Biostatistics (2021) **22**, 1, *pp.* 35–50
doi:10.1093/biostatistics/kxz016
Advance Access publication on May 20, 2019

MethylSeqDesign: a framework for Methyl-Seq genome-wide power calculation and study design issues

PENG LIU[†]

Department of Biostatistics, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA 15261, USA

CHIEN-WEI LIN[†]

Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226, USA

YONGSEOK PARK, GEORGE TSENG[*]

Department of Biostatistics, University of Pittsburgh, 130 De Soto Street, Pittsburgh, PA 15261, USA
ctseng@pitt.edu

Summary

Bisulfite DNA methylation sequencing (Methyl-Seq) becomes one of the most important technologies to study methylation level difference at a genome-wide scale. Due to the complexity and large scale of methyl-Seq data, power calculation and study design method have not been developed. Here, we propose a "MethylSeqDesign" framework for power calculation and study design of Methyl-Seq experiments by utilizing information from pilot data. Differential methylation analysis is based on a beta-binomial model. Power calculation is achieved using mixture model fitting of p-values from pilot data and a parametric bootstrap procedure. To circumvent the issue of existing tens of millions of methylation sites, we focus on the inference of pre-specified targeted regions. The performance of the method was evaluated with simulations. Two real examples are analyzed to illustrate our method. An R package "MethylSeqDesign" to implement this method is publicly available.

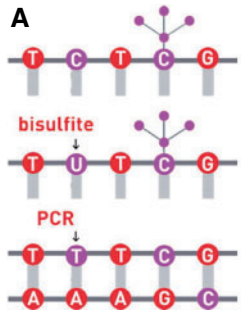
Keywords: Bisulfite sequencing; Methyl-Seq; Power calculation; Study design.

1. Introduction

DNA methylation is a chemical modification of DNA nucleotides when a methyl-group (CH_3) is attached at the 5th position of cytosine (5mC). It is one of the best characterized and the most studied epigenetic markers, which has shown to control gene expression in both normal cell development and abnormal biological process such as cancer. Particularly in gene promoter regions, hyper-methylation is shown closely related to silencing gene expressions. In mammals, such as human, DNA methylation happens almost exclusively at cytosine site that follows with guanine known as CpG site. There are tens of thousands of regions with a high frequency of CpG sites in the whole genome that are classified as CpG islands,

[†]The first two authors contributed equally to this work.
[*]To whom correspondence should be addressed.

© The Author 2019. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.



B

Single hypothesis testing	Methyl-Seq genome-wide screening
Type I error rate α	False discovery rate (FDR)
Power 1-β	Expected discovery rate (EDR)
Univariate effect size θ	Distribution of effect size of DML/DMRs

Fig. 1. (A) An illustration of sodium bisulfite modification. (B) Comparison of three elements between single hypothesis testing and Methyl-Seq genome-wide screening.

which typically exist at or near the transcription starting sites of genes. DNA methylation process has been found to link to many important biological processes, such as genomic imprinting, X-chromosome inactivation, repression of repetitive elements, aging, and carcinogenesis (Li *and others*, 1993; Paulsen and Ferguson-Smith, 2001; Robertson, 2005). In cancer studies, aberrant DNA methylation changes are considered as one of the leading factors in developing tumors (Esteller, 2005; Baylin, 2005; Delpu *and others*, 2013; Licht, 2015).

Over the past couple decades, sodium bisulfite treatment has become widely used tool to study DNA methylation at the level of single nucleotide resolution. When DNA is treated with sodium bisulfite, the unmethylated cytosines are converted to uracil and amplified by polymerase chain reaction as thymine while methylated cytosines remain protected from this conversion (see Figure 1(A)). The outcome of this treatment leads to identifying methylated and unmethylated cytosines when the sequencing reads are mapped to reference genome using special mapping pipelines such as Bismark, which consider Thymine/Cytosine mismatch. Two major technologies have been developed to quantify the DNA methylation after bisulfite conversion. One is methylation microarray, which targets on pre-selected CpG sites in certain regions, mostly within CpG islands. The total number of targeted CpG sites is relatively small, for example, Illumina HumanMethylation27 and HumanMethylation450 Bead chips cover only about 27K and 480K CpG sites) compared to over 28 million CpG sites in human genome (Schumacher *and others*, 2006). Another recently developed technology is coupling bisulfite conversion with next generation sequencing (NGS) to quantitatively query the methylation status across the whole genome. Whole genome bisulfite sequencing (WGBS) can provide accurate and quantitative estimates of the proportion of methylated cells in a population at each of the tens of millions of CpG sites across genome. However, accurate estimates of methylation level requires large number of reads to cover CpG sites of interest. Because of unevenly distributed CpG sites across the genome, a large proportion of sequencing reads do not contain any CpG sites, which results in high cost of WGBS. To overcome this disadvantage, reduced representation bisulfite sequencing (RRBS) (Meissner *and others*, 2005) was introduced to target on CpG rich regions, relying on restriction enzyme that can ensure the capture of at least one CpG site per sequencing read. Using RRBS, methylation levels of a portion of genome regions can be accurately obtained at much lower cost compared to WGBS. Here, we use "Methyl-Seq" to refer to bisulfite sequencing technology including WGBS and RRBS. Due to the popularity of Methyl-Seq and the high sequencing cost with limited budget, sample size and power calculation methods become critical for design of such studies.

Traditional power calculation methods seek the statistical power ($1-\beta$; β here is type II error) to detect the difference between groups by pre-specifying effect size (θ), type I error rate (α), and sample size (N). Alternatively, one can calculate the required sample size with pre-specified statistical power, θ and α. The effect size θ is the measure of group difference that is generally obtained from a pilot study or researchers' belief. Usually, the type I error rate is set to be 5% and the desired statistical power is

70–80% for a study design. This classical framework is based on performing a single hypothesis testing. For high-throughput genome-wide experimental data, however, many hypotheses are tested simultaneously to compare the methylation difference at the thousands of regions or millions of CpG sites. Therefore, genome-wide power calculation should be based on appropriately controlled type I error rates. One widely used in Genomic study is false discovery rate (FDR; Benjamini and Hochberg, 1995) and in this article, we use FDR to control genome-wide type I error rate. In addition, Gadbury *and others* (2004) introduced a useful concept, called expected discovery rate (EDR), to replace test power $1-\beta$ from single hypothesis to address genome-wide detection power. Since genome-wide screening considers the whole set of differentially methylated loci/regions (DML/DMRs), specifying a single effect size θ for power calculation is no longer valid. Alternatively, the distribution of effect sizes of DML/DMRs is needed, which could be estimated from pilot data. Figure 1(B) shows changes of the essential elements in genome-wide screening, compared to traditional single hypothesis testing.

There are three unique characteristics inside Methyl-Seq data that should be considered in power and sample size calculation. First, it generates random binomial data for each sample at each CpG site. A model with discrete distributions is more suitable for Methyl-seq data and both sampling and biological variations should be considered. For this reason, the beta-binomial model (Dolzhenko and Smith, 2014; Feng *and others*, 2014; Park *and others*, 2014) has gained popularity over the binomial model. Second, for Methyl-Seq experiments, researchers have choices of different sequencing depth (R) for the design. In other words, one can choose to process one sample per lane, which results in roughly 250 million reads in Illumina HiSeq 2500 platform or three samples per lane each with 83 million reads for the same sequencing cost. Therefore the power calculation problem may need to consider both N and R. Finally, there are about 28 million CpG sites in human genome. Based on the current technology, it is impossible to sequence most CpG sites with sufficient coverage even with ultra-deep sequencing depth. As a result, many CpG sites will have zero or almost zero reads in many subjects. We further discussed the coverage of single CpG sites and CpG region in the Section 1 of the supplementary material available at *Biostatistics* online. Therefore, it is not realistic to study the differences of methylation levels at all CpG sites. To circumvent this difficulty, we restrict to methylation regions by aggregating methylation data across multiple CpG sites within a particular region, such as promoter regions, for power calculation.

Many power calculation methods have been developed for RNA-Seq data, such as RNASeqPower (Hart *and others*, 2013), Scotty (Busby *and others*, 2013), and PROPER (Wu *and others*, 2015). For microarray methylation data, Tsai and Bell (2015) proposed method for study design and power calculation. To the best of our knowledge, for Methyl-seq data, no existing power calculation method has been developed so far in the literature. Here, we propose a statistical framework "MethylSeqDesign" for sample size and power calculation for studies with Methyl-seq data. The "MethylSeqDesign" R package is publicly available at `https://github.com/liupeng2117/MethylSeqDesign`.

The article is structured as follows. In Section 2, the statistical framework of MethylSeqDesign is proposed. In Section 3, we present comprehensive simulations and real data applications. Section 4 is real data application. Section 5 provides conclusion and discussion.

2. Genome-wide power calculation in Methyl-Seq

2.1. *Notations and terminology*

Consider $D_0=\{Y=(y_{gj})_{G\times(n_0+n_1)}, M=(m_{gj})_{G\times(n_0+n_1)}, X=(x_{jp})_{(n_0+n_1)\times P}\}$ $(1\le g\le G, 1\le j\le n_0+n_1)$ a pilot Methyl-Seq dataset, where y_{gj} and m_{gj} represent the methylated and total read counts for CpG region g of subject j, respectively. Let X be a design matrix of dimension $(n_0+n_1)\times P$, which contains case/control group information and other continuous or discrete covariates. n_0 and n_1 are the number of controls and cases in the pilot data. Denote N_0 and N_1 the target number of controls and cases for power

calculation. Let $R_j = \sum_{g=1}^{G} m_{gj}$ be the total number of reads observed in subject j (a.k.a. library size). We consider genome-wide power calculations under genome-wide type I error control using FDR = E (number of claimed false positives/number of claimed positives). Following Gadbury *and others* (2004), we use expected discovery rate, EDR = E (number of claimed true positives/number of total true positives), as the genome-wide power. The methylation level is the proportion of methylated cells among all cells at a particular CpG site or region. Statistical power is impacted by both sample size and sequencing depth. Therefore, the statistical framework of MethylSeqDesign becomes to estimate the genome-wide power $\widehat{\text{EDR}}(N_0, N_1, R|D_0)$ based on the pilot data (D_0 with n_0 controls, n_1 cases, and sequencing depth R_0) for designing a future experiment with N_0 controls, N_1 cases, sequencing depth R, and under a prespecified FDR level (e.g. FDR = 5%).

2.2. *Three sequential steps for genome-wide Methyl-Seq power calculation*

Park and Wu (2016) proposed the "DSS-general" method, a model-based method for detecting DML/DMRs based on beta-binomial model with arcsine link function. The estimation procedure is based on generalized least square (GLS) approach, which can significantly reduce the computation demands compared to other beta-binomial based methods (Dolzhenko and Smith, 2014; Feng *and others*, 2014). In addition, Park and Wu (2016) showed superior performance of their method in terms of DMR detection accuracy and type I error rate control. Therefore, in this article, we will adopt Park and Wu (2016)'s approach for our power calculation tool.

Below we propose three sequential steps in MethylSeqDesign to estimate EDR. In Step I, p-values and effect size distribution of all methylated regions from pilot data are obtained using DSS-general. In Step II, a beta-uniform mixture (BUM) model is applied to characterize the genome-wide p-value distribution and to estimate the proportion of true DMRs. In Step III, a parametric bootstrapping method based on DMR posterior probability is used to simulate and transform the genome-wide p-value distribution towards the targeted sample size and sequencing depth. The detailed description of our method is as follows.

Step I. Differential methylation analysis on pilot data.

To account for both sampling and biological variation, denote by Y_{gj} the methylated read count for gene $g (g = 1, 2, ..., G)$ in sample $j (j = 1, 2, ..., (n_0 + n_1))$, let q_{gj} be the underlying methylation level for gene g and sample j, $Y_{gj} \sim \text{bin}(m_{gj}, q_{gj})$ and $q_{gj} \sim \text{beta}(\alpha_{gj}, \beta_{gj})$. Marginally, $Y_{gj} \sim \text{Beta-bin}\left(m_{gj}, \pi_{gj}, \phi_g\right)$, where π_{gj} and ϕ_g are the mean and dispersion parameter of beta distribution, such that $\pi_{gj} = E(q_{gj}) = \frac{\alpha_{gj}}{\alpha_{gj}+\beta_{gj}}$, $\phi_g = \frac{1}{\alpha_{gj}+\beta_{gj}+1}$ and we assume $\phi_{gj} = \phi_g$ for all $j = 1, 2, ..., (n_0 + n_1)$. Here, we account for covariate effect as,

$$\arcsin\left(2\pi_{gj} - 1\right) = \boldsymbol{x}_j \boldsymbol{\beta}_g, \tag{2.1}$$

where $\boldsymbol{x}_j = (x_{j1}, x_{j2}, ..., x_{jp})$ is jth subject's covariate, and $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, ..., \beta_{gp})^T$ is a vector of p covariate coefficients for gth CpG region.

Denote $A_{gj} = \arcsin\left(2Y_{gj}/m_{gj} - 1\right)$. As shown in Park and Wu (2016), the expectation of A_{gj} can be approximated as $E\left(A_{gj}\right) \approx \arcsin\left[2E\left(Y_{gj}\right)/m_{gj} - 1\right] = \arcsin\left(2\pi_{gj} - 1\right) = \boldsymbol{x}_j \boldsymbol{\beta}_g$. Furthermore, the variance of A_{gj} can also be also approximated as $Var\left(A_{gj}\right) \approx \frac{1+(m_{gj}-1)\phi_g}{m_{gj}}$, which is approximately independent of the mean structure. Given dispersion parameter ϕ_g, the regression coefficients $\boldsymbol{\beta}_g$ can be estimated using GLS method, i.e. $\hat{\boldsymbol{\beta}}_g = \left(X^T V_g^{-1} X\right)^{-1} X^T V_g^{-1} A$, where $V_g = \text{diag}\left(\frac{1+(m_{gj}-1)\phi_g}{m_{gj}}\right)$ is the covariance matrix. The estimator of ϕ_g is given by $\hat{\phi}_g = \frac{(n_0+n_1)\left(\hat{\sigma}_g^2 - 1\right)}{\Sigma_j (m_{gj}-1)}$, where $\hat{\sigma}_g^2 = \frac{\Sigma_j m_{gj}\left(A_{gj} - \boldsymbol{x}_j \hat{\boldsymbol{\beta}}_g^{(0)}\right)}{n_0+n_1-p}$, $\hat{\boldsymbol{\beta}}_g^{(0)}$ is

the GLS estimator of β_g under $\hat{\phi}_g = 0$. The estimate of covariance structure is $\hat{V}_g = \text{diag}\left(\frac{1+(m_{gj}-1)\hat{\phi}_g}{m_{gj}}\right)$. Given $\hat{\phi}_g$ and $\hat{V}_g$, the estimator of variance of $\hat{\boldsymbol{\beta}}_g$ is $\hat{\Sigma}_g \equiv \hat{\text{var}}\left(\hat{\boldsymbol{\beta}}_g\right) = \left(X^T \hat{V}_g^{-1} X\right)^{-1}$.

Hypothesis testing for $H_0 : C^T\beta = 0$ vs. $H_A : C^T\beta_g \neq 0$ is based on Wald statistics

$$Z_g(\boldsymbol{C}) = \frac{\boldsymbol{C}^T \hat{\boldsymbol{\beta}}_g}{\sqrt{\boldsymbol{C}^T \hat{\Sigma}_g \boldsymbol{C}}},$$

where $\boldsymbol{C}$ can be any linear combination of the covariate effects. The statistic approximately follows a standard normal distribution under null hypothesis.

For simplicity, here we consider the study with two groups (case and control), i.e. $p = 2$ with intercept and case/control effect. Let $n_0 = n_1 = n$ be the number of subjects in each group in pilot data, and $N_0 = N_1 = N$ be the target number of subjects in each group. Here, we assume equal sample sizes for control and cases in pilot and target cohorts while the method can be easily generalized for $n_0 \neq n_1$ and $N_0 \neq N_1$ later (see the leukemia application in Section 4.2). Then the model becomes $\arcsin\left(2\pi_{gj} - 1\right) = \beta_{0g} + x_{j2}\beta_{1g}$. In this case, $\boldsymbol{C} = (0, 1)$. Here the variance of $\hat{\beta}_{1g}$ is

$$\begin{aligned}
\widehat{\text{Var}}\left(\hat{\beta}_{1g}\right) &= \frac{\sum_{j=1}^{n_0+n_1} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}}}{\sum_{j_1=1}^{n_0} \frac{m_{gj_1}}{1+(m_{gj_1}-1)\hat{\phi}} \times \sum_{j_2=1}^{n_1} \frac{m_{gj_2}}{1+(m_{gj_2}-1)\hat{\phi}}} \\
&= \frac{\sum_{j=1}^{2n} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}}}{\sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+(m_{gj_1}-1)\hat{\phi}} \times \sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+(m_{gj_2}-1)\hat{\phi}}} \\
&= \frac{n \times \left(\frac{1}{n}\sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+(m_{gj_1}-1)\hat{\phi}} + \frac{1}{n}\sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+(m_{gj_2}-1)\hat{\phi}}\right)}{n^2 \times \frac{1}{n}\sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+(m_{gj_1}-1)\hat{\phi}} \times \frac{1}{n}\sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+(m_{gj_2}-1)\hat{\phi}}} \\
&= \frac{1}{n}\frac{\bar{A}_g + \bar{B}_g}{\bar{A}_g \times \bar{B}_g} \\
&= \frac{1}{n}\Psi_g,
\end{aligned} \tag{2.2}$$

where $\bar{A}_g = \frac{1}{n}\sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+(m_{gj_1}-1)\hat{\phi}}$ and $\bar{B}_g = \frac{1}{n}\sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+(m_{gj_2}-1)\hat{\phi}}$. The Wald test statistic becomes

$$Z_g = \frac{\hat{\beta}_{1g}}{\sqrt{\hat{\text{var}}(\hat{\beta}_{1g})}} = \frac{\hat{\beta}_{1g}}{\sqrt{\frac{1}{n}\Psi_g}}. \tag{2.3}$$

where $\hat{\beta}_{1g}$ is the GLS estimator of β_{1g}.

Remark 1: A common over-dispersion parameter ($\hat{\phi}$) over all CpG regions is used which is the mean of all tag-wise dispersion parameters ($\hat{\phi}_g$) estimated from the procedure proposed by Park and Wu (2016). This

is because when sample size is small, estimation of region-specific dispersion parameter is not precise, and region-specific power calculation is very challenging.
Remark 2: One interesting finding is that the quantities $\bar{A}_g$ and $\bar{B}_g$ in (2.2) are in the mean form that depends on coverage and dispersion parameter given a region g. When m_{gj} is small, it has negative correlation with Ψ_g, while $\bar{A}_g$ and $\bar{B}_g$ are roughly independent on m_{gj} when m_{gj} is large. If we assume the dispersion parameter stay constant, Ψ_g is mostly impacted by sequencing depth. In Section 2.1 of the supplementary material available at *Biostatistics* online, we further studied the property of the quantity Ψ_g by simulation.

Step II. Mixture model fitting for p-value distribution.

A BUM model (Allison *and others*, 2002) has been proposed to fit the p-value distribution. To be specific, we use a beta distribution $f_1(p|r,s)$ with shape parameter r and s ($0 < r < 1 \leq s$) for p-values of DMRs and a uniform distribution $f_0(p)$ for p-values of non-DMRs. The mixture density of overall p-value distribution is $f(p|r,s,\lambda) = \lambda f_0(p) + (1-\lambda) f_1(p|r,s)$, where λ is the proportion of non-DMRs. The constraints for r and s is used to have a proper shape for the p-value distribution of DMRs. A proper estimation of λ is essential in fitting a BUM model. We apply censored BUM (CBUM) proposed by Markitsis and Lai (2010) to reduce the impact of extremely small p-values, since our main purpose is to estimate the proportion of true DMRs for those with relatively larger p-values. The detailed comparisons of performance between BUM and CBUM methods are included in Section 2.2 of the supplementary material available at *Biostatistics* online. The shape parameters r and s can then be estimated using maximum likelihood approach using $\hat{\lambda}$ estimated from the CBUM method.

Step III. Parametric bootstrapping based on DMR posterior probability to estimate EDR.

Theoretically, the p-value distribution for non-DMRs with zero effect size follows a uniform distribution that does not change with the sample size. However, we expect that the p-values for those DMRs will be more significant as sample sizes increase. Equations (2.2) and (2.3) reveal a transformation of Z-statistics of DMRs from the pilot data with sample size n to the targeted sample size N. When the effect size $\hat{\beta}_{1g}$ and the common over-dispersion parameter $\hat{\phi}$ stay approximately unchanged, the Wald test statistics change by a factor of $\sqrt{\frac{N}{n}} \times \sqrt{\frac{\Psi_g}{\Psi_g'}}$ (see (2.2) and item (2) below; n is the pilot sample size, N is the targeted sample size, Ψ_g is the quantity under pilot sequencing depth R_0, and Ψ_g' is the quantity under the targeted sequencing depth R). Throughout this article, we assume sequencing depth of pilot data R_0 is deep enough and the targeted sequencing depth R does not exceed R_0 (i.e. $R \leq R_0$). Since Ψ_g is a function depending on pre-estimated dispersion parameter $\hat{\phi}_g$ and count data $\{m_{gi1}, m_{gj2}; 1 \leq g \leq G, 1 \leq j \leq (n_0 + n_1)\}$, it is readily calculated from pilot data. To estimate for Ψ_g', we can randomly subsample from pilot data to achieve sequencing depth R and derive Ψ_g' by definition based on subsampled counts m_{gj}' and $\hat{\phi}_g$. The influence of sequencing depth on Ψ_g is further discussed in Section 2.1 of the supplementary material available at *Biostatistics* online. We found that when the median coverage level is above 160, Ψ_g is roughly constant and the correction term for different sequencing depth is not necessary. Otherwise, the correction term $\sqrt{\frac{\Psi_g}{\Psi_g'}}$ is needed.

Let I_g be the latent variable indicating region g a DMR (I_g=1) or non-DMR (I_g=0), and let p_g be the p-value of region g from the aforementioned Wald test in pilot data. The detailed parametric bootstrapping procedure is described as follows:

(1) Calculate the posterior probability of the DMR indicator I_g with posterior probability

$$P(I_g = 1|\hat{\lambda}, \hat{r}, \hat{s}, p_g) = \frac{(1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s})}{\hat{\lambda} + (1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s})},$$

where $\hat{\lambda}$, $\hat{r}$, and $\hat{s}$ are estimated in Step II. In the bth parametric bootstrapping ($1 \le b \le B$), draw $I_g^{(b)}$ from $P(I_g|\hat{\lambda}, \hat{r}, \hat{s}, p_g)$ for $1 \le g \le G$.

(2) Transform Z-statistics for DMRs using equation

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \sqrt{\frac{N}{n}} \times \sqrt{\frac{\Psi_g}{\Psi_g'}} + (1 - I_g^{(b)}) \times Z_g,$$

where we assume that the effect size $\hat{\beta}_{1g}$ and the common over-dispersion parameter $\hat{\phi}$ of a DMR in (2.2) and (2.3) are roughly fixed. Therefore, when $I_g^{(b)} = 1$, region g is a DMR in the bth parametric bootstrap and the Wald statistic is transformed to $Z_g \times \sqrt{\frac{N}{n}} \times \sqrt{\frac{\Psi_g}{\Psi_g'}}$. When $I_g^{(b)} = 0$, the Wald statistic remains unchanged.

(3) Compute p-value based on the 2-sided test: $p_g^{(b)} = 2 \times (1 - \Phi(|Z_g^{(b)}|))$ for a DMR region ($I_g^{(b)} = 1$), where Φ is a cumulative density function of a standard normal distribution. When $I_g^{(b)} = 0, Z_g^{(b)} = Z_g$, and $p_g^{(b)} = p_g$ remain unchanged.

(4) Control FDR at level α:

(a) In the bth simulation, calculate $\text{FDR}^{(b)}(u) = \frac{\sum_{g=1}^{G}(1-I_g^{(b)})\cdot\chi_{(p_g^{(b)} \le u)}}{\sum_{g=1}^{G}\chi_{(p_g^{(b)} \le u)}}$ for a given p-value threshold u, where $\chi(\cdot)$ is an indicator function that takes value one when the statement is true and zero otherwise. Here, by definition, the denominator is the number of detected regions under p-value threshold u, and the numerator is the number of non-DMRs among those detected regions.

(b) Let $u^{(b)} = \underset{u}{\text{argmax}}(\text{FDR}^{(b)}(u)) \le \alpha)$, where $u^{(b)}$ is the p-value threshold to keep FDR at α level for the bth simulation.

(5) Obtain the estimated EDR for the bth simulation with $\widehat{\text{EDR}}^{(b)} = \frac{\sum_{g=1}^{G} I_g^{(b)}\cdot\chi_{(p_g^{(b)} < u^{(b)})}}{\sum_{g=1}^{G} I_g^{(b)}}$. Here, by definition, the denominator is the number of total DMRs and the numerator is the number of detected true DMRs.

(6) Repeat steps (1) to (5) for B times and the robust estimator of EDR from the B simulations is $\widehat{\text{EDR}}(N|D_0) = \underset{b}{median}(\widehat{\text{EDR}}^{(b)})$. The first and third quantile of bootstrapped EDRs can also be derived to account for the variability of EDR estimation.

3. Simulation

3.1. *Simulation scheme*

We simulated data based on parameters estimated directly from the mouse pregnancy dataset (see details in Section 4.1 (Katz *and others*, 2015)). We empirically drew the total number of reads and baseline methylation level of control group from the data. Effect size (methylation level difference between two groups) was either fixed or randomly generated from $U(0.1, 0.2)$. In total, 10,000 regions were simulated, and we assigned 10% regions as DMRs. The common dispersion parameter was set to $\phi = 0.048$, which was the mean dispersion parameters estimated from the data.

The steps to simulate pilot data with sample size n and sequencing depth R_0, and targeted data with sample size N and sequencing depth R are shown below.

(1) Draw total read m_{gj} for region g and sample j randomly from the mouse pregnancy data, and baseline methylation level q_g for each CpG region from the empirical distribution estimated from the mouse pregnancy data.
(2) To simulate pilot data with sequencing depth R_0, we directly use the total reads drawn from step 1. When simulating the targeted data with sequencing depth $R \neq R_0$, we downsample the matrix of total reads based on the ratio of $\frac{R}{R_0}$.
(3) DM index: Generate random number I_g from $U(0,1)$ for each region. If $I_g \leq 0.1$ the gth region is DMR and $I_g = 1$. Otherwise, it is non-DMR and $I_g = 0$. This generates roughly 10% DMRs.
(4) Effect size $\triangle$: Draw effect size from $U(0.1, 0.2)$ for each DMR. The effect size for non-DMRs is set to 0.
(5) Generate the number of methylated reads: If the gth region is non-DMR, then $y_{gj} \sim$ Beta-bin(m_{gj}, q_g, ϕ). If the gth region is DMR, then in control group $y_{gj} \sim$ Beta-bin(m_{gj}, q_g, ϕ) while in case group, the methylated counts $y_{gj} \sim$ Beta-bin(m_{gj}, q'_g, ϕ), where $q'_g = q_g + \triangle_g$ if $q_g \leq 0.5$ and $q'_g = q_g - \triangle_g$ otherwise.
(6) Follow above steps to simulate pilot data and the targeted data.

3.2. *Performance comparison with other hypothesis testing methods*

We compared the statistical power of our proposed test statistic (3) with other three methods: Beta value $(2Y_{gj}/m_{gj})$ with t-test, M value (logit $(2Y_{gj}/m_{gj})$) with t-test, and A value ($A_{gj} = \arcsin(2Y_{gj}/m_{gj} - 1)$) with t-test.

To compare the performance, we conducted the analysis by stratifying the baseline methylation proportion in control group into three categories: low ($0 < q_g < 0.2$), medium ($0.2 < q_g < 0.8$), and high ($0.8 < q_g < 1$). In each baseline group, we simulated 20 times independent analysis, in which pilot data had 10 subjects in each group (i.e., $n_0 = n_1 = 10$), and 10 000 regions (10% are DMRs). As shown in Figure S3 of the supplementary material available at *Biostatistics* online, we compared the power based on how many true DMRs could be declared among different numbers of top declared DMRs. As a result, the result clearly shows better performance of using our arcsin transformation and Wald statistics compared to other approaches. Furthermore, we observe that the power of each method is stronger in either low or high baseline group and relatively weaker in medium baseline group, which is reasonable because the effect size is at methylation level scale and for binomial distribution, the same difference is easier to detect when methylation is close to boundary 0 or 1. Overall, the results justifies the need of using arcsine transformation and Wald test statistics for our power calculation framework.

3.3. *Performance evaluation*

We simulated B=10 pilot datasets ($b = 1, 2, ...B$) with pilot sample size n_0= 2, 4, 6, 8, 9, and 10 when R_0 are 5 million reads. For each pilot dataset with (n_0, R_0), the projected power for targeted sample size N_i=2, 6, 10, 15, 25, 50 ($i = 1, 2, ..., 6$) and R_j = 0.25, 0.5, 1, 2, 3, 4 million reads ($j = 1, 2, ..., 6$) from a power calculation method is denoted by $\widehat{\text{EDR}}(N_i, R_j; n_0, R_0)$. Since the underlying truth is known, the true EDR for each (N_i, R_j) can be estimated as $\widehat{\text{EDR}}(N_i, R_j) = \frac{\sum_{b=1}^{B} \widehat{\text{EDR}}^{(b)}(N_i, R_j)}{B}$, where $\widehat{\text{EDR}}^{(b)}(N_i, R_j)$ is the actual EDR in the bth simulation when sample size N_i and R_j are simulated. We propose the following benchmarks based on root mean squared error (RMSE) to evaluate performance of different power calculation methods:

Table 1. *Performance evaluation in simulation study stratified by different effect sizes. Performance evaluation based on RMSE of* $\hat{\text{EDR}}(D; D_0)$ *in simulation analysis. Results based on different pilot sample size (n_0 = 2, 4, 6, 8, 9, and 10) are shown in different rows. In the first three columns, stratified analysis is performed as* $\triangle$ *=0.1, 0.14, and 0.18. In the last column, "sOverall" refers to generating* $\triangle$ *from* $U(0.1, 0.2)$

	RMSE (computing time in seconds)			
Pilot n_0	$\triangle = 0.1$	$\triangle = 0.14$	$\triangle = 0.18$	Overall
2	0.27(332)	0.12(176)	0.04(171)	0.10(198)
4	0.16(190)	0.04(175)	0.04(172)	0.03(179)
6	0.08(150)	0.02(164)	0.02(144)	0.01(148)
8	0.05(153)	0.02(149)	0.01(154)	0.02(144)
9	0.03(153)	0.02(151)	0.02(158)	0.01(150)
10	0.02(116)	0.01(121)	0.03(118)	0.01(120)

(1) Consider two-dimensional power calculation from (n_0, R_0) to (N_i, R_j) $(i = 1, 2, ..., 6$ and $j = 1, 2, ..., 6)$. The RMSE of estimated EDR from power calculation is

$$\text{RMSE} = \sqrt{\frac{\sum_{b=1}^{B}\sum_{i=1}^{6}\sum_{j=1}^{6}\left[\widehat{\text{EDR}}^{(b)}\left(N_i, R_j; n_0, R_0\right) - \widehat{\text{EDR}}\left(N_i, R_j\right)\right]^2}{B \times 6 \times 6}}.$$

We first performed a stratified analysis based on different level of effect size, as we already know it will impact the EDR. $\triangle$ was set as 0.1, 0.14, and 0.18. In each setting, we generated the same number of regions to compare the performance (Figure 2 for $\triangle = 0.14$, Figures S4 and S5 of the supplementary material available at *Biostatistics* online for $\triangle = 0.1$ and 0.18). Table 1 shows the RMSEs and computing time of Figure 2, Figures S4 and S5 of the supplementary material available at *Biostatistics* online. As shown in Figure 2, similarly in Figures S4 and S5 of the supplementary material available at *Biostatistics* online, the estimated true EDR increases as the sequencing depth increases; however, the gain of EDR decreases as the sequencing depth increases. When the ratio of targeted sequencing depth to the pilot sequencing depth > 0.4 (i.e. prop > 0.4), increasing sequencing depth has almost no effect on the true EDR. The observed trend of true EDR is consistent with the trend of predicted EDR as described in Section 2.1 of the supplementary material available at *Biostatistics* online. This consistency indicates that our method can estimate EDR prediction well when sequencing depth changes. We also observed that the predicted EDR curves from MethylSeqDesign are close to the true EDR curves, and the performance improves as the sample size of pilot data (n_0) increased. The result of Table 1 shows affordable computing time (2–5 min using a regular laptop) for one run under this simulation setting. Secondly, to mimic real situation, we generated $\triangle$ from $U(0.1, 0.2)$ and compared the predicted versus true curves as shown in Figure S6 of the supplementary material available at *Biostatistics* online. We observed results similar to that in the case of fixed $\triangle$.

3.4. *Cost and benefit analysis and study design*

In this subsection, we illustrated two scenarios where our method can guide Methyl-Seq study design. In scenario 1, a desired level of EDR was given, and we would like to find the optimal combination of N

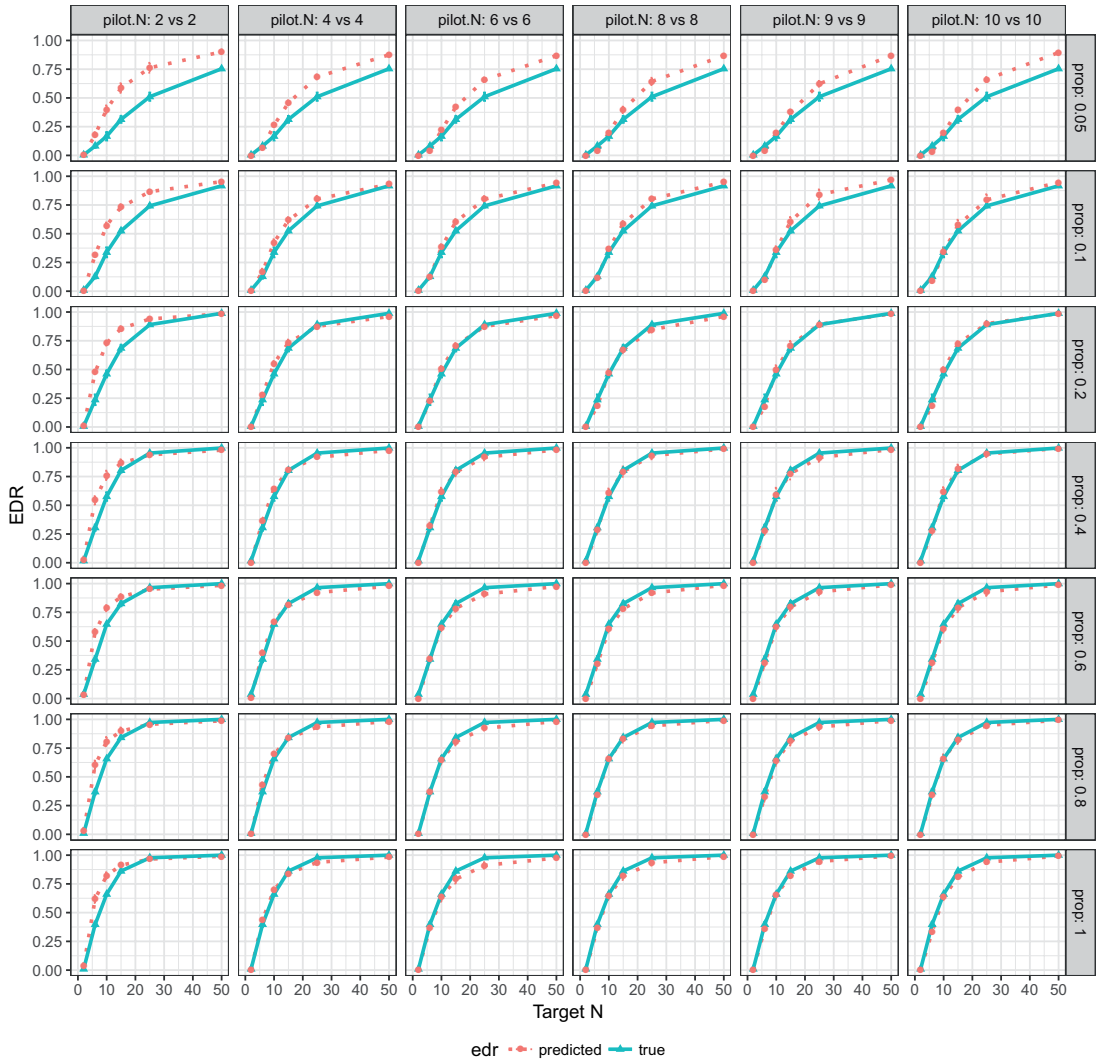


Fig. 2. EDR prediction from MethylSeqDesign compared with true EDR under different pilot data sample sizes and sequencing depth. Effect size $\triangle$ is fixed at 0.14. The pilot data sample size per group varied from 2 to 10, and $R_0 = 5M$ is fixed. The predicted EDRs by the pilot data are shown by the dotted curves. The targeted data sample size per group varied from 2 to 50, and the ratio of targed sample sequencing depth to that of pilot is prop $= \frac{R_j}{R_0}$, which varied from 0.05 to 1. The estimated true EDRs by targeted data are in solid curves.

and R that achieves the desired EDR with the minimal budget. Whereas in scenario 2, a fixed budget limit was given, and we would like to find the optimal combination of N and R which spends within the budget and maximizes the EDR. We obtained the sequencing cost information from Sequencing and Microarray Facility core at MD Anderson for this example. Price per lane (P) is \$1500 dollars when the total number of reads per lane (D) is set at 250 M with alignment rate (A) of 50%. Library preparation cost per sample (X) including bisulfite conversion treatment is \$300. The total cost can be written as:

$$\begin{aligned}\text{Total cost} &= \frac{R \cdot 2 \cdot N}{D \cdot A} \cdot P + X \cdot 2 \cdot N \\ &= \frac{R \cdot 2 \cdot N}{250 \cdot 0.5} \cdot 1500 + 300 \cdot 2 \cdot N.\end{aligned} \tag{3.4}$$

Based on (3.4), we can calculate the cost for any combination of N and R. Given this extra information, the optimal combinations of N and R in both scenarios can be derived. In scenario 1, the optimal N and R combination corresponds to the design with the lowest cost among those with the desired level of EDR; in scenario 2, the optimal design is the one with the highest EDR with costs at most the given budget.

We simulated pilot data ($n_0 = 4$ and $R_0 = 1/4$ lane) based on the settings described in Section 3.1. The targeted sample size $N = 4$ to 50 by a gap of 2, and targeted sequencing depth $R = 1/10, 1/8, 1/6, 1/4$ of one lane. The EDR for each N and R combination is estimated from pilot data and the corresponding cost for each design is calculated. The optimal design can be identified according to different constraints: (i) in scenario 1, we want to achieve at least 80% EDR; (ii) a limited budget \$20 000 dollars is given in scenario 2. For any given design (i.e. combination of N and R), if there is no other design achieving higher EDR with lower cost than the current one, it is called an admissible design (in dark color), otherwise it is an inadmissible design (in light color). Figure 3 shows the resulting N–R and cost-EDR corresponding plots. The optimal design is highlighted and circled. In scenario 1, the optimal design to achieve at least 80% EDR is to perform $N = 12$ and $R = 1/8$ lane and the design will cost \$11 700 to achieve EDR = 0.807 (see the left two plots in Figure 3). In scenario 2, with maximal budget of \$20 000, the optimal design is $N = 20$ and $R = 1/10$ lane, which will cost \$18 000 and achieve EDR = 0.922 (the two plots on the right of Figure 3).

4. Real data application

4.1. *Breast cancer mouse data*

In this subsection, we demonstrated the performance of MethylSeqDesign using the Katz's data (Katz *and others*, 2015), which was used to investigate the protective risk effect of pregnancy toward breast cancer in a mouse model. The DNA methylation data were from the mammary gland tissue. The sample library was prepared using Agilent SureSelectXT Mouse Methyl-Seq Kit. The kit design covered 109 Mb of Ensemble regulatory features (promoters, promoter flanking regions, enhancers, etc.), CpG islands, known tissue-specific DMR, and open regulatory elements. The dataset was generated with mm9 mouse reference genome (Kent *and others*, 2002). Aligned reads outside of the targeted regions (provided by Agilent SureSelectXT Mouse Kit) were removed. Data preprocessing was performed by R package "MethyKit." We only used samples from batch one, which harvested from mammary gland tissue immediately after involutions including five parous and five non-parous mice.

A total of $G = 297\,773$ methylation regions of interest (ROI) are pre-defined from the Agilent SureSelectXT kit. We randomly subsampled 2 vs. 2, 3 vs. 3, and 4 vs. 4 samples from the full data as the pilot data, and used MethylSeqDesign to calculate the predicted EDR. We repeated this procedure for 10 times. The predicted EDR from the subsampled data was compared with the reference EDR calculated using the full data. As shown in Figure 4, although the underlying true EDR is unknown, as the sample size of subsampled data increases, the predicted EDR from subsampled data converges to the predicted EDR from the full 5 vs. 5 samples.

4.2. *Chronic lymphocytic leukemia data*

Kushwaha *and others* (2016) studied hypomethylated and hypermethylated regions and how the methylation changes affect gene expression in the oncogenesis of chronic lymphocytic leukemia (CLL) (GEO

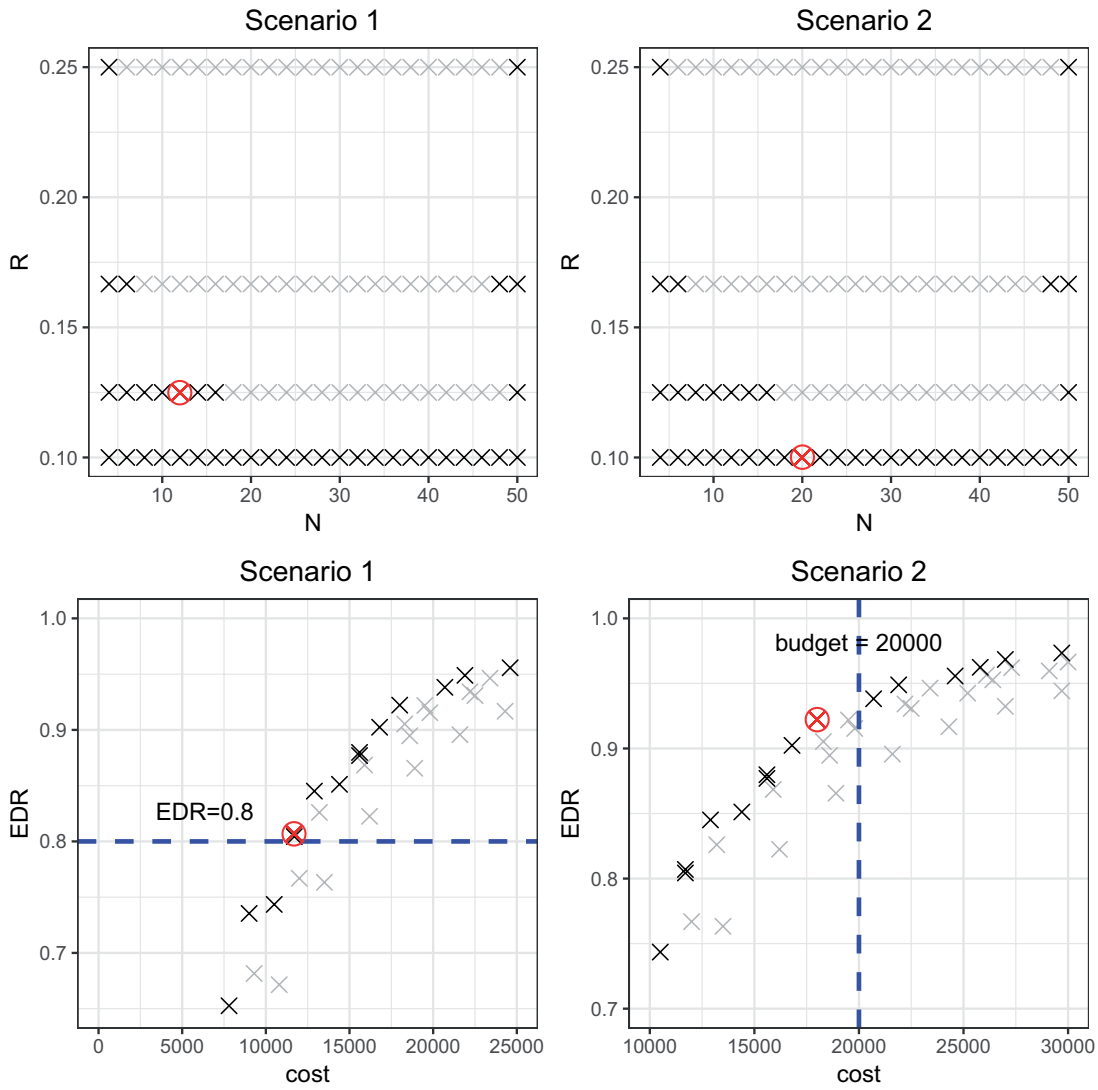


Fig. 3. Illustration of study design optimization in two scenarios. The first row plots all N and R combinations. The optimal N and R combination is highlighted and circled, the admissible combinations are highlighted in dark and inadmissible ones are in light background. The second row plots the corresponding Budget and EDR relation for N and R combinations. The dashed line marked the targeted EDR in scenario 1 and budget limit in scenario 2.

accession number GSE66167). RRBS was performed for a genome-wide DNA methylation analysis in 43 tumors and 8 controls. We implemented MethylSeqDesign to the dataset using targeted regions defined as 250 bp tiling windows with at least 10 read counts.

Similar to the previous example, the true underlying EDR is unknown in real data. We instead showed the performance of our method by comparing the predicted EDR from smaller sample size to full sample size. Since the sample size in control and tumor groups were unbalanced (number of tumor samples is roughly 5 times more than that of controls), we kept this ratio and randomly subsampled $(n_0, n_1) = (2,10)$, (4,20), and (6,30) from full dataset to treat as pilot data and repeated independent subsampling for 10 times

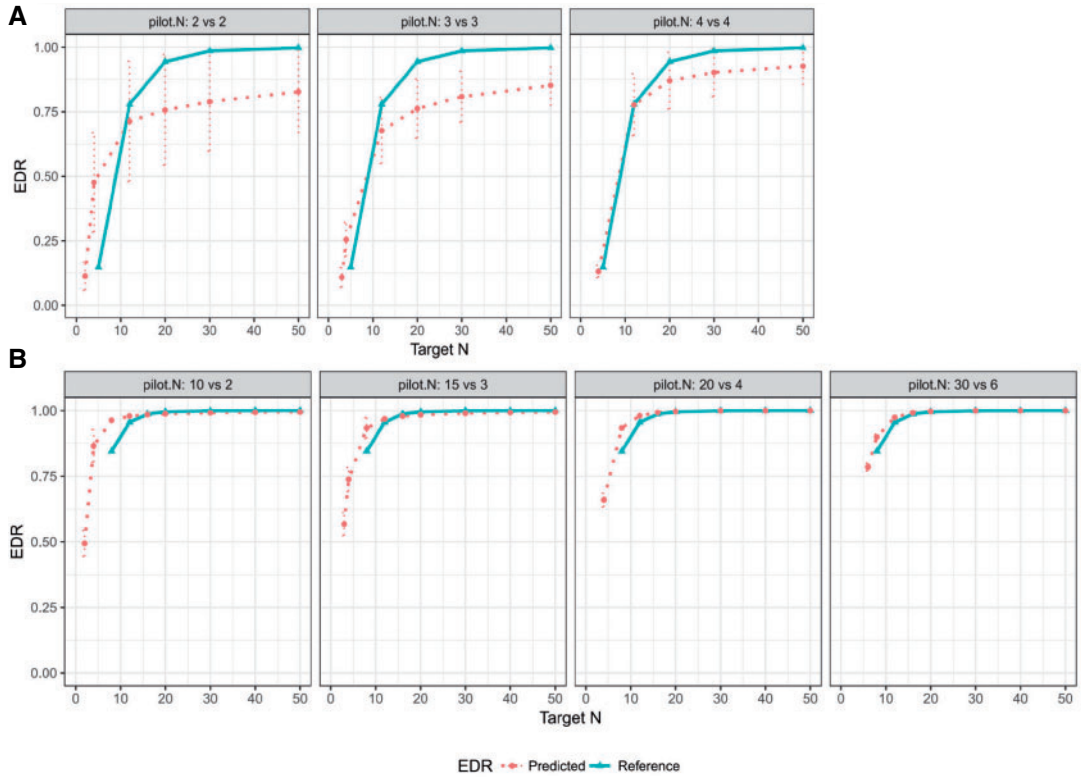


Fig. 4. (A) Real data application using mouse pregnancy dataset. (B) Real data application using CLL dataset. The mean and 95% CI of the predicted EDR from subsampled data is shown in dotted curve, and the solid curve is the reference EDR from the full data. As sample size of subsampled data increases, the predicted EDR becomes closer to the reference.

for each (n_0, n_1) pair (see formulation for unbalanced design in Section 3 of the supplementary material available at *Biostatistics* online).

For full data $(n_0, n_1) = (8, 43)$, we also derived predicted EDR and treated it as a reference to compare with predicted EDR from smaller pilot data (shown in Figure 4). Although no underlying truth was available for this application, predicted EDR from our method gave reasonably accurate results, where increased sample size in pilot data generated less variation in predicted EDR curves and converged to the result from large pilot data. In this example, power calculation using $(n_0, n_1) = (3, 15)$ pilot data is roughly sufficient. The required larger sample size is reasonable due to unbalanced design and small sample size in n_0.

5. Discussion and conclusion

NGS-based bisulfite sequencing is an increasingly important high-throughput technology to measure genome-wide methylation patterns. An important goal of Methyl-Seq is to detect DMRs, such as promoter regions and transcription biding sites. During the study design stage, it is essential to accurately estimate study power based on appropriate method, particularly when pilot data exist. Given thousands of targeted methylation regions are considered simultaneously to detect differential methylation, it is imperative that the power calculation method is able to appropriately control genome-wide type I error rate, to evaluate genome-wide statistical power and to account for varying DMR effect sizes. To our knowledge, there is no existing method for this purpose. In this article, we proposed a MethylSeqDesign statistical framework

to accommodate all three elements mentioned above with FDR, EDR, and estimating effect sizes from pilot data. This method uses a beta-binomial model to account for variations in the Methyl-Seq count data that is due to sampling variations and biological heterogeneities between subjects. We use FDR to control genome-wide type I error rate, and EDR as the genome-wide power. In addition, the use of Wald test statistic enables the transformation of statistics from pilot data to targeted sample size and sequencing depth, which allows two-way power calculation and saves computing time. Our method utilizes the pilot data to estimate the genome-wide distribution of methylation level difference between two groups (effect size) and the proportion of true DMRs, which can be efficiently avoid arbitrary guesses by researchers. Finally, with the specified cost function, we demonstrated how our method guides the selection of proper study designs in two scenarios.

The MethylSeqDesign framework needs a pilot dataset as input. It is crucial that the pilot data are technically similar to the targeted data as possible. If no pilot dataset is available in the local lab, existing datasets on the public domain with similar biological and technical setting (e.g. similar tissue, disease, and sequencing protocols) are the appropriate alternative. In general, a pilot data with larger sample size would yield a superior estimate of EDR. Although pilot sample size required for accurate power calculation depends on biological and experimental variability in each project, from our experience, $n_0 = n_1 = 5$ is usually sufficient for accurate power calculation. Our second real example shows that unbalanced n_0 and n_1 design requires larger pilot sample size.

In this article, we restricted the power calculation framework to pre-defined targeted regions since only small proportion of CpG sites (5–20%) is available for differential methylation analysis. When the sequence depth is sufficiently deep, the effect on identifying DMRs is minimal, so does on power calculation. However, sequencing depth can still play important roles when sequencing depth is not deep enough. In this article, our method allows a two-dimensional power calculation by considering both sample size and sequencing depth.

In summary, we proposed a MethylSeqDesign framework to deal with the study design and power calculation issues for epigenetic studies of DNA methylation where the ROI are prespecified. As technology advances and sequencing cost decreases, the emergence of more large-scale Methyl-Seq studies will lead to increased demand for Methyl-Seq study design and power calculation. An R package "MethylSeqDesign" is publicly available at https://github.com/liupeng2117/ MethylSeqDesign and all code and data used in this article are available at https://github.com/liupeng2117/ MethylSeqDesign_data_code.

6. SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

National Institutes of Health (NIH; R01CA190766).

REFERENCES

ALLISON, D. B., GADBURY, G. L., HEO, M., FERNANDEZ, J. R., LEE, C.-K., PROLLA T. A., WEINDRUCH, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis* **39**, 1–20.

BAYLIN, S. B. (2005). DNA methylation and gene silencing in cancer. *Nature Reviews. Clinical Oncology* **2**, S4.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300.

Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. and Marth, G. T. (2013). Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29**, 656–657.

Delpu, Y., Cordelier, P., Cho, W. C. and Torrisani, J. (2013). DNA methylation and cancer diagnosis. *International Journal of Molecular Sciences* **14**, 15029–15058.

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**(1), 215.

Esteller, M. (2005). Aberrant DNA methylation as a cancer-inducing mechanism. *Annual Review of Pharmacology and Toxicology* **45**, 629–656.

Feng, H., Conneely, K. N. and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research* **42**, e69.

Gadbury, G. L., Page, G. P., Edwards, J., Kayo, T., Prolla, T. A., Weindruch, R., Permana, P. A., Mountz, J. D. and Allison, D. B. (2004). Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research* **13**, 325–338.

Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. and Kocher, J.-P. (2013). Calculating sample size estimates for RNA sequencing data. *Journal of Computational Biology* **20**, 970–978.

Katz, T. A., Liao, S. G., Palmieri, V. J., Dearth, R. K., Pathiraja, T. N., Huo, Z., Shaw, P., Small, S., Davidson, N. E., Peters, D. G., *and others*. (2015). Targeted DNA methylation screen in the mouse mammary genome reveals a parity-induced hypermethylation of igf1r that persists long after parturition. *Cancer Prevention Research* **8**, 1000–1009.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research* **12**, 996–1006.

Kushwaha, G., Dozmorov, M., Wren, J. D., Qiu, J., Shi, H. and Xu, D. (2016). Hypomethylation coordinates antagonistically with hypermethylation in cancer development: a case study of leukemia. *Human Genomics* **10**, 18.

Li, E., Beard, C. and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365.

Licht, J. D. (2015). DNA methylation inhibitors in cancer therapy: the immunity dimension. *Cell* **162**, 938–939.

Markitsis, A. and Lai, Y. (2010). A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics* **26**, 640–646.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S. and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research* **33**, 5868–5877.

Park, Y., Figueroa, M. E., Rozek, L. S. and Sartor, M. A. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics* **30**, 2414–2422.

Park, Y. and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32**, 1446–1453.

Paulsen, M. and Ferguson-Smith, A. C. (2001). DNA methylation in genomic imprinting, development, and disease. *The Journal of Pathology* **195**, 97–110.

Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics* **6**, 597–610.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T. *and others*. (2006). Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Research* **34**, 528–542.

TSAI, P.-C. AND BELL, J. T. (2015). Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International Journal of Epidemiology* **44**, 1429–1441.

WU, H., WANG, C. AND WU, Z. (2015). PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* **31**, 233–241.

[*Received August 6, 2018; revised February 20, 2019; accepted for publication March 4, 2019*]