Halla-aho, Viivi; Lähdesmäki, Harri

Efficient statistical methods for detecting differential methylation

Published: 22/07/2017

Please cite the original version:
Halla-aho, V., & Lähdesmäki, H. (2017). *Efficient statistical methods for detecting differential methylation*. Paper presented at International Conference on Intelligent Systems for Molecular Biology , Prague, Czech Republic.

Efficient statistical methods for detecting differential methylation

Viivi Halla-aho and Harri Lähdesmäki

Aalto University School of Science
Department of Computer Science

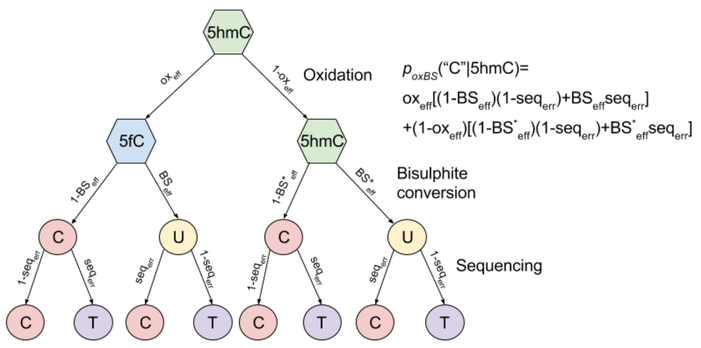
Aalto University
School of Science

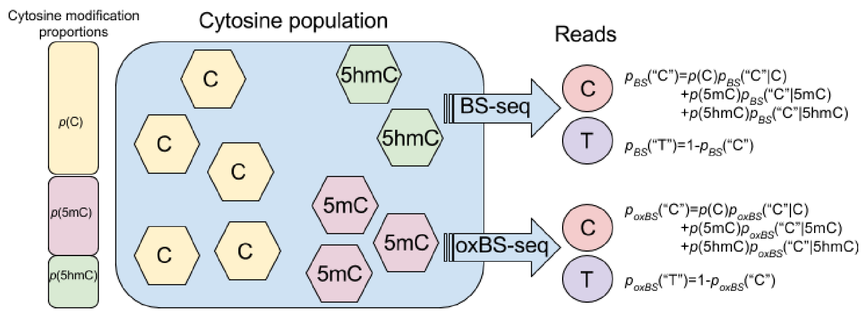
Background and motivation

- Addition of the methyl group to the 5-position of a cytosine (5mC) is the most commonly studied epigenetic modification on DNA, and its effects on different diseases and cancer have been widely studied.
- We have previously developed a hierarchical generative model, LuxGLM [1], for analysing 5mC and oxidized methylcytosine species (oxi-mC).
- LuxGLM can take into account the different experimental parameters and confounding factors along with complex experimental design.
- To enhance the computational efficiency we propose the usage of variational inference (VI) instead of Hamiltonian Monte Carlo (HMC) sampling. VI is typically faster than MCMC sampling methods.

LuxGLM

- **Read-out probabilities for single cytosine and for a population**





The general linear model is used for calculating $\theta_i = \big(p(\text{C}), p(\text{5mC}), p(\text{5hmC})\big)$ for each sample $i = 1, ..., N$.

- **General linear model**
  The linear part of the model with P covariates has the following form

$$\mathbf{Y} = \mathbf{DB} + \mathbf{E}, \tag{1}$$

  where $\mathbf{Y} \in R^{N\times M}$ gives the parameters θ_i through Softmax transformation $\theta_i = \text{Softmax}(\text{row}_i(\mathbf{Y}))$, $\mathbf{D} \in R^{N\times P}$ is the design matrix, $\mathbf{B} \in R^{P\times M}$ is the parameter matrix and $\mathbf{E} \in R^{N\times M}$ represents normally distributed, zero-centered noise term.

- **Bayes factors**
  To asses the difference in methylation between two conditions i and j the null hypothesis (no differential methylation) is

$$H_0 : \text{row}_i(\mathbf{B}) - \text{row}_j(\mathbf{B}) \equiv C_1 - C_2 = \mathbf{0}, \tag{2}$$

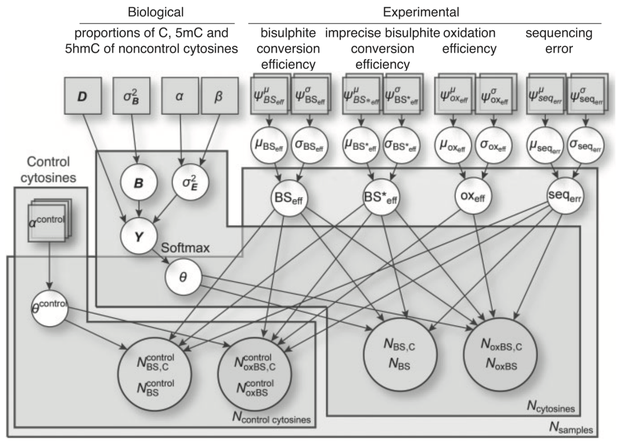
  and alternative hypothesis (differential methylation) is

$$H_1 : \text{row}_i(\mathbf{B}) - \text{row}_j(\mathbf{B}) \equiv C_1 - C_2 \neq \mathbf{0}. \tag{3}$$

  The Savage-Dickey density ratio approximates the Bayes factor between the models representing these hypotheses

$$BF \approx \frac{p(C_1 - C_2 = \mathbf{0}|H_1)}{p(C_1 - C_2 = \mathbf{0}|H_1, \mathcal{D})}. \tag{4}$$

- **Model hierarchy**
  HMC sampling from the posterior is done with Stan.



Variational inference for computation of the Bayes factors

- Variational inference approximates the posterior with a simpler distribution and to find the optimal approximative distribution, the expectation lower bound (ELBO) is maximized, which corresponds to mimizing the Kullback-Leibler distance.
- In the probabilistic programming language Stan, Automatic Differentiation Variational Inference (ADVI) algorithm has been implemented [2] and so the HMC sampling used by default in Stan can be easily switched to VI. ADVI algorithm parameters which can be tuned are number of gradient samples N_G and number of ELBO samples N_E.
- The ELBO values for the approximations can be used to calculate another BF approximation
  BF$\approx \exp(\text{ELBO}_{H_1} - \text{ELBO}_{H_0})$.

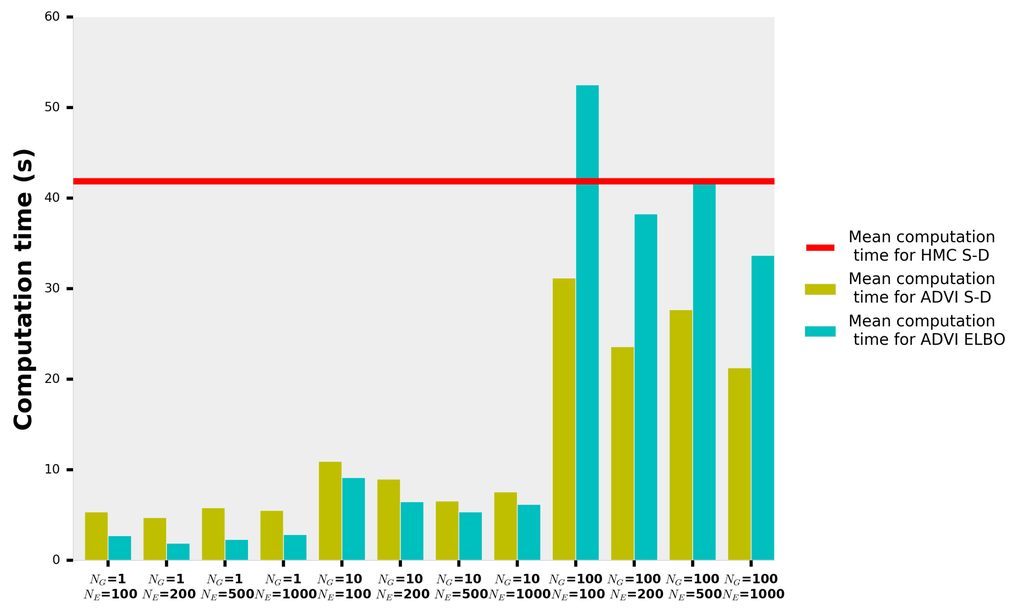
Comparison of LuxGLM and state-of-the-art methods

Comparison table of LuxGLM, RADMeth [3] and MACAU [4] from [1]. In the comparison the area under receiver operating charasteristic curve (AUROC) was calculated using simulated data sets. Perfect experimental steps and only BS-seq data were considered in the simulations, as experimental parameters and oxi-mC are not supported by the other methods.

	Number of replicates								
	6			10			20		
Number of reads	LuxGLM	RADMeth	MACAU	LuxGLM	RADMeth	MACAU	LuxGLM	RADMeth	MACAU
6	**0.674**	0.642	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**

Computation times for variational inference and comparison with HMC

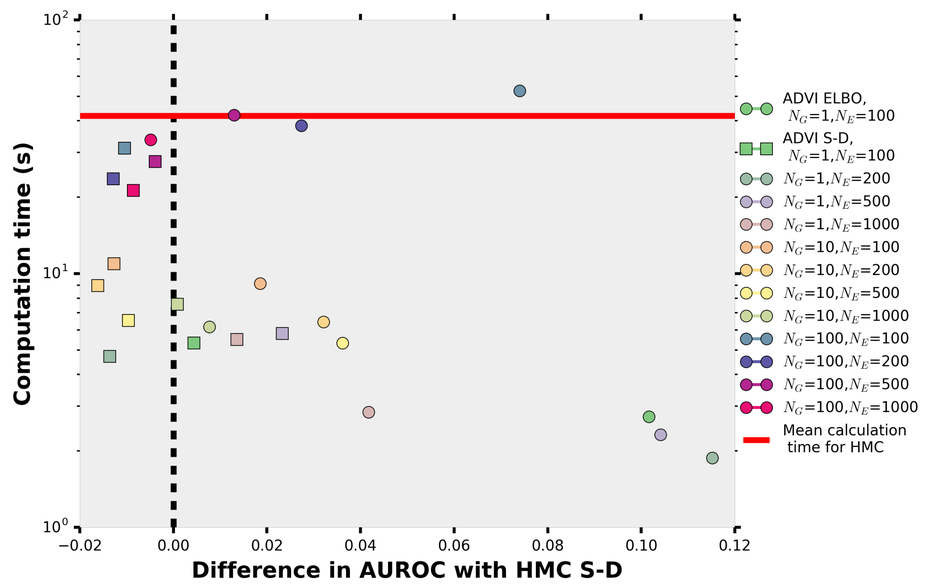
- Computation times using Stan's variational inference feature with different parameter values to compute the Savage-Dickey and ELBO approximations of the Bayes factor. The number of reads was 12 and number of replicates was 10.



- Comparison table of the AUROC values and mean computation times in seconds of the original Savage-Dickey estimate and Savage-Dickey and ELBO estimates calculated using variational inference for simulated data. The algorithm parameters were $N_G = 10$ and $N_E = 1000$ for ADVI.

	Number of replicates																	
	6						10						20					
Number	HMC S-D		ADVI S-D		ADVI ELBO		HMC S-D		ADVI S-D		ADVI ELBO		HMC S-D		ADVI S-D		ADVI ELBO	
of reads	AUROC	Time	AUROC	Time	AUROC	Time	AUROC	Time	AUROC	Time	AUROC	Time	AUROC	Time	AUROC	Time	AUROC	Time
6	**0.655**	16.98	0.607	5.93	0.595	**3.23**	0.811	36.87	**0.823**	7.54	0.778	**6.13**	0.962	130.61	0.957	**15.21**	**0.963**	21.33
12	0.765	19.10	**0.770**	5.94	0.698	**3.23**	**0.898**	42.54	0.897	7.56	**0.898**	**6.16**	**0.985**	151.14	0.978	**15.24**	**0.985**	21.27
24	0.750	23.30	**0.765**	5.92	0.699	**3.09**	0.905	52.18	**0.910**	7.52	0.901	**5.98**	**0.993**	179.56	0.986	**14.91**	0.992	20.90

- Scatterplot of the mean computation times and differences in AUROC with Savage-Dickey approximation calculated using HMC using different parameter values for ADVI. Number of reads was 12 and number of replicates was 10.



References

[1] Tarmo Äijö et al. LuxGLM: a probabilistic covariate model for quantification of DNA methylation modifications with complex experimental designs, *Bioinformatics*, 32 (17): i511-i519, 2016.

[2] Alp Kucukelbir et al. Automatic variational inference in Stan, *Advances in Neural Information Processing Systems 28*, 2015.

[3] Egor Dolzhenko and Andrew D. Smith. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments, *BMC Bioinformatics*, 15:215, 2014.

[4] Amanda J. Lea et al. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulphite sequencing data, *PLos Genetics*, 11.11 (2015): e1005650, 2015.