Bioinformatics Advance Access published August 27, 2016

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw527
Advance Access Publication Date: 13 August 2016
Applications Note

Genome analysis

oxBS-MLE: an efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA

Zongli Xu[1], Jack A. Taylor[1,2], Yuet-Kin Leung[3], Shuk-Mei Ho[3] and Liang Niu[3,*]

[1]Epidemiology Branch, [2]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA and [3]Department of Environmental Health, College of Medicine, University of Cincinnati, Cincinnati, OH 45267, USA

*To whom correspondence should be addressed

Associate Editor: Inanc Birol

Received on June 27, 2016; revised on August 4, 2016; accepted on August 8, 2016

Abstract

Motivation: 5-Methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) are important epigenetic regulators of gene expression. 5mC and 5hmC levels can be computationally inferred at single base resolution using sequencing or array data from paired DNA samples that have undergone bisulfite and oxidative bisulfite conversion. Current estimation methods have been shown to produce irregular estimates of 5hmC level or are extremely computation intensive.
Results: We developed an efficient method oxBS-MLE based on binomial modeling of paired bisulfite and oxidative bisulfite data from sequencing or array analysis. Evaluation in several datasets showed that it outperformed alternative methods in estimate accuracy and computation speed.
Availability and Implementation: oxBS-MLE is implemented in Bioconductor package ENmix.
Contact: niulg@ucmail.uc.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

5-Methylcytosine (5mC) is an essential epigenetic mark that regulates cell differentiation and gene expression. 5mC can be oxidized to 5-hydroxymethylcytosine (5hmC) by activity of the ten-eleven translocation (TET) enzyme (Ito *et al.*, 2010; Tahiliani *et al.*, 2009). 5hmC can function differently from 5mC in transcriptional control and is prevalent in embryonic stems cells (Pastor *et al.*, 2011; Song *et al.*, 2011). Thus, it is desirable to separately quantify 5mC and 5hmC levels at single base resolution.

Traditional bisulfite (BS) treatment commonly used in genome-wide methylation studies cannot distinguish 5mC from 5hmC: both are protected from conversion while unmodified cytosines are converted to uracil (U). Selective oxidation of 5hmC to 5-formylcytosine (5fC) can be achieved by treatment with potassium perruthenate (KRuO4), and 5fC converts to U following bisulfite treatment (Booth *et al.*, 2012). 5hmC level can be estimated using paired experiments on a DNA sample: bisulfite converted DNA can be used to estimate the proportion of 5mC + 5hmC and oxidative bisulfite (oxBS) converted DNA can be used to estimate the proportion of 5mC. Measurement can be done in a genome-wide manner using sequencing techniques, e.g. oxBS-seq (Booth *et al.*, 2013); or for specific loci using array techniques, e.g. oxBS-450K (Stewart *et al.*, 2015) and OxBS-array (Field *et al.*, 2015).

5hmC level at each CpG site can be estimated by taking the difference between measurements made following BS and oxBS treatments (Booth *et al.*, 2013; Field *et al.*, 2015; Stewart *et al.*, 2015). However, this naïve approach frequently leads to negative estimates of 5hmC due to measurement error, while true levels must always be non-negative. To overcome the limitation, a statistical approach, OxyBS (Houseman *et al.*, 2016) was proposed to estimate 5hmC and 5mC levels by assuming beta distributions for the two measurements. A limitation of this approach is that it is too computationally intensive for large studies.

Here we propose a computationally efficient method oxBS-MLE (MLE stands for maximum likelihood estimate) to jointly estimate 5mC and 5hmC levels from the paired experimental data. Evaluations in real datasets show that it outperforms OxyBS and the naïve method. oxBS-MLE has been implemented in the Bioconductor package ENmix (Xu *et al.*, 2016).

2 Methods

For each CpG site following either BS or oxBS treatment, the resulting measurements are proportions. If the measurement is obtained by a sequencing technique, then the measurement is M/N, where M is the number of reads with unconverted (i.e. methylated) cytosine and N is the total number of reads containing the locus of interest. If the measurement is obtained by an array technique, e.g. Illumina Infinium HumanMethylation450 BeadChip or Infinium MethylationEPIC BeadChip, then the measurement is called 'beta value', and is calculated as M/(M + U+c), where M is the methylated intensity, U is the unmethylated intensity and c (usually = 100) is a constant offset to prevent denominator from being zero. Notice that the measurements obtained by array techniques will have the same form as those obtained by sequencing techniques if we let N = M + U+c.

Let M_k be the methylated signal and N_k be the total signal at a DNA locus, where $k \in \{BS, oxBS\}$. oxBS-MLE assumes that M_k follows a binomial distribution, i.e.,

$$M_{BS} \sim B(N_{BS}, \pi_{5mC} + \pi_{5hmC})$$

$$M_{oxBS} \sim B(N_{oxBS}, \pi_{5mC})$$

Here π_{5mC} (5mc level) and π_{5hmC} (5hmC level) are between 0 and 1 and $\pi_{5mC} + \pi_{5hmC} \leq 1$. It can be proved that the MLE of (π_{5mC}, π_{5hmC}) is (see Supplementary Material):

$$(\widehat{\pi}_{5mC}, \widehat{\pi}_{5hmC}) = \begin{cases} \left(\dfrac{M_{oxBS}}{N_{oxBS}}, \dfrac{M_{BS}}{N_{BS}} - \dfrac{M_{oxBS}}{N_{oxBS}}\right); & \text{if } \dfrac{M_{BS}}{N_{BS}} \geq \dfrac{M_{oxBS}}{N_{oxBS}} \\ \left(\dfrac{M_{oxBS} + M_{BS}}{N_{oxBS} + N_{BS}}, 0\right); & \textit{otherwise} \end{cases}$$

oxBS-MLE is implemented in ENmix as a function oxBS.MLE. It takes four matrices as input: a matrix of BS measurements (M_{BS}/N_{BS}), a matrix of oxBS measurements (M_{oxBS}/N_{oxBS}), a matrix of BS total signal (N_{BS}) and a matrix of oxBS total signal (N_{oxBS}). In each matrix, the rows are corresponding to CpG loci and the columns are corresponding to samples. oxBS.MLE outputs the $(\widehat{\pi}_{5mC}, \widehat{\pi}_{5hmC})$ defined as above.

3 Results

To evaluate oxBS-MLE and compare it with alternative methods, we applied the oxBS-MLE, OxyBS and the naïve subtraction method to the methylation 450k array data (GEO accession number GSE63179) from (Field *et al.*, 2015), which have undergone the correction for Infinium I and II probe design bias using Subset-quantile With Array Normalization (SWAN) by Field *et al.* (2015). The data consist of eight arrays for four replicates of a single cerebellum sample (two arrays, BS or oxBS treated, for each replicate). The qPCR measurements of 5hmC level for 27 CpG sites on the array were also available (Field *et al.*, 2015) and thus can be used as gold standard to evaluate the accuracy of the 5hmC estimates from array data with different methods.

Evaluations showed that methods oxBS-MLE and OxyBS provided almost identical 5hmC estimates with maximum absolute difference of 0.005 across all CpG sites and all four replicates, while the naïve method produced a significant portion of negative 5hmC estimates (12–38.4% in each replicate). To evaluate the accuracy of 5hmC estimates, we first averaged the four replicate estimates with each method at the 27 CpG sites and then calculated the absolute difference between the averages and the corresponding qPCR measurements. The results showed that oxBS-MLE and OxyBS estimates are almost identical and both outperformed the naïve method (Student paired *t*-test $P < 0.025$ for either comparisons).

Method oxBS-MLE and OxyBS also produced similar estimates of 5mC at majority of the CpG sites with only 0.039% of the estimates having absolute difference greater than 0.01 between the two methods. However, when the beta value from BS experiment is smaller than that from the corresponding oxBS experiment, oxBS-MLE and OxyBS tend to produce different 5mC estimates with maximum difference of 0.28 in the evaluation. Especially when BS measurement is very small, OxyBS method can result in irrational estimates of 0 no matter what the oxBS measurement is (Supplementary Table S1). We observed the similar phenomenon in other datasets (data not shown). We also noticed that for most CpG sites (71.4%, 286 752 out of 401 866) the oxBS-MLE estimates of 5mC are closer to the measurements from oxBS experiment (oxidative bisulfite experiments aim to measure 5mC level) than the OxyBS estimates; the percentage of such CpG sites in Infinium I probes (69.9%) is similar to that in Infinium II probes (72.0%).

We note that oxBS-MLE is ~5000 times faster than OxyBS when both methods were applied to the above GEO data (0.7 second versus 62 minutes for eight arrays on a machine with 2.6GHz CPU). This is because that OxyBS uses an iterative method to find MLE while oxBS-MLE calculates MLE using an explicit analytical form. The computation in the method OxyBS is done sample by sample, and thus the computing time will increase linearly with sample size. The computing time for 1000 samples (2000 arrays) is expected to be ~257 h for OxyBS while oxBS-MLE takes only <3 minutes. It is easy to see that the oxBS-MLE method assumption is even more suitable for sequencing data. However, we were not able to find oxBS sequencing data with additional qPCR/pyrosequencing validation. In summary, oxBS-MLE is an efficient and accurate method to jointly estimate 5mC and 5hmC levels in paired BS- and oxBS-treated experiments.

Funding

This research was supported by the Center for Environmental Genetics grant (NIEHS award P30ES00606), NIEHS grant (ES022071), and Intramural Research Program of the NIH (NIEHS Z01 ES049033 and Z01 ES049032).

Conflict of Interest: none declared.

References

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.* (2013) Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat. Protoc.*, **8**, 1841–1851.

Field,S.F. *et al*. (2015) Accurate measurement of 5-methylcytosine and 5-hydroxymethylcytosine in human cerebellum DNA by oxidative bisulfite on an array (OxBS-array). *PLoS One*, **10**, e0118202.

Houseman,E.A. *et al*. (2016) OxyBS: estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. *Bioinformatics*, **32**, 2505–2507.

Ito,S. *et al*. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, **466**, 1129–1133.

Pastor,W.A. *et al*. (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, **473**, 394–397.

Song,C.X. *et al*. (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol*., **29**, 68–72.

Stewart,S.K. *et al*. (2015) oxBS-450K: a method for analysing hydroxymethylation using 450K BeadChips. *Methods*, **72**, 9–15.

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

Xu,Z. *et al*. (2016) ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res*., **44**, e20.