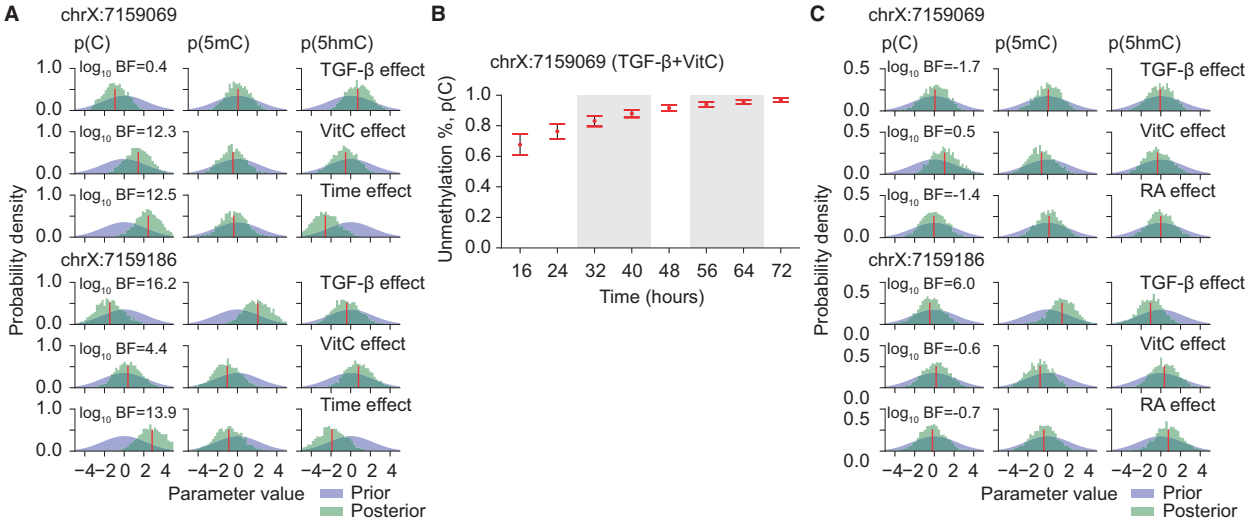

Fig. 4. (**A**) The posterior distributions of the parameter matrix **B** defined in Supplementary Equation (S15) of two CpG cytosines within the *Foxp3* CNS1 locus. The prior and posterior distributions are shaded in blue and green, respectively. The red lines depict the posterior means. The $\log_{10}$ transformed BFs of individual covariates are listed. (**B**) Predicted proportions of unmodified Cnm of the cytosine chrX:7159069 in the *Foxp3* CNS1 locus at different time points after the TGF-β and VitC stimuli. The posterior model parameters are estimated from BS-seq and oxBS-seq data at time points 16, 2, 48, 72 h and the predicted levels of unmethylated Cs at the time points 32, 40, 56, 64, h (shaded rectangles) are obtained using the posterior parameter samples of B. The means with the sSDs are depicted. (**C**) The posterior distributions of the parameter matrix **B** defined in Supplementary Equation (S18) of two CpG cytosines within the *Foxp3* CNS1 locus. The prior and posterior distributions are shaded in blue and green, respectively. The red lines depict the posterior means. The $\log_{10}$ transformed BFs are listed

measured at time points 16, 24, 48, 72 h and then produced the model output in all the conditions using the posterior samples of **B** (Fig. 4B). The model output illustrates the aforedescribed role of VitC and time in demethylation of chrX:7159069. Notably, our covariate model allows us to predict methylation levels at time points without measurements (32, 40, 56, 64h).

Additionally, we used the model, similarly as above, to predict the methylation landscapes of chrX:7159069 at 0 h with TGF-β stimulus and at 144 h with the TGF-β and VitC stimuli [Table 2; see Supplementary Equation (S17) for the model]. To validate these predictions, we used independent BS-seq data from (Yue *et al.*, 2016) (Table 2); briefly, Yue *et al.* measured (only BS-seq) three naïve $CD4^+$ T cell populations and four cell populations stimulated with TGF-β and VitC at 144 h. Although the predictions do not fully match the experimental values, our predictions are still notably accurate. It is important to remember that the assumed model [Supplementary Equation (S15)] is a simplification of the biological phenomenon; for instance, the assumed linear effect of time might be unrealistic because the effect of time on demethylation will likely dampen over time. Additionally, the prediction and experiment conditions at 0 h do not fully match; training data did not have any conditions without TGF-β, thus our prediction technically corresponds to naïve T cells with TGF-β ('Basal/TGF-β'). Nevertheless, our predictions clearly capture the biological phenomenon correctly and thus demonstrate that LuxGLM can also be used for predicting DNA methylation modifications.

Table 2. The predicted and experimentally quantified proportion of unmodified C's of the cytosine chrX:7159069 in the *Foxp3* CNS1 locus) and four (VitC 144 h) experiments

Time point			
0 h (naïve $CD4^+$)		144 h (VitC)	
Prediction	Experimental	Prediction	Experimental
0.117 ± 0.045	0.142 ± 0.021	0.997 ± 0.004	0.971 ± 0.023

The listed predictions are the posterior means with the SDs. The means and SDs of the experimental data are calculated from three (naïve $CD4^+$) and four (VitC 144 h) experiments.

3.5 Quantifying the effects of vitamin C and retinoic acid on demethylation of the *Foxp3* CNS1 locus

To study LuxGLM further on real biological data, we investigated the effects of VitC and retinoic acid (RA) on demethylation of the *Foxp3* CNS1 locus at 24 h (Yue *et al.*, 2016). Interestingly, RA has been reported to promote iTreg differentiation and suppress Th17 differentiation (Raverdeau and Mills, 2014). Similarly to above, we formulated a covariate model to study the effects of VitC and RA [the model is defined in Supplementary Equation (S18)]. Notably, the total number of samples and the number of 'RA' samples are rather small in this case [see the design matrix **D** in Supplementary Equation (S18)].

The oxidation efficiencies of the samples Sample nos. 7 and 8 were 0.888 and 0.814, respectively. To quantify the VitC and RA effects, we investigated the distributions of **B** across the CpG cytosines of the *Foxp3* CNS1 locus (Fig. 4C, Supplementary Fig. S5). The VitC and RA effects were subtle and nearly nonexistent on the *Foxp3* CNS1 methylation, respectively; VitC had minor effects (BF>1) on the cytosines chrX:7159069 and chrX:7159222 and RA did not have significant effect on any of the four cytosines. Presumably, the observed minor effect of VitC is due to the early time point and small sample size. Finally, the observed nonexistent effect of RA is supported by literature as RA has been reported to regulate *Foxp3* through histone modifications instead of DNA methylation (Lu *et al.*, 2011).

4 Discussion and conclusions

Here we described the first method which allows integrative analysis of different oxi-mC data sets with experimental parameters and arbitrary, complex experimental designs. The method is applicable in

analysis of genome-wide, reduced representation and targeted bisulphite sequencing data. Comparisons to existing methods demonstrated that LuxGLM has similar or better differential methylation detection performance than existing tools on BS-seq data. Analysis of simulated data showed that LuxGLM can provide accurate estimates of DNA methylation modifications even when confounding factors are present. Moreover, for oxi-mC species measured using complex experimental designs, LuxGLM is superior in differential methylation analysis when compared with existing methods.

Recent studies have indicated multiple functions for oxi-mC species, including intermediates in active DNA demethylation pathway as well as epigenetic marks that recruit chromatin regulators and interact with RNA polymerase. Taken together with findings of 5mC associations to several diseases, such as different cancers, interest in oxi-mCs in clinical setting is likely to emerge in near future. Clinical samples are commonly obtained using complex experimental designs and our method will allow better utilisation such data sets by enabling the control of confounding effects through covariates. Additionally, our covariate model introduces a flexible modeling framework to study and pinpoint the effects of different factors on DNA methylation modifications.

Acknowledgements

We wish to thank Maia Malonzo and Henrik Mannerström for careful reading of the paper. We acknowledge the computational resources provided by the Aalto Science-IT project.

Funding

This work was supported by Simons Foundation (to T.Ä.), Center for Computational Biology, the Academy of Finland Centre of Excellence in Molecular Systems Immunology and Physiology Research (to T.Ä. and H.L.), National Institutes of Health Research Project R01 grants AI44432 and CA151535 (to A.R.) and a Translational Research Program Award from the Leukemia and Lymphoma Society (LLS TRP 6464-15) (to A.R.).

Conflict of Interest: none declared.

References

Äijö,T. *et al.* (2016) A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways. *Genome Biol.*, **17**, 1–22.

Baylin,S.B. (2005) DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.*, **2**(**Suppl 1**), S4–11.

Booth,M.J. *et al.* (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, **336**, 934–937.

Booth,M.J. *et al.* (2014) Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat. Chem.*, **6**, 435–440.

Carpenter,B. *et al.* (in press). Stan: A probabilistic programming language. *J. Stat. Softw.*

Chen,W. *et al.* (2003) Conversion of peripheral CD4+CD25- naive T cells to CD4+CD25+ regulatory T cells by TGF-beta induction of transcription factor Foxp3. *J. Exp. Med.*, **198**, 1875–1886.

Dayeh,T. *et al.* (2014) Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion. *PLoS Genet.*, **10**, e1004160.

De Jager,P.L. *et al.* (2014) Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat. Neurosci.*, **17**, 1156–1163.

Dolzhenko,E. and 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.

Frommer,M. *et al.* (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA*, **89**, 1827–1831.

Gupta,A.K. and Nagar,D.K. (1999) *Matrix Variate Distributions, Volume 104*. CRC Press, Boca Raton.

He,Y.F. *et al.* (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, **333**, 1303–1307.

Huang,Y. *et al.* (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One*, **5**, e8888.

Ito,S. *et al.* (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, **333**, 1300–1303.

Lea,A.J. *et al.* (2015) A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genet.*, **11**, e1005650.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Lister,R. *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905.

Lu,L. *et al.* (2011) All-trans retinoic acid promotes TGF-β-induced tregs via histone modification but not DNA demethylation on Foxp3 gene locus. *PloS One*, **6**, e24590.

Lu,X. *et al.* (2013) Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J. Am. Chem. Soc.*, **135**, 9315–9317.

Nardone,S. *et al.* (2014) DNA methylation analysis of the autistic brain reveals multiple dysregulated biological pathways. *Transl. Psychiatry*, **4**, e433.

Pastor,W.A. *et al.* (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell. Biol.*, **14**, 341–356.

Plongthongkum,N. *et al.* (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat. Rev. Genet.*, **15**, 647–661.

Qu,J. *et al.* (2013) MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics*, **29**, 2645–2646.

Ramsdell,F. and Ziegler,S.F. (2014) FOXP3 and scurfy: how it all began. *Nat. Rev. Immunol.*, **14**, 343–349.

Rastogi,D. *et al.* (2013) Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. *Sci. Rep.*, **3**, 2164.

Raverdeau,M. and Mills,K.H.G. (2014) Modulation of T cell and innate immune responses by retinoic acid. *J. Immunol.*, **192**, 2953–2958.

Rein,T. *et al.* (1998) Identifying 5-methylcytosine and related modifications in DNA genomes. *Nucleic Acids Res.*, **26**, 2255–2264.

Sakaguchi,S. *et al.* (2008) Regulatory T cells and immune tolerance. *Cell*, **133**, 775–787.

Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.

Sun,D. *et al.* (2014) MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.

Tahiliani,M. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.

Wu,H. *et al.* (2014) Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.*, **32**, 1231–1240.

Yu,M. *et al.* (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, **149**, 1368–1380.

Yue,X. *et al.* (2016) Control of Foxp3 stability through modulation of TET activity. *J. Exp. Med.*, **213**, 377–397.

Zheng,Y. *et al.* (2010) Role of conserved non-coding DNA elements in the Foxp3 gene in regulatory T-cell fate. *Nature*, **463**, 808–812.