Bioinformatics, 35(7), 2019, 1094–1097
doi: 10.1093/bioinformatics/bty778
Advance Access Publication Date: 1 September 2018
Original Paper

OXFORD

Genome analysis

Differential methylation values in differential methylation analysis

Changchun Xie[1,*], Yuet-Kin Leung[1], Aimin Chen[1], Ding-Xin Long[2], Catherine Hoyo[3] and Shuk-Mei Ho[1]

[1]Department of Environmental Health, University of Cincinnati, Cincinnati, OH 45267, USA, [2]School of Public Health, University of South China, Hengyang, Hunan 421001, China and [3]Department of Biological Sciences, North Carolina State University, Raleigh, NC 27695, USA

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on April 8, 2018; revised on August 10, 2018; editorial decision on August 29, 2018; accepted on August 31, 2018

Abstract

Motivation: Both β-value and M-value have been used as metrics to measure methylation levels. The M-value is more statistically valid for the differential analysis of methylation levels. However, the β-value is much more biologically interpretable and needs to be reported when M-value method is used for conducting differential methylation analysis. There is an urgent need to know how to interpret the degree of differential methylation from the M-value. In M-value linear regression model, differential methylation M-value ΔM can be easily obtained from the coefficient estimate, but it is not straightforward to get the differential methylation β-value, $\Delta\beta$ since it cannot be obtained from the coefficient alone.

Results: To fill the gap, we have built a bridge to connect the statistically sound M-value linear regression model and the biologically interpretable $\Delta\beta$. In this article, three methods were proposed to calculate differential methylation values, $\Delta\beta$ from M-value linear regression model and compared with the $\Delta\beta$ directly obtained from β-value linear regression model. We showed that under the condition that M-value linear regression model is correct, the method M-model-coef is the best among the four methods. M-model-M-mean method works very well too. If the coefficients $\alpha_0,\ \alpha_2, \ldots \alpha_p$ are not given (as 'MethLAB' package), the M-model-M-mean method should be used. The $\Delta\beta$ directly obtained from β-value linear regression model can give very biased results, especially when M-values are not in $(-2, 2)$ or β-values are not in (0.2, 0.8).

Availability and implementation: The dataset for example is available at the National Center for Biotechnology Information Gene Expression Omnibus repository, GSE104778.

Contact: xiecn@ucmail.uc.edu

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

1 Introduction

Changes in DNA methylation patterns play a critical role in the organ development, aging and diseases such as multiple sclerosis, diabetes, schizophrenia and cancer (Laird, 2010). Advances in the high-throughput assessment of DNA methylation have enabled quantitative profiling of DNA methylation of CpG loci throughout the genome, which is crucial to understand the role of epigenetics in regulating gene expression. The microarray-based Infinium HumanMethylation27 BeadChip, the Infinium HumanMethylation450 BeadChip and the newly developed MethylationEPIC BeadChip (Infinium) microarray (850k) (Moran *et al.*, 2016; Sandoval *et al.*, 2011; Thirlwell *et al.*, 2010) are widely used commercial platforms for low-cost high-throughput methylation profiling. Both β-value and M-value have been used as metrics to measure methylation levels. β-value is defined as:

$$\beta = \frac{\max(methylated,\ 0)}{(max(methylated,\ 0) + max(unmethylated,\ 0) + \alpha)},$$

where *methylated* and *unmethylated* are intensities measured by the methylated and unmethylated probes for an interrogated CpG site and a constant offset α (by default, $\alpha = 100$) is added to regularize β-value when both methylated and unmethylated probe intensities are low (Du *et al.*, 2010). The standardized fraction, i.e. the M-value is defined as

$$M = \log_2\left(\frac{\beta}{1-\beta}\right).$$

While the M-value is more statistically valid for the differential analysis of methylation levels because the M-value is approximately homoscedastic (the β-value has severe heteroscedasticity outside the middle methylation range, which imposes serious challenges in applying many statistic models) (Du *et al.*, 2010), the β-value is much more biologically interpretable, because it corresponds roughly to the percentage of a site that is methylated, which makes the β-value very attractive when modelling the underlying biological effect (Du *et al.*, 2010).

Saadati *et al.* (2014) examined parametric methods, such as linear and beta regression, and nonparametric methods, such as rank-based regression. They found that the use of β-values in a beta regression setting may be of benefit, but only if the underlying distribution of the β-values is indeed the beta distribution, which requires that the methylated and unmethylated signal intensities are independently gamma distributed with the same scale parameter. Beta regression model seems very susceptible to the violation of the beta distribution assumption and may show an uncontrolled false discovery rate. By allowing for possible correlations between the methylated and unmethylated signal intensities, Weinhold *et al.* (2016) proposed the Ratio of Correlated Gammas (RCG) model and showed the large benefit of RCG model when the correlation is high. However, when the correlation is low ($\rho = 0.2$), the Type I error exceeds the nominal level of significance, 0.05 in all of their simulations. Currently, the M-value linear regression model is one of the most popular models in the analysis of DNA methylation data. In this article, we focus on M-value linear regression model.

Du *et al.* (2010) compared β-value and M-value approaches and demonstrated that the relationship between the β-value and M-value methods is a Log-transformation.

$$M(\beta) = log_2\left(\frac{\beta}{1-\beta}\right),\ \ \beta(M) = \frac{2^M}{1+2^M}.$$

They showed that the β-value method has severe heteroscedasticity for highly methylated or unmethylated CpG sites and recommended using the M-value method for conducting differential methylation analysis and including the β-value statistics when reporting the results to investigators.

In M-value linear regression model (see the next section for details), differential methylation value, ΔM can be easily obtained from the coefficient estimate, however, it is not straightforward to get the differential methylation β-value, $\Delta\beta$ since it cannot be obtained from the coefficient alone (see the methods below). Due to this reason, some investigators usually run both M-value linear regression model and β-value linear regression model. They use the M-value linear regression model to select the CpG sites and report the p-values, but use β-value linear regression model to report differential methylation β-value, $\Delta\beta$. This can cause inconsistent results. First, the two models have different assumptions. Second, $\Delta\beta$ can be out of the [0, 1] interval (see below for details). In this article, we suggest different methods to calculate differential methylation values, $\Delta\beta$ from M-value linear regression model and compare it with the $\Delta\beta$ directly obtained from β-value linear regression model.

The outline of this article is as follows. The four different methods (three methods from the M-value linear regression model and one from β-value linear regression model) to obtain $\Delta\beta$ are presented in Section 2. In Section 3, simulations are conducted to compare the proposed methods. In Section 4, a simple method is proposed to quickly estimate $\Delta\beta$ when M-values are in $(-2, 2)$ or β-values are in $(0.2, 0.8)$. Examples are given in Section 5 to illustrate the proposed methods. Section 6 discusses the implications and provides concluding remarks.

2 Materials and methods

Considering the following linear regression model:

$$y_i = \alpha_0 + x_{i1}\alpha_1 + x_{i2}\alpha_2 + \cdots + x_{ip}\alpha_p + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i is the methylation value (M-value or β-value), x_{i1} is the variable of interest and $x_{i2}, \ldots, x_{ip}$ are the adjusting variables (confounders) for individual i. If M-value (β-value) is used in (1), the model is called M-value (β-value) linear regression model. The variable of interest, x_{i1} and the covariates can be categorical or continuous. This model has been implemented in R package 'MethLAB' although the coefficient estimates for the covariates are not available (only the coefficient estimate for the variable of interest is available, this is also the case for most common methylation packages).

2.1. M-value linear regression model

The M-value linear regression model is the model (1), where M-value is used. For this model, the differential methylation value, ΔM is the coefficient estimate, α_1 for 1 unit increase of the variable of interest, x_{i1}. Given the value for each covariate $x_{ij} = x_{0j}$ (e.g. x_{0j} can be chosen as the mean of $x_{ij},\ i = 1, \ldots, n$), $j = 1, \ldots, p$, $\Delta\beta$ can be obtained by

$$\begin{aligned}\Delta\beta &= \beta(M_0 + \Delta M) - \beta(M_0) \\ &= \frac{2^{(M_0+\Delta M)}}{1+2^{(M_0+\Delta M)}} - \frac{2^{M_0}}{1+2^{M_0}},\end{aligned} \tag{2}$$

where $M_0 = \alpha_0 + x_{01}\alpha_1 + x_{02}\alpha_2 + \cdots + x_{0p}\alpha_p$. This method will be called 'M-model-coef' method for the rest of the article.

If the coefficients $\alpha_0,\ \alpha_2, \ldots \alpha_p$ are not given (as 'MethLAB' package (Kilaru *et al.* 2012)), M_0 can be chosen as the mean of methylation M-value, which will be called 'M-model-M-mean' method. M_0 might also be chosen as

$$M_0 = M(\beta_0) = \log_2\left(\frac{\beta_0}{1-\beta_0}\right),$$

where β_0 is the mean of methylation β-value. This method will be called 'M-model-β-mean' method.

2.2. β-value linear regression model

β-value linear regression model is the model (1), where β-value is used. For this model, the differential methylation value, $\Delta\beta$ is the coefficient estimate, α_1 for 1 unit increase of the variable of interest, x_{i1}. This method will be called 'β-model-coef' method.

Note that the β-value has lower limit, 0 and upper limit, 1 but the right side of (1) does not have any limits. Due to this reason, the

Table 1. $\Delta\beta$=0.1

β_value: β_0	M-model-coef		M-model-M-mean		M-model- β-mean		β-model-coef	
	Bias	SD	Bias	SD	Bias	SD	Bias	SD
0.05	0.0002	0.0075	0.0002	0.0100	0.0640	0.0159	−0.0138	0.0094
0.5	−0.0002	0.0118	−0.0002	0.0118	−0.0002	0.0118	−0.0113	0.0103
0.85	−0.0001	0.0046	0.0001	0.0082	0.0396	0.0094	0.0581	0.0103

Table 2. $\Delta\beta$=0.2

β_value: β_0	M-model-coef		M-model-M-mean		M-model- β-mean		β-model-coef	
	Bias	SD	Bias	SD	Bias	SD	Bias	SD
0.05	0.0001	0.0117	0.0006	0.0216	0.1574	0.0222	−0.0532	0.0134
0.5	−0.0002	0.0106	−0.0003	0.0109	−0.0002	0.0107	−0.0297	0.0087
0.75	0.0001	0.0073	0.0005	0.0193	0.0627	0.0148	0.0577	0.0112

differential methylation β-value can be outside the limits if we consider many units increase of the variable of interest, x_{i1}.

3 Results

Simulations are conducted to compare the methods proposed in Section 2 under the condition that M-value linear regression model is correct. These simulations are not to show M-value linear regression model is better than β-value linear regression model (this work has been done by Du *et al.* (2010)). Linear regression model (1) with $p = 1$ was used to generate the methylation M-values with sample size = 200. In the first simulation, we assume β_0 (β_value) = 0.05, 0.5 and 0.85 and $\Delta\beta$ = 0.1. The covariate X was generated from normal distribution with mean = 2 and standard deviation = 1. α_1 and α_0 were calculated by

$$\alpha_1 = \Delta M = M_0 + \Delta M - M_0 = \log_2\left(\frac{\Delta\beta + \beta_0}{1 - \Delta\beta - \beta_0}\right) - \log_2\left(\frac{\beta_0}{1 - \beta_0}\right),$$

$$\alpha_0 = M_0 - mean(X)\alpha_1 = \log_2\left(\frac{\beta_0}{1 - \beta_0}\right) - 2\alpha_1.$$

In the second simulation, we assume $\beta_0 = 0.05$, 0.5 and 0.75 and $\Delta\beta$ = 0.2. After M-values were generated, the β_values were calculated by

$$\beta = \frac{2^M}{1 + 2^M}.$$

Four methods (M-model-coef, M-model-M-mean, M-model- β-mean and β-model-coef) were performed on the generated data to estimate $\Delta\beta$ separately. We repeated 10 000 times. The bias and standard deviation (SD) were summarized in Tables 1 and 2 below.

Based on the simulations (Tables 1 and 2), we can see the method M-model-coef is the best among the four methods. M-model-M-mean method works very well although it has a slightly larger bias and SD than M-model-coef method. It has much less bias than M-model- β-mean method (except β_value = 0.5, we will discuss this situation in the next section) and β-model-coef method. If the coefficients $\alpha_0, \alpha_2, \ldots \alpha_p$ are not given (as 'MethLAB' package, Kilaru *et al.*, 2012), M-model-M-mean should be used.

4 A simple method when M-values are in (−2, 2) or β-values are in (0.2, 0.8)

As shown in simulations above, M-model- β-mean method has a large bias, compared with M-model-coef method and M-model-M-mean method when β_value $\neq$ 0.5. However, when β_value = 0.5, M-model- β-mean method works very well. In fact, from Figure 1 above, we can see there is roughly linear relationship between M-value and β_value when M-values are in (−2, 2) or β-values are in (0.2, 0.8). Based on this approximately linear relationship, we can roughly estimate $\Delta\beta$ from the following simple formula:

$$\Delta\beta = \frac{0.6\Delta M}{4} = 0.15\Delta M. \quad (3)$$

However, most of the methylation sites have β-values outside of (0.2, 0.8), limiting the utility and applicability of this formula.

5 Example

In this section, we will use two real studies as examples to illustrate the methods introduced above.

The first study was to determine whether maternal, postnatal, and early childhood lead exposure can alter the differentially methylated regions (DMRs) that control the monoallelic expression of imprinted genes involved in metabolism, growth, and development (Li et al., 2016). In this study, we reported that mean blood lead concentration from birth to 78 months was associated with a significant decrease in PEG3 DMR methylation. For 1 μg/L increase of the mean blood lead concentration, $\Delta\beta = -0.0014117$ if 'β-model-coef' method is used; $\Delta\beta = -0.0014489$ if 'M-model-coef' method is used; $\Delta\beta = -0.0014491$ if 'M-model-M-mean' method is used; $\Delta\beta = -0.0014495$ if 'M-model- β-mean' method is used; $\Delta\beta = -0.0012782$ if the simple method is used. There were no dramatic differences among all the methods in this example (note that mean β_value for PEG3 DMR methylation is 0.43, which is in (0.2, 0.8)).

The second study was to determine the association of CpG site changes with concentration of methylmercury (MeHg), major polychlorinated biphenyls (PCBs) and other organochlorine compounds (Leung et al., 2018). For this example, we only consider the association between PCB congener 105 (PCB105) and CpG site cg20619296. For 1 μg/g increase of PCB105, $\Delta\beta = -0.099$ if 'β-model-coef' method is used; $\Delta\beta = -0.195$ if 'M-model-coef' method is used; $\Delta\beta = -0.195$ if 'M-model-M-mean' method is used; $\Delta\beta$

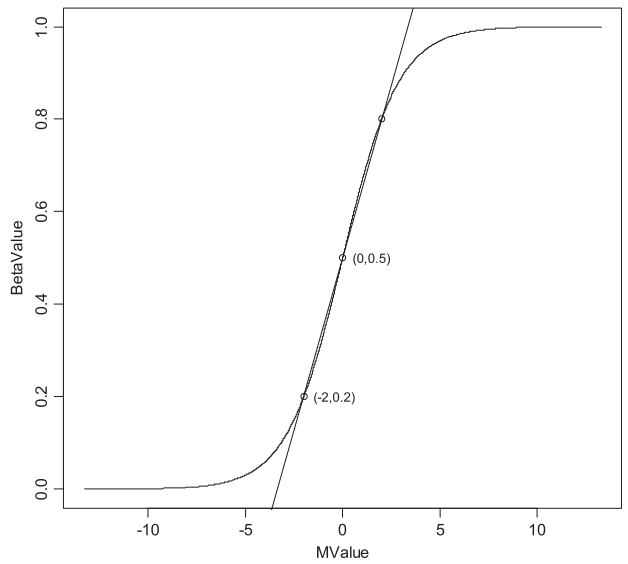

Fig. 1. The relationship curve between M-value and Beta-value

$=-0.203$ if 'M-model- β-mean' method is used; The simple method is not suitable for this case since the mean β_value for CpG site cg20619296 is 0.9736, which is not in (0.2, 0.8). For this example, we can see there is a large difference between 'β-model-coef' method and 'M-model-coef' method (or 'M-model-M-mean' method). $\Delta\beta$ given by 'M-model-coef' method (or 'M-model-M-mean' method) is almost a double of $\Delta\beta$ given by 'β-model-coef' method.

6 Discussion

Both β-value and M-value are commonly used to measure methylation levels. The M-value is more statistically valid for the differential analysis of methylation levels in relation to exposure x. However, the β-value is much more biologically interpretable to show how much methylation was changed. In this article, we proposed three different methods to calculate differential methylation values, $\Delta\beta$ from M-value linear regression model. Under the condition that M-value linear regression model is correct, we showed that the method M-model-coef is the best among the four methods. M-model-M-mean method works very well too. If the coefficients $\alpha_0,\ \alpha_2, \ldots \alpha_p$ are not given (as 'MethLAB' package, Kilaru *et al.*, 2012), the M-model-M-mean method should be used. The $\Delta\beta$ directly obtained from β-value linear regression model can give very biased results, especially when M-values are not in $(-2, 2)$ or β-values are not in (0.2, 0.8). Note the β-value distribution across the methylome are not uniform, but more like 'U-shape' between 0 and 1. That means, most of the methylation sites have β-values outside of (0.2, 0.8), and the conclusion from this article can provide very valuable suggestions for better estimating the change of methylation level.

Acknowledgements

The author wishes to thank Dr Philippe Grandjean and the reviewers for their insightful and constructive comments that have greatly improved this article.

Funding

This work was in part supported by the National Institute of Environmental Health Sciences of the National Institutes of Health [P30-ES006096].

Conflict of Interest: none declared.

References

Du,P. *et al.* (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Kilaru,V. *et al.* (2012) MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics*, **7**, 225–229.

Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.

Leung,Y.K. *et al.* (2018) Identification of sex-specific DNA methylation changes driven by specific chemicals in cord blood in a Faroese birth cohort, *Epigenetics*, **13**, 290–300.

Li,Y. *et al.* (2016) Lead exposure during early human development and DNA methylation of imprinted gene regulatory elements in adulthood. *Environ. Health Perspect.*, **124**, 666–673.

Moran,S. *et al.* (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.

Saadati,M. and Benner,A. (2014) Statistical challenges of high-dimensional methylation data. *Stat. Med.*, **33**, 5347–5357.

Sandoval,J. *et al.* (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.

Thirlwell,C. *et al.* (2010) Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip. *Methods*, **52**, 248–254.

Weinhold,L. *et al.* (2016) A statistical model for the analysis of beta values in DNA methylation studies. *BMC Bioinformatics*, **17**, 480.