POWER CALCULATION AND STUDY DESIGN IN RNA-SEQ AND METHYL-SEQ

by

Chien-Wei Lin

MS., National Chiao Tung University, Taiwan, 2007

BS., National Tsing Hua University, Taiwan, 2005

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Chien-Wei Lin

It was defended on

April 14th 2017

and approved by

George C. Tseng, ScD

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Yong Seok Park, PhD

Assistant Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Robert Krafty, PhD

Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Daniel E. Weeks, PhD

Professor

Department of Human Genetics

Graduate School of Public Health

University of Pittsburgh

Dissertation Director: **George C. Tseng, ScD**

Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

POWER CALCULATION AND STUDY DESIGN IN RNA-SEQ AND METHYL-SEQ

Chien-Wei Lin, PhD

University of Pittsburgh, 2017

ABSTRACT

Next generation sequencing (NGS) technology has emerged as a powerful tool in characterizing genomic profiles. Among several applications, RNA sequencing (RNA-Seq) and Methylation sequencing (Methyl-Seq) have gradually become standard tools for transcriptomic and epigenetic monitoring respectively. Although the costs of NGS experiments have constantly decreased, high sequencing cost and bioinformatic complexity remain obstacles for many biomedical projects. Unlike earlier microarray technologies, modeling of NGS data should consider discrete count data. In addition to sample size, sequencing depth is also directly related to experimental costs. Consequently, given a total budget and a pre-specified unit experimental cost, the study design issue in RNA-Seq/Methyl-Seq is a multi-dimensional constrained optimization problem rather than a one-dimensional sample size calculation in a traditional hypothesis setting. In the first part of this dissertation, we proposed a statistical framework, namely "RNASeqDesign", to utilize pilot data for power calculation and study design of RNA-Seq experiments. The approach was based on a mixture model fitting of the p-value distribution from pilot data and a parametric bootstrap procedure to infer genome-wide power for optimal sample size and sequencing depth. We further illustrated five practical study design tasks for practitioners. We performed simulations and real data applications to evaluate performance and compare to existing methods.

In the second part, we proposed another statistical framework, namely "MethylSeqDesign", specifically for Methyl-Seq data. There were mainly two challenges. Firstly, the statis-

tical modeling for Methyl-Seq data required a powerful statistical test using beta-binomial model for conducting power calculation. Secondly, there is an extremely high number of CpG sites (about 30M) in the human genome, which results in many CpG sites with very shallow coverage. Hence, we focused on a region-/capture-based method which produced more counts in a region/window such that power calculation became feasible.

Public health significance: As sequencing costs keep dropping, RNA-Seq and Methyl-Seq experiments will become more prevalent and more projects with large sample size will be expected. We believe our work will provide practical guidance for future study design to understand disease mechanism and improve disease diagnosis and treatment.

Keywords: Power calculation, Sample size calculation, Sequencing depth, RNA-Seq data, Methyl-Seq data, Next Generation Sequencing (NGS), p-value mixture model, Parametric bootstrap.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Microarray technology has gained tremendous popularity in genomic research for its high-throughput quantitative representation and cost-effectiveness in the last decades (Reimers, 2010). While microarray experiments provide access for biologists to a range of applications, including the development of new diagnostic tools, discovery of novel disease subtypes, and identification of underlying mechanisms of disease or drug response, statistical analysis has played an active and significant role in the whole process. Statisticians correspondingly, have taken a enthusiastic interest in developing statistical tools that could lead to more profound biological interpretation to a certain research question (Slonim and Yanai, 2009; Kerr and Churchill, 2007). Next-generation sequencing, based on randomly amplifying and shotgun sequencing, is another revolutionary technology first marketed in 2004, making genomic profiles available in much higher resolution and in extremely high parallel (Fang and Cui, 2011). Although error and biases might be introduced in major steps of the experimental preparation process, next-generation sequencing has been hailed as the future of genetic research since it provides higher sensitivity than microarrays and could potentially generate an unlimited dynamic range. It is generally expected that research will gradually shift from microarray technologies to next-generation sequencing (Shendure, 2008). From statistical point of view, many methodologies developed under the microarray context could still be extended to NGS, while we will also face new challenges in data analysis.

In a biological study, the procedure of exploring a research topic usually starts from the study design, where a major component is sample size and power calculation. The purpose of such careful design is obvious: to improve efficiency and reduce cost. Methods for power and sample size calculation in clinical and microarray data are rich in the field (Lee and Whitmore, 2002; Gadbury et al., 2004), but it is still limited in methods developed

specifically for sequencing data. As sequencing technology is still not quite affordable to the majority of researchers, it is significant to ensure desirable power of biomarker detection (a.k.a. differential expression (DE) analysis) from the earlier phase of study (herein called "pilot study").

In this introduction, we will first go over the significance of gene expression and DNA methylation quantification with both traditional and advanced technology. Then, we will introduce the structure of gene expression data and DNA methylation data, and review methods of differential analysis for each data type. Furthermore, we will distinguish the difference between traditional power and genome-wide power definitions, and review some existing methods for microarray and RNA-Seq data power calculation (to the best of our knowledge, there is no existing statistical method for Methyl-Seq data power calculation). Finally, the major motivation of developing our methods will be addressed.

1.1 QUANTIFICATION OF GENE EXPRESSION AND DNA METHYLATION LEVEL

Gene expression, which is the procedure of mRNA synthesis from a set of genes, has been extensively used in the characterization of human disease, identification of novel disease subtypes, and potential drug target for treatment. Understanding the dynamic changes of gene expression of a given subject is important for us to study biological processes ranging from inflammation to human aging.

DNA methylation is a process in which methyl group attaches to the cytosine followed by a guanine on the DNA sequence, known as CpG sites. In the genome, there are certain regions enriched with these spots, e.g., CpG islands. Many of those regions are related to gene regulatory regions. It is well known that DNA methylation alters the gene expression level, typically repressing it. This process has been found to be involved in many important biological systems, including genomic imprinting, X-chromosome inactivation, repression of repetitive elements, aging and carcinogenesis (Li et al., 1993; Paulsen and Ferguson-Smith, 2001; Robertson, 2005).

By comparing gene expression/methylation data between different groups of subjects, we can explain if any of the biological pathways were altered by certain disease mechanisms. Hence, the quantification of the high-throughput data plays an important role. Here we review traditional and advanced high-throughput technology.

Advances in molecular and computational biology have led to the development of powerful, high-throughput methods for the analysis of differential gene expression. These tools have opened up new opportunities in disciplines ranging from cell and developmental biology to drug development and pharmacogenomics.

1.1.1 Microarray - Hybridization based approaches

With the increased popularity of high throughput technology in mid 90's, microarrays became the prominent tool for quantifying genomic changes. The ability of these arrays to simultaneously interrogate thousands of transcripts has led to important advances in a wide range of biological problems, including the identification of gene expression differences among diseased and healthy tissues, and new insights into developmental processes, pharmacogenomic responses, and the evolution of gene regulation. The principle of a microarray experiment is that mRNA from a given tissue is used to generate a labelled target, which is then hybridized in parallel to a large number of DNA sequences, immobilized on a solid surface in an ordered array (Schulze and Downward, 2001). The data generated from microarray experiment typically consist of a long list of measurements for spot intensities or intensity ratios. Nonetheless, it suffers from the following limitations: (1) background noise from hybridization limits the measurement of expression, especially for probes with low abundance; (2) heterogeneity of probes with respect to their hybridization properties will reduce the accuracy of measurements; (3) the assay is limited to transcripts with known probes (Marioni et al., 2008).

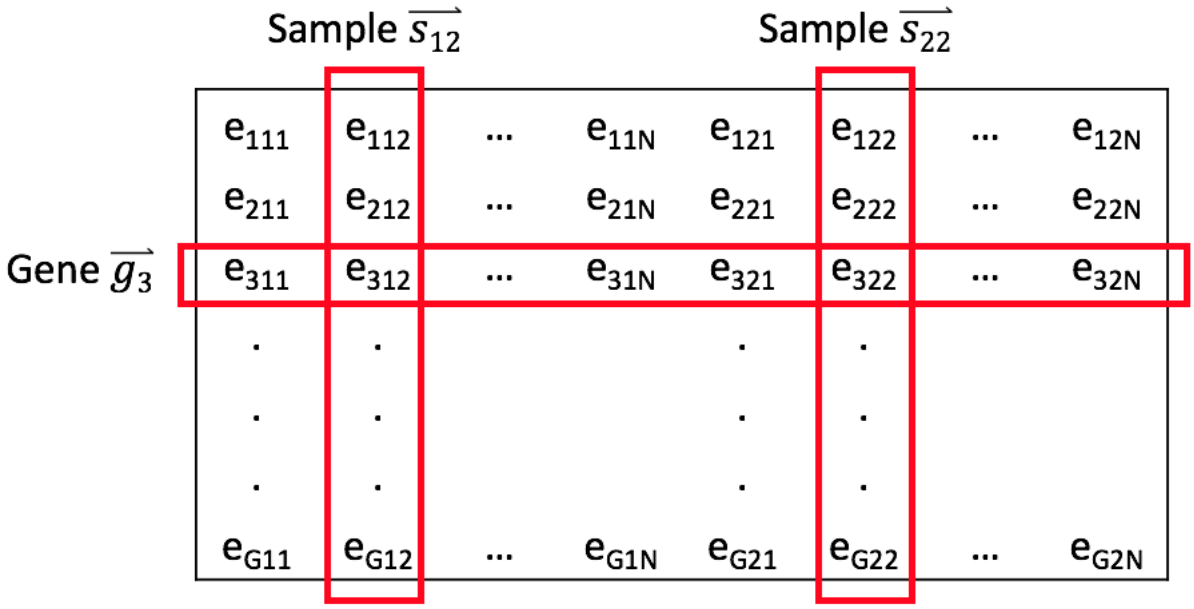
1.1.2 Next generation sequencing (NGS)

In the field, the traditional sequencing technique, Sanger sequencing, was the most widely used method for more than 30 years since it was developed in 1977. In 2001, the Human

Genome Project sequenced the human genome and it greatly motivated researchers to explore human genetic mechanisms from a sequencing perspective. However, the method is not only expensive and labor intensive, but also only able to target several specified genes with known primer sequences. Thus researchers have been eager for high-throughput techniques. Next generation sequencing utilizes high-throughput DNA sequencing techniques: DNA sequences are smashed into fragments and sequenced in parallel, generating multiple reads of each fragment and yielding substantial throughput. Alignment algorithms assemble these short reads to the reference genome. By reconstructing the whole genome, we are able to know the exact nucleotide order present in DNA and the coverage of segment at every position. Therefore a wide variety of genomic features can be measured. Through deep sequencing, it is possible to detect SNP/indel, structural variation and somatic mutations on a genome-wide scale. Through coverage, we are able to detect copy number variation and mRNA expression. By some extra bisulfite treatment technique, sequencing can also measure methylation. In addition, novel genomic features such as isoforms of mRNA and fusion genes can be detected. Nowadays, millions of fragments of DNA from a single sample can be sequenced in parallel and the entire genome can be sequenced within one day. This technique has dramatically accelerated our understanding of the human genome.

1.2 DATA STRUCTURE OF MICROARRAY, RNA-SEQ AND METHYL-SEQ EXPERIMENT

A genomic study typically assesses a large number of DNA sequences (or genetic features) under multiple conditions, e.g., a collection of different tissue samples. For transcriptomic applications, the output data from the experiment after proper preprocessing (including normalization, transformation...etc) is a gene expression matrix $M = \{e_{gij} | 1 \leq g \leq G, 1 \leq i \leq k, 1 \leq j \leq n_k\}$, where the rows ($G = \{\overrightarrow{g_1}, ..., \overrightarrow{g_G}\}$) form the expression patterns of genes, the columns ($S = \{\overrightarrow{s_{11}}, ..., \overrightarrow{s_{1n_1}}, ..., \overrightarrow{s_{kn_k}}\}$) represent the expression profiles of $\sum_{i=1}^{k} n_i = n$ samples, and each cell e_{gij} is the measured expression level of gene g in sample j of group i. For illustration, we first assume the genomic study is a balanced design but the assumption can



An example of genomic data with genes in rows and samples in columns. Balanced two group design ($n_1 = n_2 = N$). Total number of genes is G and total number of samples is n.

Figure 1 An example of genomic data

be relaxed later. Figure 1 illustrates the case where there are two groups of interest ($k = 2$), and each of them has N ($n_1 = n_2 = N$) samples. In other words, subject $\{\overrightarrow{s_{11}}, ..., \overrightarrow{s_{1N}}\}$ are in group 1 with class label $x_j = 0$ ($j = 1, ..., N$), while subject $\{\overrightarrow{s_{21}}, ..., \overrightarrow{s_{2N}}\}$ are in group 2 with class label $x_j = 1$ ($j = 1, ..., N$).

In a transcriptomic microarray study, e_{gij} is typically either log2 of raw intensity or intensity ratio (in a two-colors design) of gene g in subject j of group i, which is a continuous variable. In a DNA methylation microarray study, for each methylation site g in subject j of group i, we have methylated M_{gij} and unmethylated U_{gij} intensities. Hence, e_{gij} is the methylated proportion of methylation site g, one can use either Beta-value ($= M_{gij}/(M_{gij} + U_{gij})$) or M-value ($=\log_2(M_{gij}/U_{gij})$) (Du et al., 2010). For RNA-Seq data, e_{gij} is the read counts of gene g in subject j of group i. For Methyl-Seq data, for each CpG site g in subject j of group i, we have M_{gij} and U_{gij} for methylated read counts and unmethylated read counts respectively. The variable of interest is again the methylated proportion using either Beta-value or M-value.

For both technologies, sample size is the most influential factor in determining power for detecting differentially expressed genes (DE genes) or differentially methylated sites. For NGS data, read/sequencing depth, which is proportional to total reads (R), is another important factor that impacts power (Rapaport et al., 2013). Higher sequencing depth generates more reads mapped to the same chromosome locations, which will give higher base quality and thus increase statistical power to detect DE genes (Sims et al., 2014). Throughout this dissertation, we refer to (sequencing) depth and coverage inter-changeably since both meaning how many reads are assigned to a particular genomic location.

1.3 BIOMARKER DETECTION IN MICROARRAY AND NGS DATA

1.3.1 RNA

Cui and Churchill (2003) provided a comprehensive review for the popular methods for statistical tests and issues that are addressed in microarray gene expression data. Perelman et al. (2007) compared several alternative methods including t-test, modification of t-test (significance analysis model, SAM) for differential expression analysis. Smyth (2004) proposed a method called "Limma" applying an empirical Bayes approach that adopts a global variance estimator computed on the basis of all genes' variances to stabilize the variance of each individual gene. These methods are all based on the Gaussian assumption for log2 transformed gene expression.

Due to the different characteristics of sequencing data, the statistical methods for detection of DE genes are more complicated and diverse according to different assumptions and applications. A number of proposals have been made for identifying differentially expressed genes from RNA-Seq data in the case of a two groups comparison. The methods can be summarized into three major categories:

(1) Method based on Gaussian assumptions: Bloom et al. (2009) applied t-tests to the total-count normalized data. AC't Hoen et al. (2008) performed a square-root transformation for the total-count normalized data to stabilize the variance and applied t-tests afterwards.

In DEGseq (Wang et al., 2009), the authors assumed that log ratios of the counts between two different samples follow an approximate normal distribution and the p-value was derived based on the conditional normal distribution.

(2) Methods based on Poisson assumptions: Marioni et al. (2008) proposed a Poisson log-linear model and performed a likelihood ratio test (LRT) for differential expression gene detection. Normalization based on total-counts were performed implicitly. Bullard et al. (2009) applied an external quantile normalization step rather than doing total-count normalization. "Poissonseq" (Li et al., 2012) is based on a Poisson log-linear model, and can be applied to not only two-class outcome but also multiple-class and even quantitative outcome.

(3) Methods based on negative binomial assumptions: Generalized linear model (GLM) based on a negative binomial distribution has been applied in order to handle overdispersed counts in RNA-Seq data. Robinson et al. (2009) developed edgeR by extending from previous methods for SAGE data (Robinson and Smyth, 2008). In their method, the dispersion parameters can be estimated for each gene or can be common across all genes. An empirical Bayes method is applied to shrink the dispersion toward a common value by borrowing information across multiple genes (same idea in Limma (Smyth, 2004)). P-values are derived using exact test. DESeq (Anders and Huber, 2010), is another method that imposes a negative binomial assumption and uses local regression to estimate the relationship between the variance and the mean. baySeq (Hardcastle and Kelly, 2010) applies empirical Bayesian approach theory to estimate the posterior probabilities of each of a set of models that define patterns of differential expression. NOISeq (Tarazona et al., 2011) is a nonparametric and data-adaptive method. To remove the dependency on sequencing depth, it models the noise distribution from the actual data. Therefore, it can better adapt to the size of the dataset compared to other methods.

Comparative studies (Rapaport et al., 2013) have indicated that no single method appears to be favorable in all settings but methods based on negative binomial assumption (e.g., DESeq, edgeR, and baySeq) have superior specificity and sensitivities as well as good control of false positive errors. Furthermore, Nookaew et al. (2012) found that edgeR could uniquely identify more differential gene expression (DGE) than Cuffdiff, baySeq, DESeq and NOISeq.

1.3.2 DNA methylation

For DE analysis of DNA methylation data, we want to identify differentially methylated loci (DML) or regions (DMRs) that show different methylation levels across distinct groups, e.g., cases and controls. Robinson et al. (2014) gives a comprehensive review of existing DE analysis methods for microarray and sequencing platforms.

For microarray platforms, the data used for the analysis is either Beta-value (methylated proportion) or M-value (log2 ratio of methylated to unmethylated intensity). Du et al. (2010) suggests using M-values because the transformation of the data allows directly applying existing methods for gene expression continuous data, like Limma (Smyth, 2004). Many methods have been developed for both upstreaming (preprocessing) and downstream (DE) analysis (Price et al., 2013; Aryee et al., 2014). Due to the complexity of the methylation data, these types of tools often use simple t-test to do the DE analysis. Wang et al. (2012) develops an R package "IMA" which applies Wilcoxon rank sum test on Beta-values. Robinson et al. (2014) concludes that moderated t/F-statistics on the normalized log-ratios of intensities perform well in microarray data.

For sequencing platforms (here we refer to Bisulfite sequencing, BS-Seq/Methyl-Seq), there are different methods according to different distributional assumptions. Hansen et al. (2012) applied t-test for the DE analysis and many other methods based on generalized linear model (GLM) have been proposed (Akalin et al., 2012; Dolzhenko and Smith, 2014; Feng et al., 2014; Park et al., 2014). Akalin et al. (2012) uses binomial GLM (essentially logistic regression), and Dolzhenko and Smith (2014); Feng et al. (2014); Park et al. (2014) use beta-binomial GLM to better account for both biological and sampling/technical variation. These methods suffer large computational burdens because the estimation procedures rely on iterative steps to maximize the likelihood function. Park and Wu (2016) proposes a novel method based on a beta-binomial GLM with an arcsine link function. The estimation procedure is based on generalized least square approach without iterative steps, which helps reduce the computational demands dramatically.

1.4 SAMPLE SIZE, POWER, GENOME-WIDE POWER

One of the most common tasks for statisticians requested by investigators is to perform sample size and power calculations. In general, sample size is the number of subjects under certain conditions enrolled in a study, e.g., control and patients. It is also usually referred to as biological replicates. Power is referred to as the statistical power/sensitivity of the test rejecting the null hypothesis when the alternative hypothesis is true. In general, increasing the sample size will certainly help increase the power. However, larger sample size often comes along with practical issues (increasing cost, limited resources, ...etc). Hence, the balance between sample size and power that investigators want to achieve has to be taken into consideration as early as possible.

In order to calculate the sample size needed, it is required to have some prior knowledge or expected conditions in a study. For example, we need to decide the desired power under certain effect size (the difference between different groups) and significance level for the hypothesis test. The greater the variability in the data, the larger the sample size that is required to assess whether or not an observed effect is a true effect. For example, larger sample size is usually needed in human studies than animal models. On the other hand, if the tested treatment/comparison is more effective (larger effect size), then a smaller sample size is needed to detect this positive or negative effect (Noordzij et al., 2010).

Traditionally, the definition of power often refers to the framework based on one hypothesis test. That is, when one only tests for one biomarker/treatment. For example, assume we are interested in testing $H_0 : \mu_A - \mu_B < 2$ against $H_1 : \mu_A - \mu_B \geq 2$, where μ_A and μ_B are different group means for group A and B, both of which have same number of subjects. To achieve a statistical power of $1-\beta$ under significance level α, sample size can be calculated as $n = \frac{(s^2_{\bar{Y}_A - \bar{Y}_B})(z_\alpha + z_\beta)^2}{(\bar{Y}_A - \bar{Y}_B)^2}$, where z_α is the critical value of the standard normal distribution with tail area of α. $\bar{Y}_A - \bar{Y}_B$ refers to effect size (which equals 2 in this example), indicating the difference between two groups of interest, and $s^2_{\bar{Y}_A - \bar{Y}_B}$ is the variability of group difference, which usually gets smaller as the sample size increases.

However, in genomic applications, gene expression matrices from transcriptomic studies usually constitute of more than 20,000 probes or genes, and for methylation matrices from

epigenetic studies, the number of CpG sites could increase to more than 450,000 or larger. Given this multiple hypothesis testing setting, we can quantify the power of detecting multiple genomic changes in a whole-genome scale, by the concept of "genome-wide power". The key question directly from it is how we can maintain type I error while we control the power, since there are multiple comparisons. The family-wise error rate and false discovery rate (FDR) (Benjamini and Hochberg, 1995) are widely used to address this kind of problem. Gadbury et al. (2004) proposed the concept of expected discovery rate (EDR) as average power across multiple comparisons to quantify the "genome-wide power." We describe these concepts below.

Based on the multiple testing framework in Benjamini and Hochberg (1995), we have the two by two contingency table with G hypotheses in total to perform (see Table 1). The numbers G_0 and G_1 of false and true null hypotheses are unknown parameters, A and R are observable random variables and A_0, A_1, R_0, R_1 are unobservable random variables. In the context of genomic studies, we would like to minimize the number R_0 of false positives (Efron, 2007; Ge et al., 2009). Following Gadbury et al. (2004), genome-wide power is defined as $EDR = E(\frac{R_1}{G_1})$, and FDR is defined as $FDR = E(\frac{R_0}{R})$. In most genomic applications, one controls type I error by FDR under a certain pre-specified threshold (e.g., FDR=0.05) to obtain the DE gene list. In the power calculations throughout this dissertation, we refer to EDR as genome-wide power and pursue it under a pre-specified FDR control level.

Table 1 Multiple testing framework

True hypothesis	Test declaration:		Number of genes
	non-DE	DE	
non-DE H_0	A_0	R_0	G_0
DE H_1	A_1	R_1	G_1
Total	A	R	G

On the row is the unknown underlying true status for each marker/gene. On the column is the testing results from a certain statistical test.

1.5 EXISTING SAMPLE SIZE AND POWER CALCULATION METHODS

1.5.1 The use of pilot study in power calculation

In general, power calculation methods can be purely model-based, that is, pre-specified the parameters such as effect sizes, variances, ...etc, or one can estimate those information from a pilot study (usually of a relatively small sample size). Model-based methods are straightforward and more economical, however, the downside is that assumptions may easily fail in the real data. Hence, by conducting pilot studies, we can reasonably estimate the variability and effects from the pilot data and further infer the proper sample size or power. There are many different types of variability coming from, for example, biological replicates, technical replicates, experimental and batch effects. Especially for genomic applications, pilot studies are of greater importance. Therefore, the key question is how to take advantage of pilot data in the sample size and power calculations under a multiple testing framework.

In this dissertation, we assume that a pilot study with sample size N (in each group) and total reads R (for each subject) is available to tackle the problem of power prediction. We will also discuss the potential approach of power calculation when there is no pilot study in the Discussion. In the following sections, we will briefly review the existing methods for transcriptomic data for microarray and NGS platforms. To the best of our knowledge, there are no existing methods for power calculation for Methyl-Seq data.

1.5.2 Existing methods for RNA microarray data sample size calculation

The significance of performing power and sample size calculations for genomic data was first addressed by Lee and Whitmore (2002). They started from a common setting of ANOVA model in microarray data analysis:

$$Y_b = \gamma_0 + \gamma_1(b_1) + ... + \gamma_L(b_L) + \sum_{\ell=1}^{L}\sum_{k>\ell}^{L}\gamma_{\ell k}(b_\ell, b_k) + ... + \epsilon_b, \tag{1.1}$$

where $\ell = 1, ..., L$ denotes a set of L experimental factors. Parameter $\gamma_\ell(b_\ell)$ denotes a main effect for factor ℓ when it has level b_ℓ, for $\ell = 1, ..., L$, respectively. Similarly, parameters $\gamma_{\ell k}(b_\ell, b_k)$ denote pairwise interaction terms for factors ℓ and k when they have

their respective levels b_ℓ and b_k, with ℓ, $k = 1, ..., L$. The error term is denoted by ϵ_b. The index b is a vector of the form $(b_1, ..., b_L)$ where b_ℓ denotes the level of factor ℓ. Let $I_g = (I_{gc}, c = 1, ..., C)^t$ denote the vector of parameters for gene g, e.g., I_{gc} is denoted as the effect of a covariate/condition c for gene g. Suppose I^d is a non-zero vector representing differential expression levels of gene g that is expected to detect. When testing $H_0 : I_g = 0$ (gene g is not differentially expressed) against $H_1 : I_g = I^d$ (gene g is differentially expressed), F-statistics, χ^2 or z-statistics could be constructed for single gene under different study designs. In their paper, ways to control multiple comparisons were discussed when genes are correlated and not correlated. Microarray studies usually involve simultaneous tests of thousands of genes. Therefore the probability of producing incorrect conclusions must be controlled.

Family-wise error rate (FWER) $\alpha_F = P(R_0 > 0)$ (where R_0 is the false positives in Table 1), is discussed in details for application in multiple comparison issues in Lee and Whitmore (2002). Both (1) Sidak approach: assuming independent estimation errors; and (2) Bonferroni procedure: assuming dependent estimation errors are considered. Notice that this approach does not consider the heterogeneity across different genes. They simply assumed all genes have the same variance and effect sizes for alternative hypothesis. In addition to the ANOVA model, the authors also mentioned the possibility solving the power calculation problem from a Bayesian perspective, where a mixture model is introduced as:

$$f(v) = p_0 f_0(v) + p_1 f_1(v), \tag{1.2}$$

where p_0 is the proportion of non-DE gene, and $p_1 = 1 - p_0$. Here v is the summary statistics for each gene, $f_0(v)$ is the density for non-DE component, and $f_1(v)$ is the density for DE component. This approach was not investigated enough until the methodological work of PowerAtlas (Gadbury et al., 2004).

PowerAtlas is a popular web tool for power and sample size calculation proposed by Page et al. (2006). They considered the variability of mean expressions and effect sizes across different genes by directly modeling the p-value distribution of all test statistics. They introduced the concept of expected discovery rate (EDR) as we reviewed in the previous section, which can be viewed as the average power across all genes with true effects. Since

genomic studies inherit multiple comparison issues, they also defined true positive rate (TP) and true negative rate (TN) as below (the notations are consistent with Table 1).

$$EDR = E(R_1/G_1) \tag{1.3}$$

$$TP = E(R_1/R) \tag{1.4}$$

$$TN = E(A_0/A) \tag{1.5}$$

If μ_{iA} and μ_{iB} are the underlying true expression for gene i in group A and B respectively, we want to test whether the expression of group A and B are different with $H_0 : \mu_{iA} - \mu_{iB} = 0$ against $H_0 : \mu_{iA} - \mu_{iB} \neq 0$. Given a set of pilot data, their procedures started from a set of p-values of t-statistics for testing differential expression for each gene. This pilot dataset is expected to represent similar experimental characteristics as the future/larger data. It could either be generated in a pilot study or directly from a public database. The t-statistic of gene i can be written in the following form:

$$t_i = \frac{\bar{e}_{iA} - \bar{e}_{iB}}{S_{e_{i0}x_{i1}}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \tag{1.6}$$

where $S_{e_{iA}e_{iB}} = \frac{(n_A-1)S^2_{e_{iA}}+(n_B-1)S^2_{e_{iB}}}{n_A+n_B-2}$, assuming equal variance. When the two groups have equal sample size $n_A = n_B = N$, the t-statistic becomes:

$$t_i = \frac{\bar{e}_{iA} - \bar{e}_{iB}}{\sqrt{(S^2_{e_{iA}} + S^2_{e_{iB}})/N}} \tag{1.7}$$

With the assumption that p-value distribution from DE analysis from microarray experiment is a mixture of a beta distribution component (for DE genes) and an uniform distribution component (for non-DE genes), a Beta-Uniform mixture (BUM) model is fitted with p-values from $t_i (i = 1, ..., G)$. The fitted model is as 1.2. Then a parametric bootstrap procedure is performed to obtain an updated set of p-values according to a targeted sample size N'. The key step of this method is the transformation of t-statistics by:

$$t_i^* = t_i\sqrt{N'/N} \tag{1.8}$$

The underlying assumption is that the group-wise mean ($\bar{e}_{iA}$, $\bar{e}_{iB}$) and group-wise variance ($S^2_{e_{iA}}$, $S^2_{e_{iB}}$) of each gene remain the same under different sample sizes. By directly modeling the p-value distribution, the heterogeneity across genes can be maintained. However, this method cannot be directly applied to RNA-Seq and Methyl-Seq data because different characteristics of the data. They also did not control FDR at a fixed level, but instead they impose an arbitrary p-value cut-off.

1.5.3 Existing methods for RNA-Seq data sample size calculation

One of the biggest distinction between RNA-Seq and microarray gene expression data is the types of expression values. Microarray data produces continuous intensities, while RNA-Seq data produces read counts for each gene. Hence, the distributional assumptions differ: Gaussian assumption for microarray (usually after log2 transformation), and Poisson/negative binomial distribution for RNA-Seq data. Many methods have been proposed for RNA-Seq data under different distributional assumptions.

1.5.3.1 Methods based on Poisson assumptions Many literature have discussed various Poisson tests: (1) asymptotic test based on normal approximation: (a) unconstrained maximum likelihood estimate (MLE); (b) constrained maximum likelihood estimate (CMLE); (2) tests based on approximate p-value methods; (3) exact conditional test and mid-p conditional test; (4) likelihood ratio test. See Gu et al. (2008) for a comprehensive review for Poisson rate tests. Li et al. (2013a) developed methods (we call it "Poisson model" for later reference) for sample size and power calculation, based on different types of Poisson tests. They used false discovery rate (FDR) for multiple comparisons (Storey and Tibshirani, 2001; Storey, 2002).

1.5.3.2 Methods based on negative binomial assumptions Poisson tests are widely used, while they ignore the nature of over-dispersion in real sequencing data. We have reviewed methods to detect DE genes based on an over-dispersed Poisson model. Among them, edgeR (Robinson et al., 2009) and DEseq (Anders and Huber, 2010) have been two

most popular methods to perform DE analysis based on exact tests. Extensive comparative studies have shown the superiority of these two tests in detecting biomarkers over other tests. However, it is clear that the two exact tests don't have a closed form for sample size and power calculation.

Until now, there are two methods proposed for RNA-Seq power calculation based on negative binomial distributional assumption: (1) RNASeqPower (Hart et al., 2013); (2) method based on the exact test (Li et al., 2013b). The two methods are similar in that they both require the estimation or pre-specification of fold changes, mean counts, coefficient of variations and the dispersion parameter.

RNASeqPower has a basic formula:

$$n = 2(z_{1-\frac{\alpha}{2}} + z_{\beta})\frac{1/\mu + \sigma^2}{ln(\Delta^2)}, \tag{1.9}$$

where α and β are type I error and power respectively; z_x is the x quantile of standard normal; and Δ is the fold change or effect size. μ and σ are read coverage and coefficient of variation (CV) between biological replicates (gene specific). The derivation of this formula is based on a generalized linear model framework. CV is estimated by edgeR ($\sigma = \frac{1}{\sqrt{\delta}}$, where δ is the dispersion parameter). μ, σ and δ are required to be fixed across all genes for a given study, and are often determined by external requirements. Apparently, this method is designed for single gene based hypothesis testing. However, the authors suggest when considering multiple genes scenario, one can simply take $\sigma_{0.60}$ (60% quantiles of CV as the overall CV) and the quantile of depth distribution across gene for sample size calculation. An R package "RNASeqPower" is available through the bioconductor website. Although this method is straightforward, it comes with several disadvantages: (1) it does not consider multiple comparisons issue since the the power is only computed based on one single hypothesis test; (2) it fails to incorporate the variability across genes and instead uses summary statistics for effect size, dispersion, coverage of each gene, etc.

Li et al. (2013b) proposed a method for power calculation based on the exact test. Instead of deriving the distribution of test statistic under the alternative hypothesis, the authors adopt the method proposed by Krishnamoorthy and Thomson (2004) to calculate the power for the exact test based on a given p-value. Following the same quantile-adjusted conditional

maximum likelihood procedure in Robinson and Smyth (2008), pseudo/normalized-counts are generated to conduct the exact test. For a given p-value $p(e_A, e_B)$ of a gene, where e_A and e_B are the observed sum of pseudo-counts, the power could be calculated by solving following equation:

$$\xi(N, \rho, \mu_A, \delta, \omega, \alpha) = \sum_{e_A=0}^{\infty} \sum_{e_B=0}^{\infty} f(N\omega\rho\mu_A, \frac{\delta}{N}) f(N\mu_A, \frac{\delta}{N}) I(p(e_A, e_B) < \alpha) = 1 - \beta, \quad (1.10)$$

where N is the sample size in group A, $\omega = \frac{d_1^*}{d_0^*}$ is the ratio of the geometric means of normalization factors between group A and B, ρ is the fold change, μ_A is the average read counts in group A and $f(\mu, \delta)$ is the probability mass function of negative binomial model with mean μ and dispersion δ. α is the significance level and $I(\cdot)$ is the indicator function. The required sample size N to attain the given power $1 - \beta$ at level of significance α can then be derived from numerically solving equation 1.10 by using gradient descent or bisection algorithm.

Considering the practical case of multiple genes, the authors provided two approaches. In the first approach, μ_{iA}, ρ_i, δ_i can be estimated from pilot data for each prognostic gene i which is known. Then we could use the numerical method to solve the equation to derive required sample size:

$$r_1 = \sum_{i \in M_1} \xi(N, \rho_i, \mu_{iA}, \delta_i, \omega, \alpha^*), \quad (1.11)$$

where r_1 is the expected number of true rejections, M_1 is the set of prognostic genes, and $\alpha^* r_1 f/(m_0\,(1-f))$ is the type I error when FDR is controlled at f, where m_0 is the number of null genes. In the second approach, when the parameters of prognostic genes are unknown (which is usually the case), we can specify a desired minimum fold change ρ^*, a minimum average read count μ_{iA} and a minimum dispersion δ_i. By replacing $\alpha^* = r_1 f/(m_0\,(1-f))$ and $\beta^* = 1 - r_1/m_1$ in equation 1.10, one can derive the required sample size in the case of multiple gene comparisons. Although this method provides a way to account for the heterogeneity across different genes, the parameter setting is still arbitrary and relying on information which is hard to retrieve.

Wu et al. (2015) proposed a method called "PROPER" for sample size and power calculation, which is a fully simulation-based method. Users have to decide every parameter

to run the procedure, including mean counts distribution, effect size distribution, dispersion parameter distribution, etc. This method suffers large computation needs and requires accurate parameter specification.

1.5.4 Methods based on Gaussian assumptions

"Scotty" (Busby et al., 2013) is an online-tool developed for interactive power calculation. Given a pilot data, it first assesses the mean counts distribution and the effect size distribution. Secondly, it builds up a matrix of combination of parameters to represent different designs for hypothesis testings. Finally, the design with highest power under a user-specified parameter will be selected. The input parameters include the number of biological replicates, read depth and cost. While Scotty provides novel ways in the study design for RNA-Seq experiment, the framework is established based on the Gaussian assumption and the statistical power is calculated based on non-central t-tests. The authors argued that by using t-tests unbiased calls of differential expression will be produced and the closed-form formula for calculating power based on non-central t-test can be easily derived. However, to justify this statement or the performance of t-test, the authors compared t-test to DESeq (Anders and Huber, 2010) using limited simulated data. They concluded that when sample size is small (N=2), DESeq has higher power in detecting DE genes, while sample size increases to greater than 5, t-test has slightly greater power in detection. However, the paper did not evaluate the false positive of tests, the accuracy of power prediction, and selection of optimal experiment configuration since they did not consider to generate a true power curve to compare with their simulation results. Furthermore, they did not take into consideration of FDR but instead using an arbitrary p-value cut-off to declare DE genes.

1.6 OVERVIEW

Inspired by the limitations of existing methods, here we proposed two novel sample size and power calculation methods: (1) RNASeqDesign for RNA-Seq data; (2) MethylSeqDesign for

Methyl-Seq data. The following chapters will be arranged as described below. In chapter 2, we will present the entire RNASeqDesign paper including comparative simulation and real data analysis. In chapter 3, we will present the second method, MethylSeqDesign, and show comprehensive simulation analysis and real data analysis.

2.0 RNASEQDESIGN: A FRAMEWORK FOR RNA-SEQ GENOME-WIDE POWER CALCULATION AND STUDY DESIGN ISSUES

This work has won 2016 ENAR student paper award and been submitted to JASA (under review).

2.1 INTRODUCTION

With the advent of next-generation sequencing (NGS) technology, RNA-Seq has been rapidly developed to characterize transcriptomic profiling, which is now impacting almost every field of life science (Ozsolak and Milos, 2011; Conesa et al., 2016). Compared to the once popular microarray technology, RNA-Seq has many advantages, such as higher per-base resolution, better reading accuracy, wider detection range, and ability to discover novel transcripts/isoforms. As the sequencing cost has constantly dropped, quantification of expression profiles by RNA-Seq experiment has become more feasible, which provides more accurate detection of differentially expressed (DE) genes. When designing RNA-Seq experiments, sample size calculation is critical because of the still high sequencing cost and limited budget.

Traditional power calculation considers relationships between four elements: effect size, α (type-I error), statistical power (1 type-II error (β)) and sample size. For example, for a given effect size (usually estimated from pilot or published data) and α (normally 5%), one is interested in calculating the sample size to reach a pre-specified statistical power (e.g. power=80%) or, equivalently, to estimate statistical power given certain sample size (e.g. $N = 50$). When analyzing high-throughput genome-wide experimental data, the situation

becomes more complicated because of the well-known multiple comparison consideration. Since thousands of hypotheses are tested simultaneously, controlling type-I error rate and reducing false discovery in a genome-wide sense becomes critical. As a result, conservative family-wise error rate (FWER) and the scientifically more applicable false discoverate rate (FDR; Benjamini and Hochberg (1995)) have been proposed in the literature to replace type-I error α. Lee and Whitmore (2002) first addressed the importance of power and sample size calculation for microarray data and provided a procedure based on ANOVA for controlling FWER. Since then, several other methods were proposed to control FWER for microarray power calculation (Jung et al., 2005; Dobbin and Simon, 2005; Jung and Young, 2012). In addition, Ferreira and Zwinderman (2006), Liu and Hwang (2007) and Van Iterson et al. (2009) incorporated the concept of FDR and utilized pilot data to account for genome-wide scenario for more realistic power calculation. Gadbury et al. (2004) introduced the concept of expected discovery rate (EDR; see definition in Section 2.1) to replace univariate power 1-β for addressing genome-wide detection power. They proposed a method combining parametric mixture modeling and parametric bootstrap to estimate the required sample size. However, their method only considered an arbitrary p-value cut-off to declare DE genes instead of controlling FDR. Conceptually, FDR is the genome-wide analogue of type-I error α from univariate hypothesis testing and EDR is the genome-wide analogue to statistical power $1-\beta$. Since genome-wide screeing considers the whole set of DE genes, specifying one single effect size for power calculation is not adequate and considering the effect size distribution of DE genes is biologically more reasonable. A good power calculation method for high-throughput experimental data should replace α and $1-\beta$ with FDR and EDR and consider the distribution of effect sizes among DE genes.

Compared to microarray power calculation, RNA-Seq data have three unique features that bring new statistical challenges and require novel study design concepts. Firstly, RNA-Seq analysis aligns randomly sequenced short reads to the transcribed regions of each gene and produces count data by nature. Continuous measurements from microarray intensities are often modeled with a normality assumption after log transformation. This normality assumption restricted direct extension of many methods from microarray to RNA-Seq experiments. RNA-Seq data need to be modeled with discrete distributions, and both sam-

pling and biological variation should be considered. For this reason, the negative binomial model (Robinson et al., 2009) has gained popularity compared to the Poisson distribution. Secondly, the genome-wide distribution of expression levels tends to be skewed with the majority of sequencing reads focused on a small portion of highly expressed genes (often housekeeping genes), leaving most genes with low counts. Transcripts with short lengths also tend to have low sequencing counts. A desirable power calculation method should accommodate the detection bias for low-expressed or short-length transcripts. Finally and most importantly, RNA-Seq experiments usually adopt a multiplex sequencing technique by adding "barcode" sequences to each sample so they can be distinguished and sorted in downstream data analysis. For example, the Illumina HiSeq 2500 platform runs eight lanes in each experiment (known as a flow cell) with a fixed cost. A researcher can choose to process one sample per lane, which results in roughly 250 million reads or three samples per lane each with 83 million reads. In other words, a researcher can choose to triple the sample size (denoted by N hereafter) by reducing the sequencing depth (denoted by R) to one-third for the same sequencing cost. The power calculation solution space increases from a classical one-dimensional (sample size, N) decision to a two-dimensional (N and R) optimization that leads to a new study design problem. Figure 2A illustrates a scenario for RNA-Seq power calculation and study design. Denote by $C = B(N, R)$ the pre-specified cost function based on selected sample size N and sequencing depth R, and $Pow(N, R)$ the estimated genome-wide power EDR. The problem can be formulated as searching the best combination of N^* and R^* to optimize genome-wide power Pow(N, R) under a given budget constraint $C = B(N, R)$. Figure 2B illustrates the importance of extending the four key elements of univariate power calculation (type-I error α, power 1-β, univariate effect size, sample size N) to RNA-Seq genome-wide power calculation using FDR, EDR, effect size distribution among DE genes, and two dimensional sample size N and sequencing depth R, respectively.

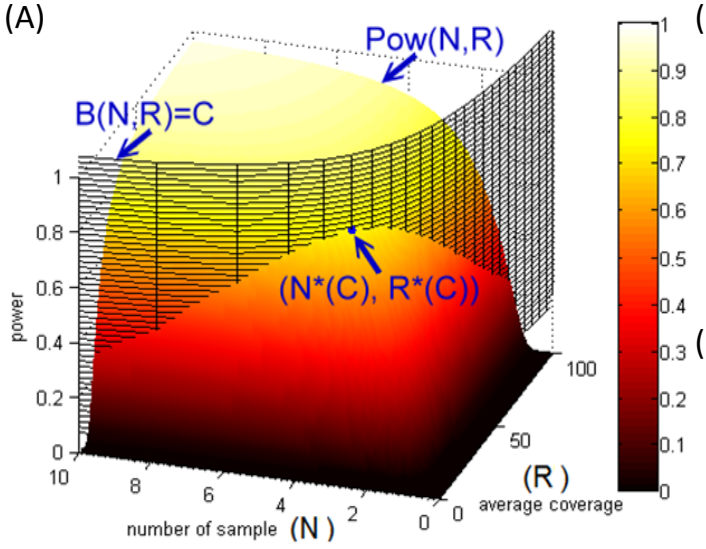
Several methods have been proposed for RNA-Seq power calculation. Busby et al. (2013) developed an interactive web tool, called "Scotty", for RNA-Seq power calculation, where the method was based on t-test with normal assumption. Hart et al. (2013) developed the "RNASeqPower" package and proposed a power calculation formula based on negative bi-

nomial model and a score test under generalized linear model framework. This method, however, is based on univariate hypothesis testing and does not consider FDR and EDR control. Li et al. (2013b) proposed a power calculation method based on exact test of negative binomial model. The method failed to consider multiple comparisons, genome-wide power and sequencing depth. Wu et al. (2015) proposed a fully simulation-based method "PROPER", which considered sequencing depth as a stratified factor rather than a second dimension for power evaluation. In this paper, we develop a unified statistical framework, namely "RNASeqDesign". Table 2 shows comparison of features of existing RNA-Seq power calculation tools and RNASeqDesign. We will consider multiple comparison FDR control, genome-wide EDR power, distribution of DE gene effect sizes, count data modeling, simultaneous optimization of sample size and sequencing depth, and incorporation of pilot data information. Additionally, RNASeqDesign allows unequal case/control sample sizes, provides a variability estimate of the power curve, and performs study design tasks and cost-benefit analysis. Unlike other methods, RNASeqDesign estimates all required parameters in the model from pilot data and does not need user-defined arbitrary input parameters such as fold change cut-off, proportion of null genes or fold change distribution of DE genes.

The chapter is structured as follows. In Section 2, we present the statistical framework of RNASeqDesign using Wald test from pilot data, model fitting of the resulting p-value distribution, parametric bootstrapping and two-dimensional smoothing for fast N and R optimization. In Section 3, we apply the methodology to develop practical cost-benefit analysis and solve five selected study design tasks. Simulations and three real data applications are shown in Section 4. Section 5 provides final conclusions and discussion. An R package "RNASeqDesign" and all data and code are available on authors website[1] and github[2].

[1]http://tsenglab.biostat.pitt.edu/software.htm
[2]https://github.com/MasakiLin/RNASeqDesign

(A)



(B)

Univariate power calculation	RNA-Seq genome-wide power calculation
Type-I error control α	FDR
Power 1-β	EDR
Univariate effect size	Effect size distribution among DE genes
Sample size N	Sample size N and sequencing depth R

(C)

	Declared non-significant	Declared significant	Total
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m-$**R**	**R**	m

(A) Illustration of two-dimensional optimal design. The X and Y axes represent sample size N and sequencing depth R respectively, and Z axis is genome-wide power EDR. Genome-wide power hypersurface $Pow(N, R)$ is a function of N and R. Increasing N and R produces higher power (from dark color to light color). When given a fixed cost/budget C, we seek the optimal design($N^*(C)$, $R^*(C)$) that maximizes $Pow(N, R)$ under the constraint of $B(N, R) \leq C$. (B) Comparison of the four key elements between univariate power calculation and RNA-Seq genome-wide power calculation. (C) Multiple testing comparisons framework.

Figure 2 Two-dimensional optimal design

2.2 GENOME-WIDE POWER CALCULATION IN RNA-SEQ

2.2.1 Notations and terminology

Consider D_0=$\{Y = (y_{gj})_{G\times(n_0+n_1)}, X = (x_j)_{1\times(n_0+n_1)}\}$ $(1 \leq g \leq G,\ 1 \leq j \leq n_0 + n_1)$ a pilot RNA-Seq dataset, where y_{gj} represents the read count for gene g of subject j, x_j is the case-control indicator (x_j=0 for controls and x_j=1 for cases), and n_0 and n_1 are the number of controls and cases in the pilot data. Denote by $\theta_p = n_1/n_0$ the sample size ratio between the number of cases (n_1) and controls (n_0). Let $R_j = \sum_{g=1}^{G} y_{gj}$ be the total number of reads observed in subject j (a.k.a. library size). For simplicity, we assume equal library size R_0 for all pilot subjects. As discussed in the previous section, we consider genome-wide power calculation under genome-wide type-I error control using FDR=E(number of claimed

Table 2 Comparison between existing methods and RNASeqDesign

	Li et al.	RNASeqPower	Scotty	PROPER	RNASeqDesign
Consider multiple comparison (FDR)	Yes	No	No	Yes	Yes
Consider genome-wide power (EDR)	No	No	Yes	Yes	Yes
Consider effect size distribution of DE genes	No	No	Yes	Yes	Yes
Consider N and R simultaneously	No	No	Yes	No	Yes
Model count data adequately	Yes (NB)	Yes (NB)	No (t-test)	Yes (NB)	Yes (NB)
Incorporate pilot data information	Minimal*	Minimal*	Minimal*	Minimal*	Yes
Allow unequal case/ control sample sizes	No	Yes	Yes	Yes	Yes
Provide variability estimate of power curve	No	No	No	No	Yes
Perform study design and cost-benefit analysis	No	No	No	No	Yes
Require user-defined input parameter	FC cut DE prop.	FC cut p-value cut	FC cut p-value cut	MC dist. Disp dist. FC dist DE prop.	No

NB: negative binomial distribution was used in the model. FC: fold change. MC: mean counts. Disp: dispersion. *: pilot data were partially used to estimate selected key parameters.

false positives/number of claimed positives) (i.e. FDR = $\mathbf{V/R}$ in Figure 2C). Following Gadbury et al. (2004), we define expected discovery rate, EDR=E(number of claimed true positives/number of total true positives) (i.e. EDR = $\mathbf{S}/(m - m_0)$ in Figure 2C), as the genome-wide average power that we aim to estimate. The basic statistical framework of RNASeqDesign is to estimate the genome-wide power $\widehat{\text{EDR}}(N_0, N_1, R|D_0)$ (equivalent to the notation $Pow(N, R)$ in Section 2.1) based on the pilot data D_0 for designing a future experiment with targeted sample sizes in control and case groups (N_0 and N_1; denote $\theta = N_1/N_0$ as the case-control ratio in targeted samples) and targeted sequencing depth R, under certain FDR control (e.g. FDR=5%). We assume equal sequencing depth R for all subjects in the planned experiment.

2.2.2 Four sequential steps for genome-wide RNA-Seq power calculation

We propose four sequential steps in RNASeqDesign to estimate EDR as the desired genome-wide power. In Step I, p-values and effect size distribution of all genes from pilot data are obtained using a negative binomial generalized linear model and Wald test. In Step II, a two beta mixture model is applied to characterize the genome-wide p-value distribution and to estimate the proportion of true DE genes. In Step III, a parametric bootstrapping method based on DE posterior probability is used to simulate and transform the genome-wide p-value distribution towards the targeted sample size and sequencing depth. In the final step, two-dimensional smoothing and hypersurface fitting is applied to stabilize the estimation of $\widehat{\text{EDR}}(N, R|D_0)$ for any N and R. Below, we describe the details of each step.

Step I. Differential expression analysis on pilot data We assume that $y_{gj} \sim \text{NB}(\mu_{gj}, \delta)$, where μ_{gj} is the mean and δ is a common dispersion parameter for gene g and subject j. The probability mass function of y_{gj} is

$$P(y_{gj}) = \frac{\Gamma(\delta + y_{gj})}{\Gamma(\delta) y_{gj}!} (\frac{\delta^{-1}\mu_{gj}}{1 + \delta^{-1}\mu_{gj}})^{y_{gj}} (\frac{1}{1 + \delta^{-1}\mu_{gj}})^{\delta},$$

where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$. Based on a generalized linear model framework, we adopt a link function $\log(\mu_{gj}) = \log(R_j) + \beta_{g0} + \beta_{g1} \cdot x_j$, where R_j is the library size for subject j and x_j is the case-control indicator. The log-likelihood becomes:

$$L_g = \sum_{j=1}^{n} [\log \frac{\Gamma(\delta + y_{gj})}{\Gamma(\delta) y_{gj}!} + y_{gj} \log(\delta^{-1}\mu_{gj}) - (y_{gj} + \delta) \log(1 + \delta^{-1}\mu_{gj})].$$

β_{g0} and β_{g1} can be estimated using maximum likelihood estimator (MLE) and variance-covariance matrix of the MLE is approximated by inverse of the estimated Fisher information matrix. To simplify the formula, we assume all samples have the same total reads R_0, i.e. $R_j = R_0$, $1 \leq j \leq n$.

For a given gene g, our goal is to test $H_0 : \beta_{g1} = 0$ versus $H_1 : \beta_{g1} \neq 0$. The Wald test procedure is used in our method in order to apply a parametric boostrapping approach in Step III. The Wald test statistic Z_g approximately follows a standard normal distribution, i.e.

$$Z_g = \frac{\hat{\beta}_{g1}}{\sqrt{\text{Var}(\hat{\beta}_{g1})}} \sim N(0, 1)$$

under the null hypothesis that no differential expression exists in gene g. MLE estimators $\hat{\beta}_{g0}$ and $\hat{\beta}_{g1}$ can be obtained by solving the equations

$$\begin{cases} \sum_{i=0}^{1}\sum_{j=1}^{n_i}[y_{gij} - (y_{gij}+\delta)\times\frac{q_{ij}}{(1+q_{ij})}] = 0 \\ \sum_{i=0}^{1}\sum_{j=1}^{n_i}[x_{ij}y_{gij} - (y_{gij}+\delta)\times\frac{x_{ij}q_{ij}}{(1+q_{ij})}] = 0 \end{cases},$$

where $q_{ij} = \delta^{-1}R_0 e^{\beta_{g0}+\beta_{g1}x_{ij}}$.

The Fisher information matrix for gene g is

$$F_g = -E(\frac{\partial^2 L_g}{\partial(\beta_{g0},\beta_{g1})}) = \begin{pmatrix} \frac{n_0 R_0 e^{\beta_{g0}}}{1+\delta^{-1}R_0 e^{\beta_{g0}}} + \frac{n_1 R_0 e^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1}R_0 e^{\beta_{g0}+\beta_{g1}}} & \frac{n_1 R_0 e^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1}R_0 e^{\beta_{g0}+\beta_{g1}}} \\ \frac{n_1 R_0 e^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1}R_0 e^{\beta_{g0}+\beta_{g1}}} & \frac{n_1 R_0 e^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1}R_0 e^{\beta_{g0}+\beta_{g1}}} \end{pmatrix}.$$

The covariance matrix of $(\hat{\beta}_{g0},\hat{\beta}_{g1})^T$ is

$$\mathrm{Cov}\begin{pmatrix}\hat{\beta}_{g0}\\ \hat{\beta}_{g1}\end{pmatrix} = F_g^{-1}(\hat{\beta}_{g0},\hat{\beta}_{g1})$$

Therefore variance estimator of $\hat{\beta}_{g1}$ is

$$\mathrm{Var}(\hat{\beta}_{g1}) = \frac{1}{n_0}\times(\frac{1+\theta_p e^{\hat{\beta}_{g1}}}{\theta_p R_0 e^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{1+\theta_p}{\theta_p\hat{\delta}}), \tag{2.1}$$

where $\theta_p = n_1/n_0$ (Zhu and Lakkis, 2013).

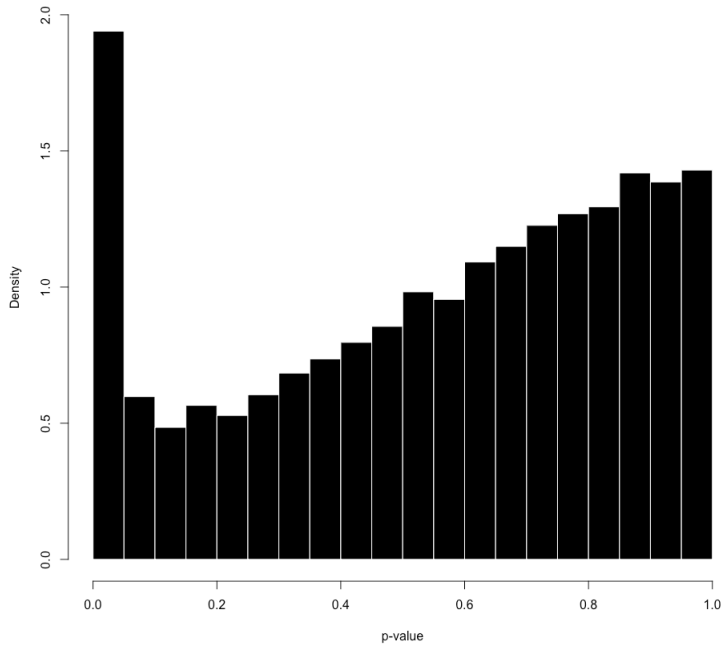
Here, the over-dispersion parameter is estimated by the conditional maximum likelihood in "edgeR" (Robinson and Smyth, 2008) assuming common dispersion across all genes. The reason for using common dispersion parameter is that when sample size is small (which is usually the case in pilot studies), estimation of tag-wise dispersion parameter is not precise with large variation. Denote by p_g the p-value of gene g from the aforementioned Wald test. As we will see in Step III, the format of variance estimator from Wald test statistic (Z-statistics) in Equation (2.1) has a convenient form to project the observed Z-statistics distribution from pilot data to the targeted sample size N_0 and N_1 and sequencing depth R. In Section 4, we will compare performance of Wald test with exact test and likelihood ratio test to justify that Wald test not only provides a convenient mathematical form for power calculation but also generates comparable hypothesis testing performance.

Step II. Mixture model fitting for p-value distribution Traditionally, a beta-uniform mixture (BUM) model (Allison et al., 2002) was used to fit the p-value distribution; that

is, using a beta distribution for DE gene p-values and a uniform distribution for non-DE gene p-values. However, when sample size is small (which is often the case of pilot data for power calculation) and data are discrete instead of continuous, the p-value distribution of non-DE genes often shows a heavy right tail, which suggests an improper fitting of uniform distribution under null hypothesis (see example in Figure 3). For example, Efron (2004) mentioned the need to estimate the null distribution and applied a non-parametric approach. Here, we propose to fit another beta distribution for non-DE genes, which results in a two beta mixture model to fit the p-value distribution $(p_1, p_2, \cdots, p_G)$ from pilot data in Step I. Specifically, the two beta model contains a non-DE gene $f_0(p)$ component and a DE gene $f_1(p)$ component. Among non-DE genes, f_0 is a beta distribution with shape parameter r_0 and s_0 with constraints $1 \leq r_0$ and $s_0 \leq 1$. When $r_0 = s_0 = 1$, the non-DE beta distribution is equivalent to a uniform distribution. For DE genes, f_1 is a beta distribution with shape parameter r_1 and s_1 with constraints $r_1 \leq 0.9$ and $1 \leq s_1$. The density of p-value distribution is $f(p|r_0, s_0, r_1, s_1, \lambda) = \lambda f_0(p|r_0, s_0) + (1-\lambda) f_1(p|r_1, s_1)$, where λ is the proportion of non-DE genes. Note that the constraints for r_0, r_1, s_0, and s_1 are necessary to guarantee the proper shapes of the p-value distributions of non-DE and DE genes, respectively.

For estimation of the five parameters (r_0, s_0, r_1, s_1, and λ), due to the fact that the estimation of λ is the most critical parameter, we adopted a robust two-step estimation procedure: we first used maximum likelihood approach and checked if the resulting non-DE beta distribution deviated from the uniform distribution, by comparing the cumulative distribution function (CDF) from 0.5 to 1. If the difference was less than 0.1, then we used a convex decreasing density estimate (CDD) method (Langaas et al., 2005) (implemented by "convest" function in R package "limma") to re-estimate λ. Then we fixed λ and used MLE again to estimate the four remaining shape parameters. If the CDF difference between non-DE beta distribution and uniform distribution was large enough (> 0.1), we simply reported the MLEs of the five parameters based on a two beta model.

Step III. Parametric bootstrapping based on DE posterior probability to estimate EDR Conceptually, the p-value distribution for non-DE genes with zero effect size follows a uniform distribution (or a beta distribution as estimated in Step II) and does not change when the sample size and sequencing depth change. On the other hand, the p-value distribution for



P-value distribution from real data showed heavy right tail, indicating a non-uniform distribution of non-DE genes.

Figure 3 P-value distribution with heavy right tail

those DE genes become more significant as sample sizes and/or sequencing depth increase. Equation (2.1) is the key formula to allow transformation of Z-statistics of DE genes to the targeted sample size N_0 and sequencing depth R. Let I_g be the latent variable representing gene g to be DE (I_g=1) or non-DE (I_g=0). We compute the posterior probability of I_g based on the estimated two beta mixture model from Step 2. Then p-values are drawn from the posterior probability of I_g to transform the Z-statistics distribution to a new Z distribution at targeted N_0 and R. Note that only p-values of DE genes should be transformed, while p-values of non-DE genes stay unchanged. Parametric bootstrapping procedures are described as below.

1. The posterior probability of the DE indicator I_g is calculated as

$$P(I_g = 1|\hat{\lambda}, \hat{r_0}, \hat{s_0}, \hat{r_1}, \hat{s_1}, p_g) = \frac{(1-\hat{\lambda})\hat{f}_1(p_g|\hat{r_1}, \hat{s_1})}{\hat{\lambda}\hat{f}_0(p_g|\hat{r_0}, \hat{s_0}) + (1-\hat{\lambda})\hat{f}_1(p_g|\hat{r_1}, \hat{s_1})},$$

where $\hat{\lambda}$, $\hat{r_0}$, $\hat{s_0}$, $\hat{r_1}$ and $\hat{s_1}$ are estimated from Step 2. In the b-th simulation ($1 \leq b \leq B$), we randomly simulate $I_g^{(b)}$ from $P(I_g = 1|\hat{\lambda}, \hat{r_0}, \hat{s_0}, \hat{r_1}, \hat{s_1}, p_g)$ for $1 \leq g \leq G$.

2. Only Z-statistics from DE genes are transformed. Specifically we derive

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \sqrt{\frac{N_0 \times (\frac{1+\theta_p e^{\hat{\beta}_{g1}}}{\theta_p R_0 e^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{(1+\theta_p)}{\theta_p \hat{\delta}})}{n_0 \times (\frac{1+\theta e^{\hat{\beta}_{g1}}}{\theta R e^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{(1+\theta)}{\theta \hat{\delta}})}} + (1 - I_g^{(b)}) \times Z_g.$$

In Equation (2.1), $\text{Var}(\hat{\beta}_1)$ can be considered as a function of n_0, θ_p and R_0. If we assume the effect size of a DE gene remains the same as n_0 and R_0 change, the above formula can transform the test statistics of DE genes to targeted N_0, θ and R.

3. Compute p-value based on the 2-sided test: $p_g^{(b)} = 2 \times (1 - \Phi(|Z_g^{(b)}|))$ if gene g with $I_g^{(b)} = 1$, where Φ is a CDF of a standard normal distribution. When $I_g^{(b)} = 0$, $p_g^{(b)} = p_g$.
4. Control FDR at level α:
   a. In the b^{th} simulation, calculate $\text{FDR}^{(b)}(u) = \frac{\sum_{g=1}^{G}(1-I_g^{(b)})\cdot\chi(p_g^{(b)}\leq u)}{\sum_{g=1}^{G}\chi(p_g^{(b)}\leq u)}$ 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.
   b. Let $u^{(b)} = \underset{u}{\text{argmax}}(\text{FDR}^{(b)}(u)) \leq \alpha)$, where $u^{(b)}$ is the p-value threshold that controls FDR at α level for the b^{th} simulation.
5. The estimated EDR for the b^{th} simulation can be calculated as $\widehat{\text{EDR}}^{(b)} = \frac{\sum_{g=1}^{G} I_g^{(b)}\cdot\chi(p_g^{(b)}<u^{(b)})}{\sum_{g=1}^{G} I_g^{(b)}}$. Finally, the robust estimated EDR for all B simulations is: $\widehat{\text{EDR}}(N_0, N_1, R|D_0) = median_b(\widehat{\text{EDR}}^{(b)})$. The first and third quantile of estimated EDR can be also derived and used to account for the variability of EDR estimation. For simplicity of presentation, we assume $N_0 = N_1 = N$ hereafter although the restriction can be relaxed easily and use $\text{EDR}(N, R|D_0)$ to represent $\text{EDR}(N_0, N_1, R|D_0)$.

Step IV. Two-dimensional smoothing and hypersurface fitting The inverse power law model has been widely applied in the machine learning field to model learning accuracy curves with increasing sample size (Mukherjee et al., 2003; Ding et al., 2014). Here we propose a two-way inverse power law hypersurface model to fit the EDR hypersurface:

$$\text{EDR}(N, R|D_0) = 1 - b \times N^{-c} - d \times R^{-e}. \tag{2.2}$$

We first calculate $\hat{\text{EDR}}(N, R|D_0)$ from Step I-III for grid selections of N and R. The inverse power law hypersurface is then fitted by minimizing sum of squared errors using BFGS

quasi-Newton method (Lewis and Overton, 2009) in R with "optim" function to estimate parameters b, c, d and e. With smoothing and hypersurface fitting, the EDR estimation is more stable and can be calculated for any N and R that are used for cost-benefit analysis in the next section. In Step III, we use a small $B = 10$ for faster computing and rely on hypersurface smoothing to reduce variability.

2.3 COST-BENEFIT ANALYSIS AND STUDY DESIGN

2.3.1 Cost function and cost-benefit analysis

In most power calculations, the experimental cost grows linearly with sample size and no complicated cost-benefit analysis is needed. In RNA-Seq analysis, however, the trade-off between sample size N and sequencing depth R brings new challenges. A common goal of study design is to pursue the best N and R balance to achieve the maximum EDR under a fixed budget constraint, as described in Figure 2. The decision is given by $(N^*, R^*) = \arg\max_{(N,R)} \widehat{\text{EDR}}(N, R|D_0)$ under the constraint that $B(N, R) \leq C$, where C is a pre-specified maximum budget.

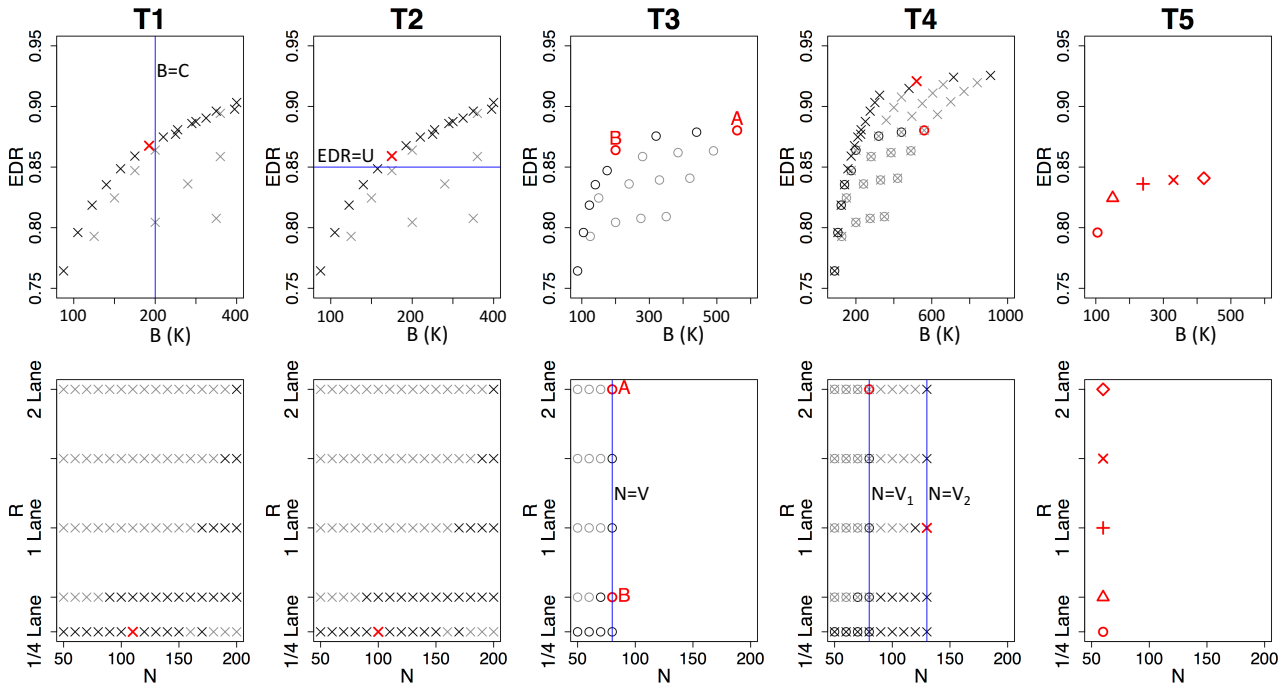
To illustrate the cost function $B(N, R)$, we checked the cost of an RNA-Seq experiment at the Sequencing and Microarray Facility core at MD Anderson. As of Dec 30, 2016, the sample preparation cost is \$500 dollars per sample. Sequencing cost for HiSeq 2000 100bp single-end reads is \$1,500 per lane for their internal users. Each lane normally generates ~250 million single-end reads. For illustration purposes, we take the most popular sequencing platforms Illumina HiSeq 2000/2500 as an example and define R as the number of lanes used for a sample. For example, if we pool three samples per lane, $R = 1/3$ lane. The resulting cost function becomes: $B(N, R) = A + (2 \times N) \times S_1 + (2 \times N) \times S_2 \times R$, where A is the fixed cost (e.g. personnel expense), S_1 is the sample collection cost per sample that includes cost of tissue collection, sample preparation and bioinformatics cost and so on, and S_2 is the sequencing cost per lane.

In Figure 4, we develop paired plots to display dynamic changes of four variables N, R, $B(N,R)$ and $EDR(N,R)$, and to illustrate the cost-benefit analysis. The lower panel of N-R plots show all feasible designs of (N,R) and the upper panel of B-EDR plots show the corresponding budget $B(N,R)$ versus genome-wide power EDR(N,R) for each (N,R) design. In practice, R is allowed with limited discrete selections (e.g. R=1/6, 1/5, 1/4, 1/3, 1/2, 1, 1.5 and 2). For a given (N,R) design, if there exists another allowed (N,R) design with lower cost and larger EDR, we consider this (N,R) design inadmissible. (N,R) design is admissible if and only if there exists no other feasible (N,R) on Quadrant II when (N,R) is treated as the origin in the B-EDR plot (see Figure 5 for illustration). In Figure 4, admissible (N,R) designs are plotted in black and inadmissible designs are shown in light grey.

2.3.2 Study design issues

As shown in Table 2, existing methods for RNA-Seq data fail to consider many key features relevant to experimental data distribution, genome-wide inference and biological objectives. Particularly, no other tool has developed guidance on decision making under different practical scenarios. This is understandable because no other method has simultaneously studied N, R, $B(N,R)$ and $EDR(N,R)$ under pre-specified FDR control. Here we propose the following five tasks under practical scenarios and utilize our power calculation framework to provide study design guidance. Figure 4 shows corresponding output of N-R and B-EDR paired plots for each task, which is implemented in the "RNASeqDesign" R package. Results in the figure are generated from simulated pilot data and will be discussed further in Section 2.4.

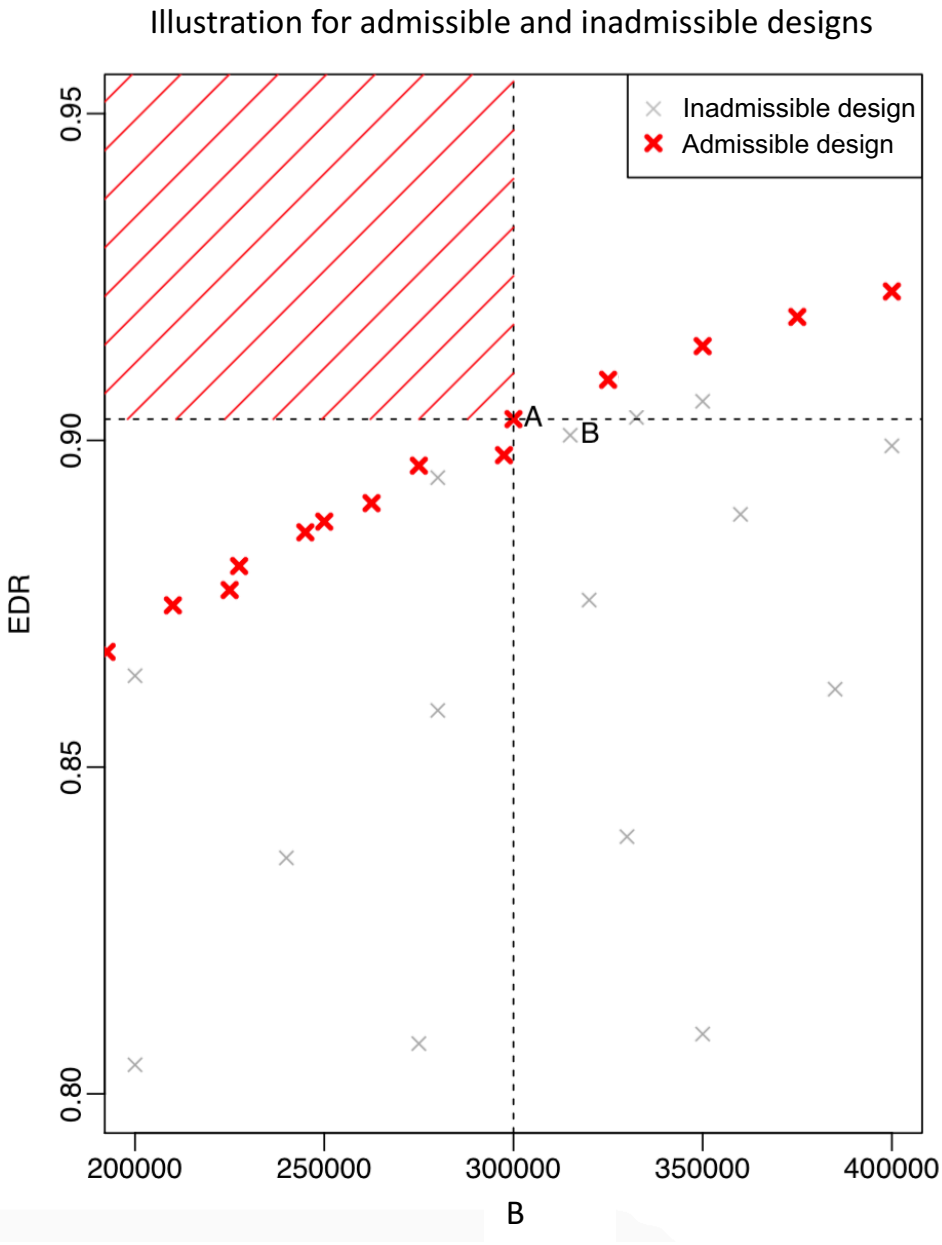
- Task 1 (T1): Given a fixed maximum budget C, what is the *optimal design* (N*, R*)? This is the most common design that has been illustrated in the Introduction Section. The red cross in Column T1 of Figure 4 shows the optimal design.
- Task 2 (T2): For a *desired power* U, how much *money* should I request in the grant and what is the corresponding design? The decision is given by $(N^*, R^*) = \arg\min_{(N,R)} B(N,R)$ such that $EDR(N,R) \geq U$. In Column T2 of Figure 4, the red cross shows the solution.



First row demonstrates B-EDR plots which show the relationship between cost and EDR, and second row is N-R plot which shows the corresponding designs (N, R). Each symbol represents a specific combination of N and R. There is an one-to-one correspondence between each symbol in B-EDR plot and N-R plot. Gray symbols refer to inadmissible designs and black ones refer to admissible designs. Blue lines illustrate fixed conditions in each task, e.g., fixed budget in T1. Symbols in red refer to the final decision in T1 and T2, and examples to illustrate ideas in T3, T4 and T5. In T3, symbols A and B are two designs given same maximum sample size where design A achieves maximum EDR with deepest sequencing. Although design A has higher EDR than B, it costs a lot more as well. In T4, circle symbols represent maximum sample size as 80 whereas cross symbols represent the case of more samples are available (up to $N = 130$). Consequently, recruiting more samples can achieve higher EDR and reduce cost. T5 evaluates the gain of power as sequencing depth gets deeper.

Figure 4 Illustration of two-dimensional optimal design in five tasks (T1-T5)

- Task 3 (T3): For a *fixed maximum sample size* V (e.g. due to limitation of tissue availability), what is the maximum achievable EDR and the corresponding cost and R? The decision is given by $(N^*, R^*) = \arg\max_{(N,R)} EDR(N, R)$ such that $N \leq V$. The red circle A in Column T3 of Figure 4 shows the highest achievable EDR.
- Task 4 (T4): For a given available sample size (maximum number of currently available tissues) V_1, is it worthwhile to *recruit more samples* (i.e. increase number of available tissues to V_2)? Will it significantly increase power and reduce cost? In this case,



For a given (N, R) design, if there exists another (N, R) design with lower cost and larger EDR, we consider this (N, R) design inadmissible. (N, R) design is admissible if and only if there exists no other feasible (N, R) on Quadrant II (area shaded in red) when (N, R) is treated as the origin in the B-EDR plot. For example, for design A, no other design exists in Quadrant II; hence design A is admissible. Design B is inadmissible because design A achieves higher EDR with lower cost. All admissible designs are highlighted in red and all inadmissible designs are highlighted in gray.

Figure 5 Illustration of admissible and inadmissible designs

we will calculate $(N^{(1)}, R_1) = \arg\max_{(N,R)} EDR(N,R)$ such that $N \leq V_1$ and similarly $(N^{(2)}, R_2) = \arg\max_{(N,R)} EDR(N,R)$ such that $N \leq V_2$. The additional power we will gain is $\Delta EDR = EDR(N^{(2)}, R_2) - EDR(N^{(1)}, R_1)$ and the cost difference is $\Delta B = B(N^{(2)}, R_2) - B(N^{(1)}, R_1)$ (or plus the recruitment cost of additional samples if necessary). In Column T4 of Figure 4, increasing allowable N from V_1 to V_2 helps achieve higher EDR with lower budget (red cross versus red circle).

- Task 5 (T5): For an existing RNA-Seq experimental data with sample size N and sequencing depth R, is it worthwhile to sequence deeper (increase from R to R) to gain more power, if remaining tissues of these N samples are available for additional sequencing? In this scenario, consider multiple possible $R > R$. Calculate the gain of power $\Delta EDR = EDR(N,R) - EDR(N,R)$ and evaluate the additional cost $\Delta B = B(N,R) - B(N,R)$ (see Column T5 of Figure 4).

Note that in all the five tasks above, maximizing power under constraint is often the goal but it may not always be the case. In many real applications, spending a lot more (ΔB) for a small increase in detection power (i.e. ΔEDR) is not desirable. To compare two potential design choices (N,R) and (N,R), the marginal utility index $\Delta U((N,R),(N,R)) = \Delta EDR/\Delta B = (EDR(N,R) - EDR(N,R))/(B(N,R) - B(N,R))$ indicates the additional detection power that can be gained while spending an additional unit of budget and can provide a good measure for decision. If $\Delta U((N,R),(N,R)) < 0$, (N,R) is inadmissible and should never be chosen.

2.4 SIMULATION AND REAL DATA ANALYSIS

2.4.1 Simulation

Simulation setting To simulate data mimicking real situations, parameter settings for our simulations are based on estimates from a real dataset downloaded from GEO (GEO accession number: GSE47474). The RNA-Seq study was designed to detect differential expression in brain regions of F344 control rats and HIV-1 transgenic rats (Li et al., 2013c). RNA tran-

scripts were sequenced in three brain regions: prefrontal cortex (PFC), hippocampus (HIP), and striatum (STR) for both HIV-1Tg and F344 rats. It includes 72 RNA samples in total (12 animals per group $\times$ 2 strains $\times$ 3 brain regions).

Similar to the data processing procedure used in the original paper, Bowtie /Tophat /Cufflinks (version 2.0.10) suites were applied to align the reads onto gene regions with the Rn4 rat reference genome. Htseq-count was used to summarize number of reads aligned to each gene. We further applied normalization method in "EDASeq" to perform within-lane normalization procedures to adjust for GC-content effect (or other gene-level effects) on read counts (Risso et al., 2011).

We started with HIP data (sample size N=12, total number of reads R=$\mu\times$G$\approx 8.85\times10^6$, total number of genes G=14,750, average counts for each gene $\mu \approx 600$). Mean count per gene(μ_g) was calculated to generate its empirical distribution. For log fold change distribution (denoted as lfc, in log2 scale), we sampled from the tails of a truncated normal distribution with mean 0, standard deviation 0.2 and truncated at 0.49/-0.49 (corresponding to at least 40% fold change) for DE genes with positive effect and negative effect in simulation study respectively. The total number of genes (G) was set to 25,000. The proportion of DE genes was set to 15%. Among DE genes, 50% had positive effect sizes and the other 50% had negative effect sizes. The average number of mean counts was set to 800 (we scaled the empirical distribution of mean counts proportionally to match this number). The total number of reads per sample was simulated with 20 million (about 4 samples per lane). The common dispersion parameter from rat dataset was estimated as 50 which reflected less biological variation in the data. To evaluate the performance in a human dataset, which tends to have larger biological variation, we simulated the data with common dispersion parameter as 5.

The detailed steps to simulate pilot data with (n_0, R_0) and targeted data with (N, R) are shown below.

1. Mean counts: Randomly sample mean counts μ_g for each gene from empirical distribution estimated by HIP data.
2. DE index: Generate random number r_g from Uniform(0,1) for each gene, if $r_g \leq$0.15 then gth gene is DE gene, otherwise, it is non-DE gene.

3. Log fold change(lfc): Generate lfc from truncated normal distribution for each DE gene. In other words, the lfc for non-DE genes is set to 0.
4. Generate count data for pilot data for each sample: If the g-th gene is DE and the j-th subject belongs to case group, then the count $y_{g1j} \sim NB(\mu_g \times 2^{lfc_g/2}, \delta)$. If the subject belongs to control group, then $y_{g0j} \sim NB(\mu_g \times 2^{-lfc_g/2}, \delta)$. If the g-th gene is non-DE gene, then $y_{g.j} \sim NB(\mu_g, \delta)$.
5. Generate count data for true data for each sample: If the g-th gene is DE and the j-th subject belongs to case group, then the count $y_{g1j} \sim NB(\mu_g \times 2^{lfc_g/2} \times \frac{R}{R_0}, \delta)$. If the subject belongs to control group, then $y_{g0j} \sim NB(\mu_g \times 2^{-lfc_g/2} \times \frac{R}{R_0}, \delta)$. If the g-th gene is a non-DE gene, then $y_{gi.} \sim NB(\mu_g \times \frac{R}{R_0}, \delta)$

Comparison of Wald test with exact test and likelihood ratio test The Wald test is applied using approximations (plugged-in standard deviation and chi-squared approximation) that sometimes raise concerns of accuracy as compared to the exact test. To demonstrate validity of Wald test, we first compared it with exact test under negative binomial distribution using "exactTest" function in "edgeR" package. Since the true labels of DE genes under simulation settings are known, we can compare the receiver operating characteristic (ROC) curve and the area under curve (AUC) of the Wald test and the exact test. We compared the two tests under 12 different simulation settings: common dispersion parameter was chosen to be 40, 50, or 60; fold change was chosen to be $\geq$ 1.15, 1.20, 1.25, or 1.30. Pilot data sample size n_0 was fixed as 4 for illustration of small sample size. In each setting, 50 datasets were generated to assess the performance variation. Given each simulated data, two tests were performed separately and ROC curves were generated by comparing the declared genes with the true DE labels. Figure 6 showed the ROC curves (with boxplot of 50 datasets). The ROC curves of exact and Wald test almost overlapped with each other, indicating good concordance between these two tests even when n_0 is as small as 4. The mean and standard deviation of AUC were presented in Table 3 for both tests.

In addition to comparisons using simulation data, we also used rat sequencing data mentioned above to compare the p-value distribution of Wald test and exact/likelihood ratio test (implemented by R function "exactTest" and "glmLRT" in "edgeR", respectively). The results (shown in Figure 2.4.1) indicated an almost perfect concordance of p-value

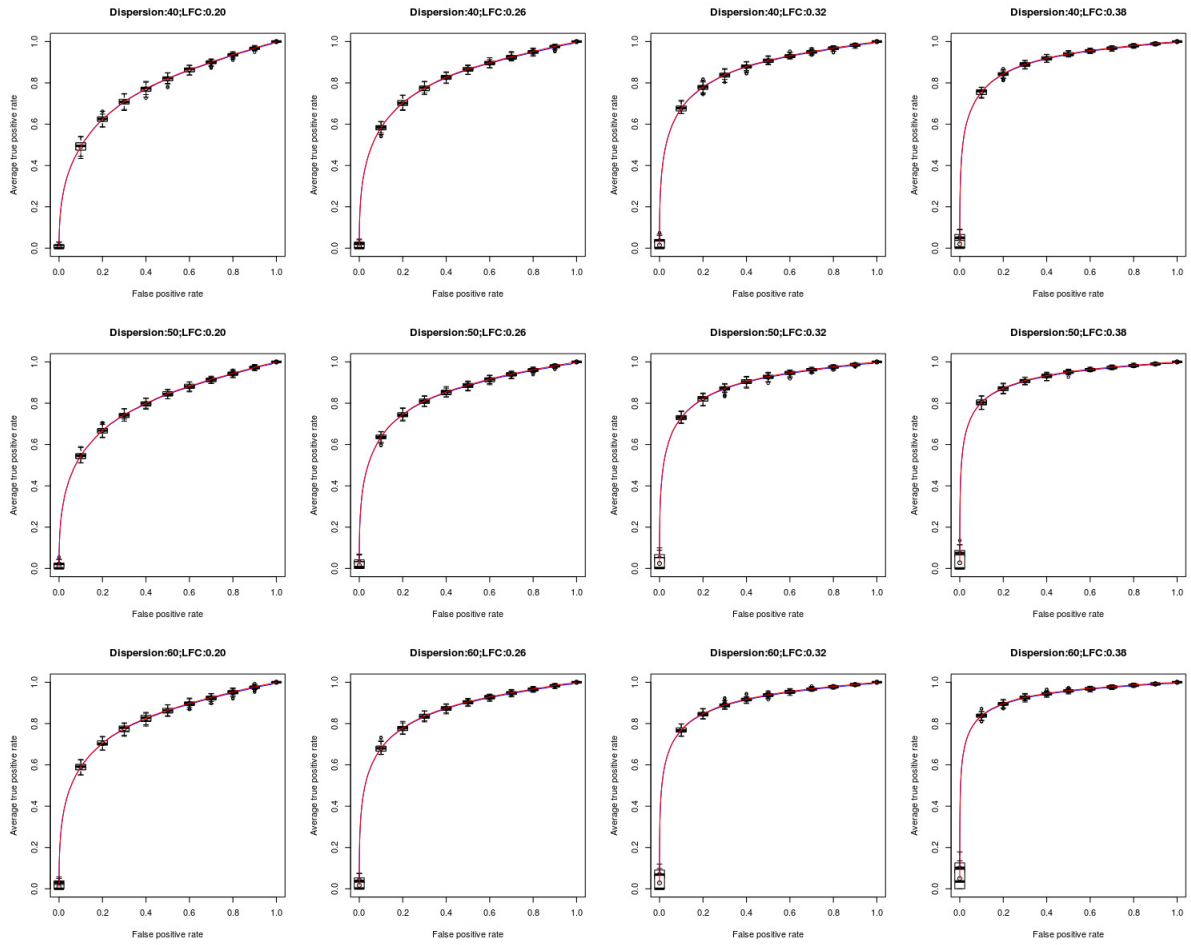


Figure 6 ROC curves comparing the exact (blue color) and Wald (red color) tests under 12 simulation settings.

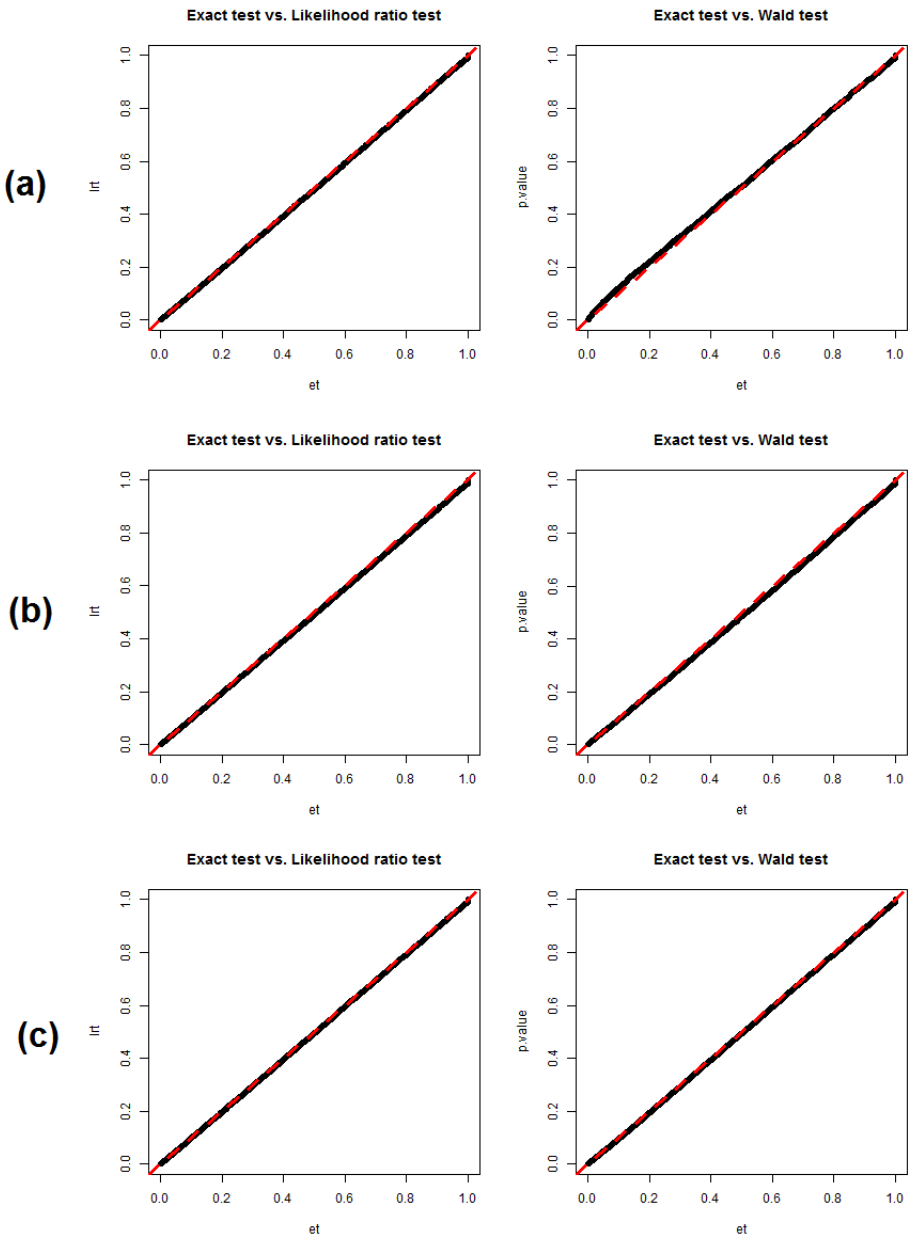
Table 3 Summary table for AUC in 12 simulation settings

fold-change(fc)	δ=40		δ=50		δ=60	
	Exact	Wald	Exact	Wald	Exact	Wald
≥ 1.15	0.77(0.01)	0.77(0.01)	0.79(0.01)	0.80(0.01)	0.82(0.01)	0.82(0.01)
≥ 1.20	0.81(0.01)	0.82(0.01)	0.84(0.01)	0.84(0.01)	0.86(0.01)	0.86(0.01)
≥ 1.25	0.86(0.01)	0.86(0.01)	0.89(0.01)	0.89(0.01)	0.90(0.01)	0.90(0.01)
≥ 1.30	0.90(0.01)	0.90(0.01)	0.92(0.01)	0.92(0.01)	0.93(0.01)	0.93(0.01)

distribution between (1) Exact test vs. likelihood ratio test and (2) Wald test vs. exact test in all three brain regions.

Implementation of existing methods In Section 2.2, we introduced several existing methods for power calculation of RNA-Seq data. Here, we compare our proposed method with five other methods: (1) Poisson exact test (Li et al., 2013a), (2) RNASeqPower, (3) Negative binomial (NB) test (Li et al., 2013b), (4) Scotty's method, and (5) PROPER under simulation setting mentioned in Section 2.4.1. Different methods have their own model specifications and parameters. For a fair comparison, we either estimated input parameters from simulated pilot data or provided the underlying truth directly to the existing methods if needed.

Genes with average count less than 5 were filtered out to reduce inflated p-value density around 1. The implementation of Poisson exact test method was based on the R code provided by original authors. True proportion of DE genes was provided to this method, which was 15% of total genes and the minimum number of mean counts in control group was estimated from the pilot data (at least 5). The FDR was set to 0.05 and the true minimum DE gene fold change of 1.4 was used as the input parameter. RNASeqPower is performed using function "rnapower" in R package "RNASeqPower". Sequencing depth is estimated by averaging the read count aligned to each gene across all samples. Biological coefficient of variation(BCV) is estimated as $\sqrt{1/\delta_{0.50}}$, where $\delta_{0.50}$ is the median of tag-wise dispersion (δ_g) obtained from function "estimateTagwiseDisp" in R package "edgeR". Effect size was set to 1.40 with alpha as 0.0001 (for a rough control of multiple comparison). For NB method, we



(a) PFC data; (b) HIP data; (c) STR data. Left panel is Quantile-Quantile plot of p-value from exact (x axis) and likelihood ratio test (y axis). Right panel is Quantile-Quantile plot of p-value from exact (x axis) and Wald test (y axis). Red dashed lines are 45 degree reference lines.

Figure 7 QQ-plot for comparisons between Exact test and likelihood ration test and Wald test

used the R function "est_power" in package "RnaSeqSampleSize". The parameter setting is similar to those in Poisson exact test except it requires additionally specifying the estimate of maximum tag-wise dispersion parameter as obtained from"edgeR". Scotty's method was implemented in MATLAB with code downloaded from https://github.com/mbusby/Scotty. Similar to RNASeqPower, effect size was set to 1.40 and p-value cutoff was set to 0.0001. Maximum number of reads to test the experiment was set to 30M. Lastly, the PROPER method was implemented by R package "PROPER". Each pilot data was used to generate empirical distribution of log mean counts and log over-dispersion. "edgeR" was used for DE analysis. FDR was set to 0.05 and fold change larger than 40% were used to declare DE genes. The rest of input parameters applied the default settings (including log fold change distribution).

Performance Evaluation We simulated B=20 pilot datasets ($b = 1, 2, ...B$) with pilot sample size n_0= 2, 4 and 8 with R_0=20M reads. For each pilot dataset with (n_0, R_0), the projected power for target sample size N_j=5, 10, 20, 30, 40, 50, 100 (j=1, 2, ..., 7) and R from a power calculation method is denoted as $\widehat{\text{EDR}}(N_j, R; n_0, R_0)$. Since the underlying truth is known, the true EDR for each (N_j, R) can be estimated as $\widehat{EDR}\,(N_j, R) = \frac{\sum_{b=1}^{B} \widehat{EDR}^{(b)}(N_j,R)}{B}$ where $\widehat{EDR}^{(b)}\,(N_j, R)$ is the actual EDR in the b-th simulation when sample size N_j and R are simulated. We propose the following benchmarks based on root mean squared error (RMSE) to evaluate performance of different power calculation methods:

1. Benchmark 1: Consider $R = R_0$ for one-dimensional power calculation from n_0 to N_j $(j = 1, 2, ..., 7)$. The RMSE of estimated EDR from power calculation is

$$\sqrt{\frac{\sum_{b=1}^{B}\sum_{j=1}^{7}\left[\widehat{EDR_j}^{(b)}\,(N_j, R; n_0, R_0) - \widehat{EDR}\,(N_j, R)\right]^2}{B \cdot 7}}.$$

2. Benchmark 2: Similar to Benchmark 1 but consider RMSE of $\hat{N}^{(b)}\,(EDR = 80\%, R; n_0, R_0)$, the sample size needed to achieve 80% EDR in the b-th simulation.

3. Benchmark 3: Similar to Benchmark 1 but consider different targeted sequencing depth from R_0=20M (quarter lane) to R_i=40M (half lane), 80M (one lane), 120M (one and

half lanes), 160M (two lanes) ($i = 1, 2, 3, 4$) lane. The RMSE of EDR becomes

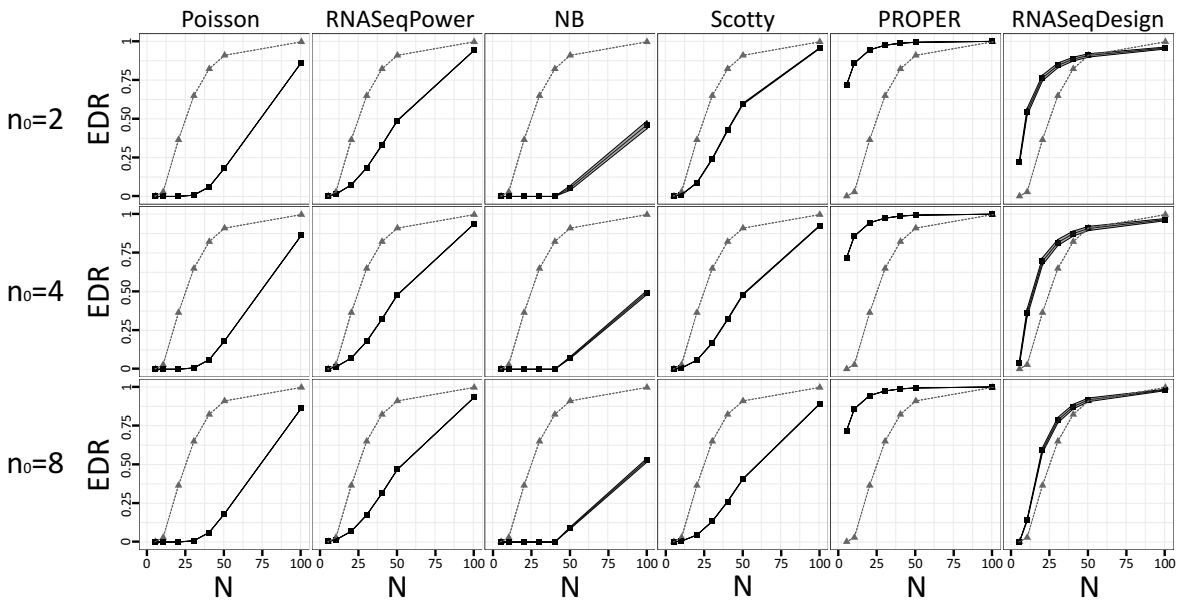
$$\sqrt{\frac{\sum_{b=1}^{B}\sum_{j=1}^{7}\sum_{i=1}^{4}\left[\widehat{EDR_{ij}}^{(b)}(N_j, R_i; n_0, R_0) - \widehat{EDR}(N_j, R_i)\right]^2}{B \cdot 7 \cdot 4}}.$$

Figure 8 shows estimated EDR curves and true EDR curves when comparing RNASeqDesign with five other existing methods. Table 4 shows the RMSEs of Benchmark 1 and Benchmark 2 and computing time. The predictive EDR curves from RNASeqDesign was closest to the true EDR curve and the performance improved when sample size of pilot data (n_0) increased, as expected. The result also showed affordable computing time (6-7 minutes using a regular laptop) for this realistic simulation setting. We altered dispersion parameter from 5 to 2, 10 and 20 and the result showed similar conclusions in Figure 9. Smaller dispersion values corresponded to larger biological coefficients of variation. In our experience, we have observed larger dispersion in TCGA human experiments and smaller variation in the HIV rat data.

In Benchmark 3, we address the goodness of fit of predicted EDR to true EDR in two-dimensional (N and R) scheme. Most existing methods do not take into account the varying sequencing depths except for Scotty. Since Scotty already performed poorly in varying N in Figure 8 and Table 4, we did not expect them to perform well in Benchmark 3. Thus, we only presented Benchmark 3 for RNASeqDesign. Here we varied fold change to 1.15, 1.20 and 1.25, and repeated 10 times. As shown in Figure 10, RNASeqDesign generated small RMSE of estimated EDR when simultaneously varying N and R and the performance improved when pilot sample size (n_0) increased.

Experimental design and N-R/B-EDR paired plots In Section 2.3 we introduced the N-R and B-EDR paired plot as a decision tool to help researchers in different design tasks. Here we illustrate with simulation results how RNASeqDesign can guide in decision making for each task in Figure 4.

For task T1 in Figure 4, say when the budget was limited to \$200,000, how can we find the optimized EDR and corresponding design (N^*, R^*)? From B-EDR plot, the optimized EDR was 0.87 and from N-R plot, the corresponding design was ($N^* = 110, R^* = 1/4$ lane). For task T2, which was in the situation when a desired power was set to 0.85, we were interested in how much minimum money was needed. From B-EDR plot, the minimum cost



Different columns represented different methods. Different rows represented different pilot data sample size (n_0=2, 4, and 8 respectively). Gray curve was true EDR and black curve was predicted EDR from each method. We ran 20 pilot data under same settings to assess the variation of performance. Confidence interval of the predicted EDR at each targeted sample size (one of N=5, 10, 20, 30, 40, 50, 100) was derived by mean predicted EDR plus/minus 1.96 $\times$ standard error of mean predicted EDR.

Figure 8 Methods comparison in simulation study

was \$175,000 and from corresponding N-R plot, the design was ($N^* = 100, R^* = 1/4$ lane). For task T3, when only limited sample size was available, say 80, the maximum EDR and corresponding cost and R were of interest. If two design decisions (A and B) were compared, where A referred to $R = 2$ lanes and B was $R = 1/2$ lane. From B-EDR plot, decisions A and B reached similar level of EDR (EDR = 0.88 for A and EDR = 0.86 for B). However, the cost of decision A was 2.5 times of the cost from decision B. Therefore, best design when $R = 1/2$ lane should be a better choice of final design. For task T4, suppose we have an initial cohort started with $N = 80$ (represented as circles) and we were considering whether recruiting 50 more samples (represented as crosses) would help to increase power and/or reduce cost. From the plots, we can see that recruiting more samples with shallower depth actually reached higher power with lower cost than smaller sample size with deeper

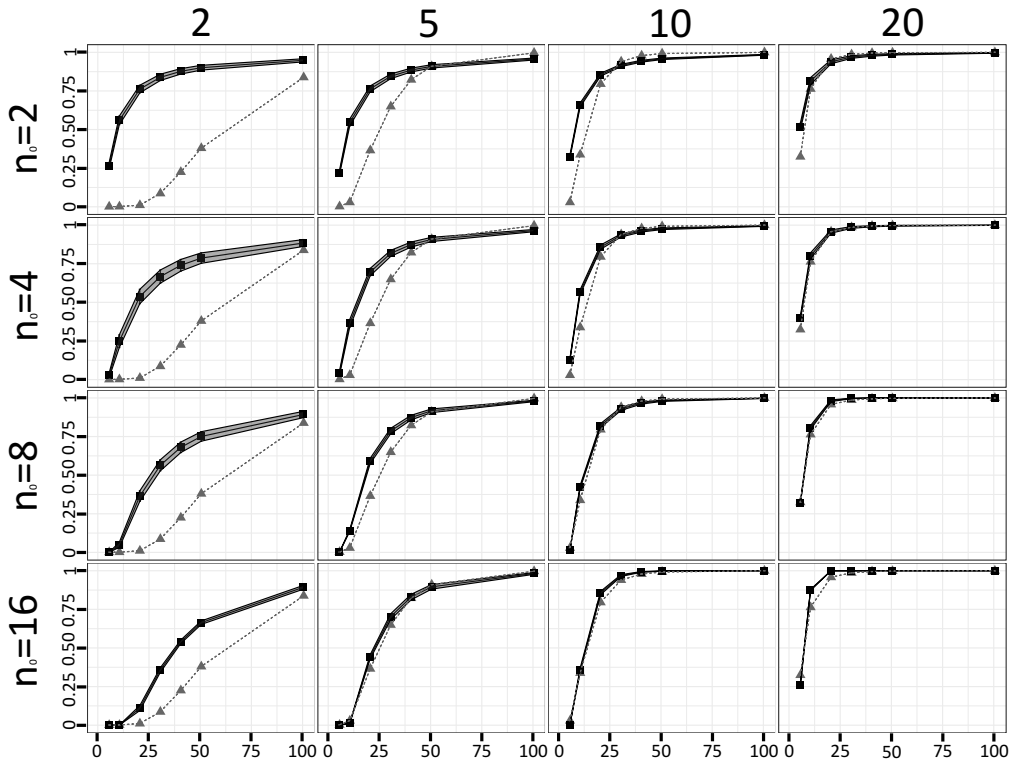


Figure 9 Performance of RNASeqDesign under different dispersion parameter settings ($\delta = 2, 5, 10, 20$)

sequencing ((N, R) = (130, 1) to reach B = \$520,000 and EDR = 0.93 versus (N, R) = (80, 2) to reach B = \$560,000 and EDR = 0.88). For task T5, it mimicked the situation when there was only a fixed number of samples, say 60, and ask if sequencing deeper would gain more power. Different symbols corresponded to different sequencing depths (circle: R=1/4 lane; triangle: R=1/2 lane; plus: R=1 lane; cross: R=3/2 lane; diamond: R=2 lane). From the plots, the EDR increased as the sequencing depth became deeper. However, after R reached 1 lane, the increment of EDR started to diminish and sequencing deeper became a waste of budget.

2.4.2 Three real applications

In this section, we demonstrate the performance of RNASeqDesign compared to other methods in three real applications: HIV-transgenic rat data, TCGA ER+ versus ER- data, and

Table 4 Performance evaluation for different methods in simulation study

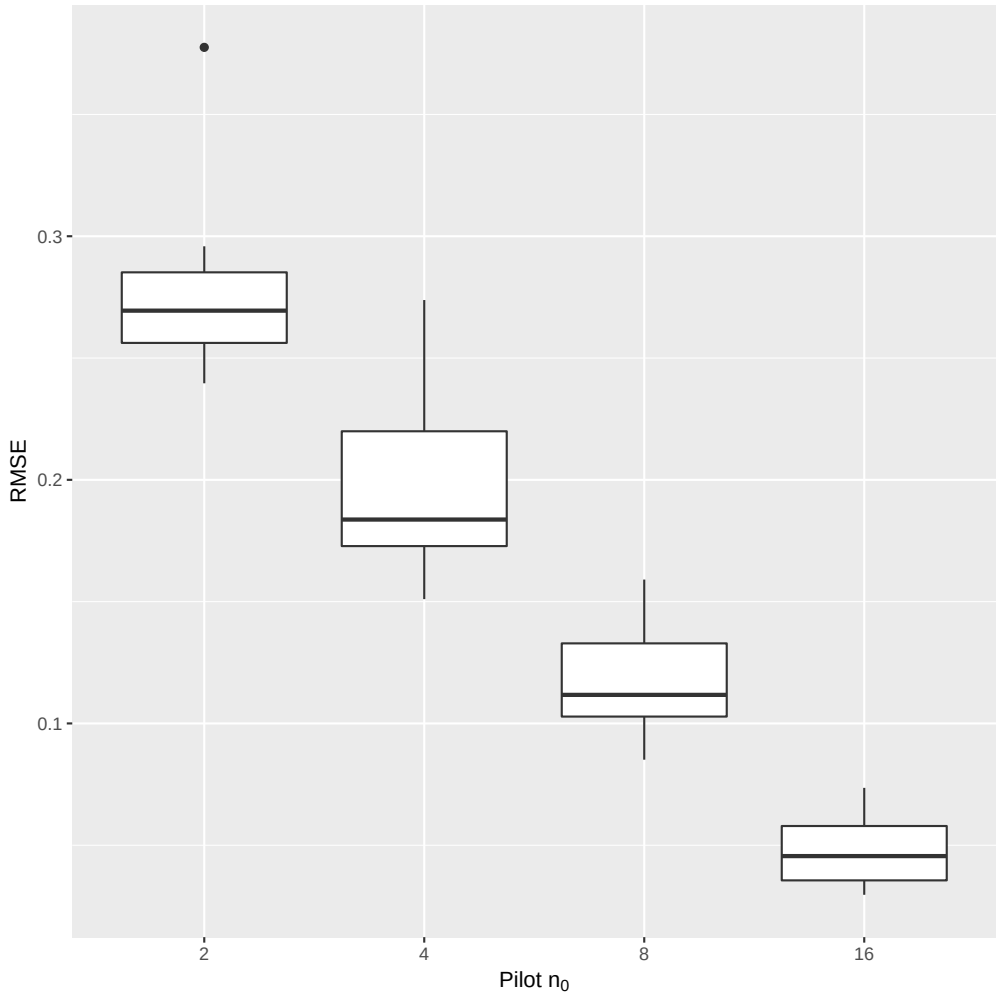
	Benchmark 1			Benchmark 2			Computing time (min.)		
Method	n_0=2	n_0=4	n_0=8	n_0=2	n_0=4	n_0=8	n_0=2	n_0=4	n_0=8
Poisson	0.49	0.49	0.49	55.0	55.0	55.0	< 1	< 1	< 1
RNASeqPower	0.32	0.33	0.33	37.9	38.9	40.0	< 1	< 1	< 1
NB	0.57	0.56	0.55	8910.4	9962.0	9962.0	< 1	< 1	< 1
Scotty	0.27	0.33	0.37	29.8	39.7	47.2	60	60	60
PROPER	0.49	0.49	0.49	31.0	31.0	31.0	42	42	42
RNASeqDesign	0.27	0.19	0.11	14.7	10.8	7.5	6	7	7

Performance evaluation based on RMSE of $\widehat{\text{EDR}}(D; D_0)$ (Benchmark 1), RMSE of $\hat{N}_{D_0, EDR^*=0.8}$ (Benchmark 2) with fixed R and computing time (unit in minutes in one simulated data) respectively in simulation analysis. Results based on different pilot sample size (n_0 = 2, 4, 8) are shown in different columns.

TCGA early versus late stage data. The settings of the other existing methods were similar to what were used in simulations unless otherwise specified.

HIV-transgenic rat data We first used aforementioned rat data from HIP brain region. Since it is a relatively weak signal dataset, we set the input parameter of proportion of DE genes to 10% for other methods. In real applications, the true underlying EDR is unknown. We instead showed how different methods performed by comparing the predicted EDR from smaller sample size (pretended as pilot data) to full sample size (N=12). We randomly subsampled n_0=2, 4 and 10 from full data to treat as pilot data and repeated independent subsampling for 10 times for each n_0. For full data, we also derived predicted EDR and treated it as a reference to compare with predicted EDR from pilot data (shown in the Figure 11(A)).

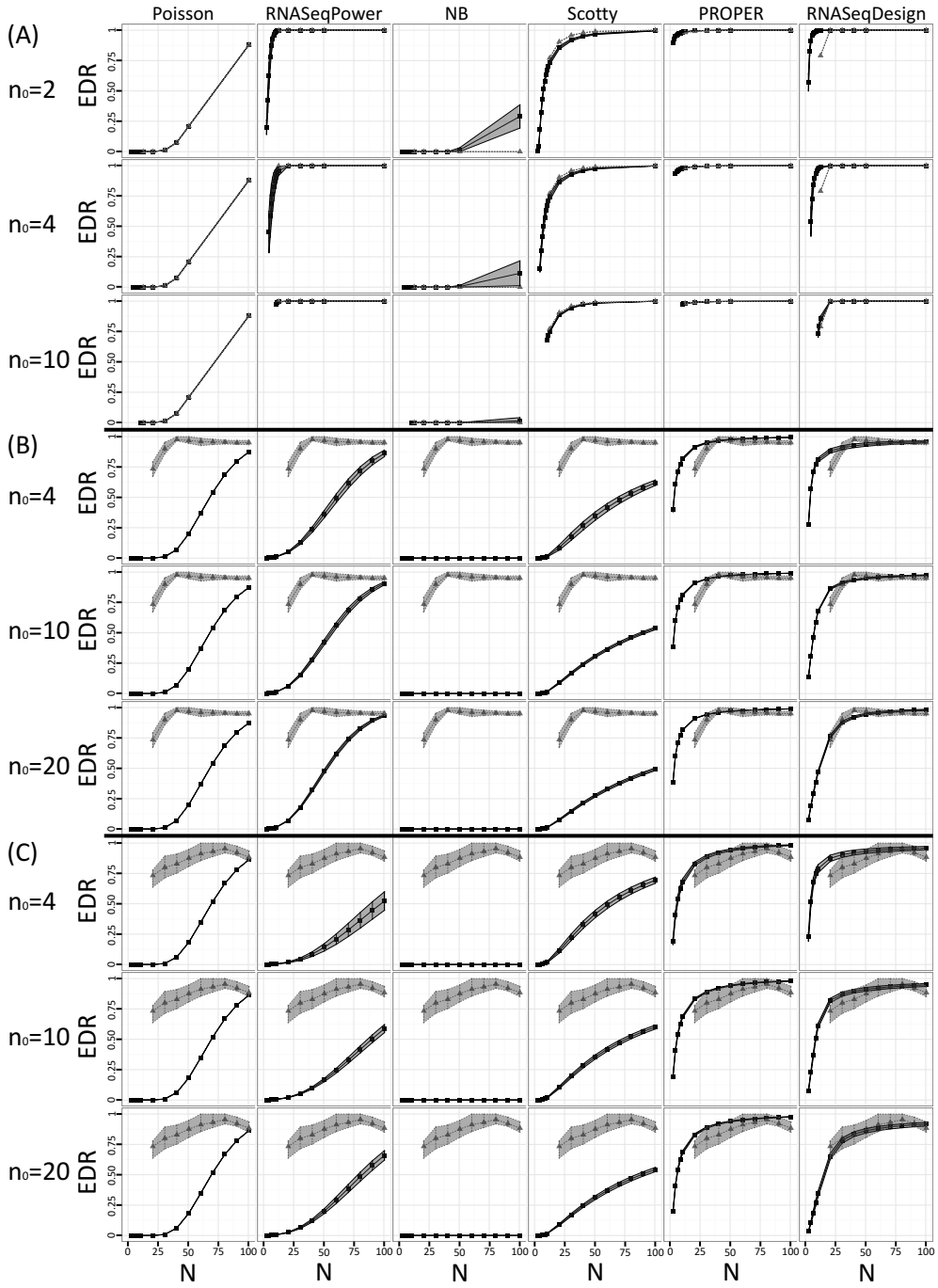
This comparison was repeated for each method separately. When $N_0 = 10$, we observed that RNASeqPower and PROPER reached almost 100% predicted EDR at small targeted sample size (e.g. $N = 2$), which was not reasonable in practical setting. For exact test



Each boxplot represents the RMSE between predicted EDR and true EDR for 20 repeatments. Pilot sample size n_0 varies from 2, 4, 8, to 16.

Figure 10 Two-dimensional goodness-of-fit

based methods (Poisson and NB), the predicted EDR was only about 15% to 25% even with targeted sample size N'=100. For most existing methods, when pilot sample size increased, the predicted EDR curves from pilot data did not change accordingly. We suspected that it was because those methods only utilized minimal level of effective information from pilot data and hence the benefits from incorporating larger pilot data were limited. Although no underlying truth was available for this application, predicted EDR from RNASeqDesign seemed to give more reasonable results.



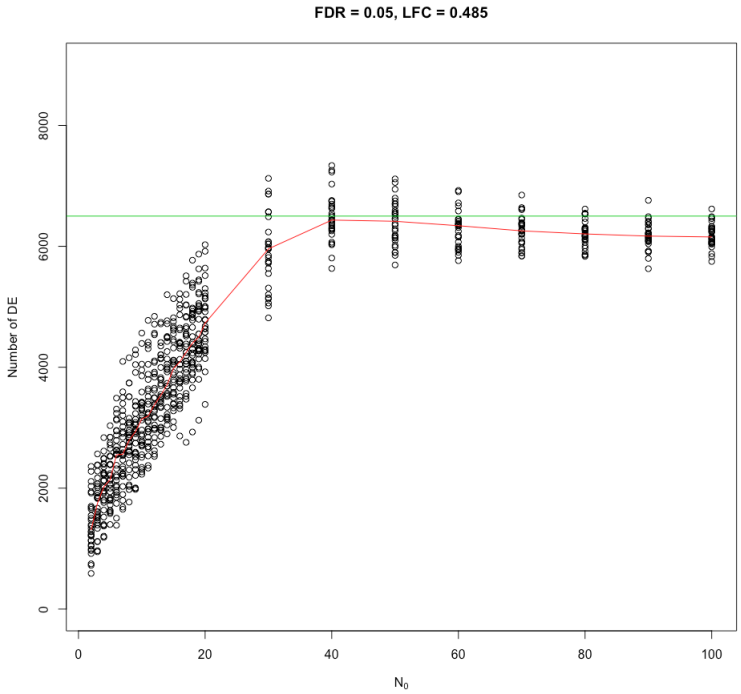
(A) HIP dataset (B) TCGA ER dataset, (C) TCGA Stage dataset. Different columns correspond to different methods. Rows refer to different pilot sample sizes. Predicted EDR from each method is in black color, and the predicted EDR from full data (HIP, N=12) or surrogate EDR (TCGA ER and Stage dataset) are in gray color.

Figure 11 Three real data applications

TCGA breast cancer ER+ versus ER- We next evaluated different methods on the publicly available TCGA breast cancer dataset with 775 tumor samples. In breast cancer, ER status is probably the most indicative biomarker in disease progression, survival prediction and treatment selection. Based on clinical information, 567 ER+ and 171 ER- tumor samples (about 3:1 ratio) were available. Although the underlying true EDR curve is still unknown, with the large sample size in this example we applied a subsampling technique to estimate the converging number of true DE genes and a "surrogate EDR curve", which we reasonably believe is close to the underlying true EDR curve. We first subsampled N=2 to 100 ER- samples with proportional ER+/ER- ratio at 3 folds (i.e. N=6 to 300 ER+ samples) and detected DE genes using Wald test at FDR=5% with more than 40% fold change to remove spurious genes with small biological effect. We performed 30 independent subsampling for each varying sample size, the number of detected DE genes were then multiplied by 0.95 to remove false positives (since FDR was controlled at 5%), and the scatter plot was shown in Figure 12. By calculating the median number of detected DE genes at each sample size, we estimated the converging number of DE genes to be around 6500 and obtained the "surrogate EDR curve" by scaling the curve to [0,1] (i.e. y-axis divided by 6500).

To perform power calculation, we subsample n_0=4, 10 and 20 from ER- patients and proportionally 3 folds (i.e. 12, 30 and 60) of ER+ patients as pilot data. We kept the sample size ratio between ER+ and ER- groups for a realistic power calculation and to demonstrate the capability of RNASeqDesign on handling unbalanced sample size design. ER status comparison is well-known to be a strong contrast with many DE genes and thus we set input parameters of 30% of DE genes when implementing power calculation with the five existing methods. From Figure 11(B)), we compared the predicted EDR from each methods to the surrogate EDR curve. The result clearly show best performance of RNASeqDesign, followed by PROPER. In RNASeqDesign, the performance improved as n_0 increased from 4 to 20, as expected. For PROPER, the predicted EDR curves remained the same for different n_0 and deviated greatly from the surrogate EDR curve at a critical sample size region around $N = 20 \sim 40$.

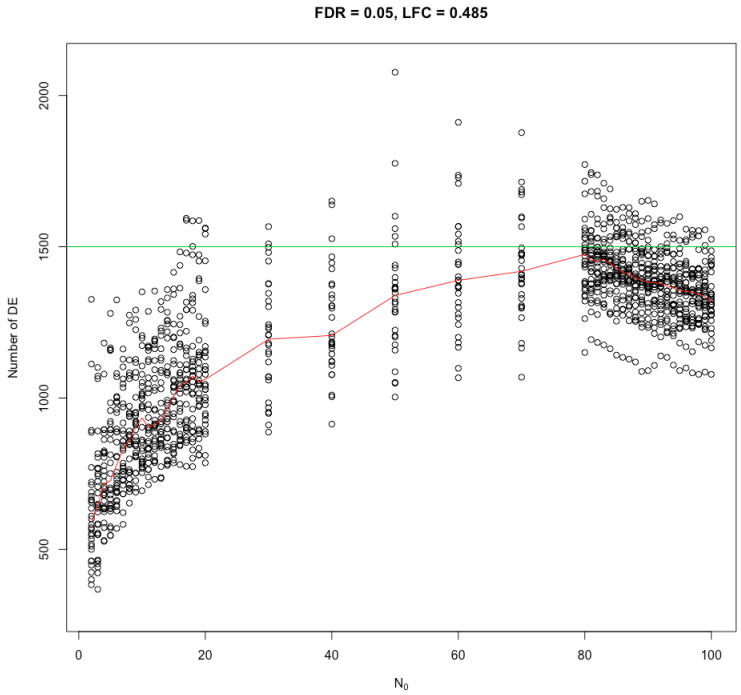
TCGA breast cancer early versus late stage For early versus late stage comparison, we classified stage I, IA and IB as early stage tumors (N=126), removed patients with stage II



At each N, we subsampled 30 times and connected the median to see the trend (red line). The observed saturated number of DE genes is 7000 (green line).

Figure 12 Number of DE genes for TCGA ER positive vs. negative dataset

and assigned stage III, IIIA, IIIB, IIIC and IV as late stage tumors (N=181). The input parameter of proportion of DE genes is set as 10% for the five existing power calculation methods. We similarly subsampled n_0=4, 10 and 20 pilot data for each group and repeated for 10 times. To obtain a surrogate EDR curve, we followed the similar subsampling procedure in ER comparison and presented the result in Figure 13. The converging number of DE genes was estimated at 1,500. Figure 11(C) shows the surrogate EDR curve and predicted EDR curves from each power calculation method. Similar to the ER comparison example, RNASeqDesign clearly performed the best and the performance improved when pilot sample size n_0 increased.



At each N, we subsampled 30 times and connected the median to see the trend (red line). The observed saturated number of DE genes is 1600 (green line).

Figure 13 Number of DE genes for TCGA early stage vs. late stage dataset

2.5 DISCUSSION AND CONCLUSION

Careful power calculations and study design are critical in high-throughput experiments to save cost and maximize the yield of experimental effort. Due to simultaneous testing of thousands of genes, power calculation of high-throughput experiments needs to consider multiple comparison control, genome-wide statistical power and distribution of effect sizes in DE genes. For RNA-Seq, we need to further consider the nature of count data and the balance between sample size and sequencing depth, which leads to a constrained optimization problem in the study design. Although several methods have been proposed for RNA-Seq power calculation, these methods miss many of the necessary elements described above. In this paper, we propose a RNASeqDesign statistical framework to accommodate all the features using information from a pilot dataset. RNASeqDesign have several unique advantages over existing methods: (1) better model fitting: Our method is based on widely accepted negative

binomial model for count data, instead of Poisson or Gaussian assumption; (2) genome-wide type I error control and power: We use genome-wide type I error control (FDR) and genome-wide power (EDR), which consider DE gene detection in the realm of whole genome, instead of at the single gene level; (3) better accuracy: Simulation and real data analysis demonstrate high accuracy of RNASeqDesign; (4) optimal study design guidance: We consider cost-benefit analysis and the influence of both sample size and read depth on genome-wide power, which provides guidance for scientists decision making in five commonly encountered study design tasks (i.e. T1 to T5); (5) better utilization of pilot data: Our method performs self-learning and better utilizes the pilot data, compared to existing methods and there is no need to specify arbitrary fold change for DE gene detection or proportion of true DE genes. To our knowledge, RNASeqDesign is the first statistical tool that comprehensively addresses the power calculation and study design issues for RNA-Seq data with the key elements mentioned in Table 1. As the sequencing cost keeps dropping, RNA-Seq experiments will become more and more prevalent and projects of large sample sizes will be expected. Thoughtful study planning, including the five tasks included in this chapter, will be essential. We believe RNASeqDesign will provide guidance for an economical and effective study design under a realistic setting.

The current RNASeqDesign framework needs a pilot dataset as an input for inference. If no pilot data exists from the same lab, one can seek existing public datasets with similar biological setting (e.g. similar tissue, disease or treatment). If possible, a recommended strategy is to perform a two-stage design by first generating suitable pilot data (e.g. n_0=n_1=6 with adequately deep sequencing). RNASeqDesign can then help determine the optimal sample size and sequencing depth needed to achieve the optimal power under a certain budget. The budget description and modelling in this paper implicitly used the popular Illumina HiSeq platforms but the framework is generalizable and applicable to any single-end or paired-end NGS platform.

There are a few technical considerations and limitations in our model and performance evaluation. In RNASeqDesign, we consider problem of thousands of simultaneous hypothesis tests and we use EDR (defined as the proportion of true detected positives among all true DE genes) as the genome-wide power. In contrast, in RNASeqPower and other meth-

ods, they usually pursue statistical power or type I error control for a single test setting. Consequently, the power from different methods are not directly comparable although we have made best effort to match them. Secondly, the Wald test in RNASeqDesign is based on Gaussian approximation of the Z-statistics. Our simulation result has shown comparable performance of Wald test with the exact test used in edgeR. One possible reason of the good approximation and performance in Wald test could be the nature of sequencing, i.e., most genes have large enough number of counts. Hence although the pilot sample size could be small, rich counts sufficed for the normality approximation of the Wald test to hold. Finally, RNASeqDesign adopts the mixture model fitting on p-value distribution. It is possible that some applications may generate p-value distributions that deviate from the parametric mixture model. More sophisticated semi-parametric model fitting will be necessary for realistic power calculation. An R package "RNASeqDesign" and all source code are available on the author's website (http://tsenglab.biostat.pitt.edu/software.htm) and github (https://github.com/MasakiLin/RNASeqDesign) for reproducing results in this paper or applying to future applications.

3.0 METHYLSEQDESIGN: A FRAMEWORK FOR METHYLATION-SEQ GENOME-WIDE POWER CALCULATION AND STUDY DESIGN ISSUES

3.1 INTRODUCTION

DNA methylation is a process in which methyl group attaches to the cytosine followed by a guanine on the DNA sequence, known as CpG sites. In the genome, there are certain regions enriched with these spots, e.g., CpG islands. Many of these regions are related to gene regulatory regions. It is well known that DNA methylation alters the gene expression level, typically represses it. This process has been found to be involved in many important biological systems, like genomic imprinting, X-chromosome inactivation, repression of repetitive elements, aging and carcinogenesis (Li et al., 1993; Paulsen and Ferguson-Smith, 2001; Robertson, 2005). In many cancer studies, aberrant DNA methylation changes are treated as the putative leading mechanism (Esteller, 2005; Baylin, 2005; Delpu et al., 2013; Licht, 2015).

There are mainly two types of technologies used to quantify the DNA methylation. One is methylation microarray (Schumacher et al., 2006) and the other is methylation sequencing (Methyl-Seq). Here we use "Methyl-Seq" to refer to whole genome bisulfite sequencing (WGBS), which which relies on bisulfite conversion of unmethylated cytosine to uracil during library preparation. Next generation sequencing (NGS) techniques are applied afterwards. Read counts are generated and after the alignment to the reference genome we know the percentage of methylated cytosine for each base. Reduced representation bisulfite sequencing (RRBS) (Meissner et al., 2005) is another NGS technology which only focus on 1% of the entire genome, which are enriched for CpG contents (CpG islands). Agilent SureSelect Methyl-Seq is a target enrichment system that enables researchers focus on biologically in-

teresting or functional regions on the genome, e.g., CpG islands, gene regulatory regions. Compare to earlier developed microarray technology, Methyl-Seq has shown superior performance in many aspects as higher per-base resolution, better accuracy, and less background noise (Hurd and Nelson, 2009). However, due to the high sequencing cost and limited budget, not many large scale epigenome-wide studies have been conducted using Methyl-Seq. Hence, sample size and power calculation methods become particularly important.

Traditional power calculations consider the relationships between effect size, α (type-I error), statistical power ($1 - type - II$ error (β)) and sample size. For example, for a given effect size (usually estimated from pilot or published data) and α (normally 5%), one is interested in calculating the sample size to reach a pre-specified statistical power (e.g. 80%) or, equivalently, to estimate statistical power given certain sample size. When analyzing high-throughput genome-wide experimental data, the situation becomes more complicated because of well-known multiple comparison problems. Since thousands of hypotheses are tested simultaneously, controlling type-I error rate and reducing false discovery in a genome-wide sense becomes critical. As a result, conservative family-wise error rate (FWER) and the scientifically more applicable false discoverate rate (FDR; (Benjamini and Hochberg, 1995)) have been proposed in the literature. Gadbury et al. (2004) introduced expected discovery rate (EDR) to replace univariate power $1 - \beta$ by addressing genome-wide detection power. Conceptually, FDR is the genome-wide analogue of type-I error α from univariate hypothesis testing and EDR is the genome-wide analogue to statistical power $1 - \beta$. Since genome-wide screening considers the whole set of differentially methylated loci/regions (DML/DMRs), specifying a univariate effect size for power calculation is not adequate and considering the effect size distribution of DML/DMRs is biologically more reasonable. A good power calculation method for high-throughput experimental data should replace α and $1 - \beta$ with FDR and EDR and consider the distribution of effect sizes among DML/DMRs.

Methyl-Seq has three unique characteristics that we need to take into account for sample size and power calculation. First, it generates randomly sequenced short reads and produces count data by nature. Methyl-Seq data need to be modeled with discrete distributions, and both sampling and biological variation should be considered. For this reason, the beta-binomial generalized linear model (GLM) (Dolzhenko and Smith, 2014; Feng et al., 2014;

Park et al., 2014) has gained popularity over the binomial GLM. Secondly, for Methyl-Seq experiments, one can choose different reads/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 or three samples per lane each with 83 million reads for the same sequencing cost. This means the power calculation problem changes from a classical one-dimensional (sample size, N) estimation to a two-dimensional (N and R) optimization problem. Thirdly, there are about 30 million CpG sites in human genome, based on the current technology, it is impossible to sequence every CpG site even with very deep sequencing depth. As a result, many CpG sites will have very low counts in many subjects, hence makes the inference for power calculation becomes extremely difficult. To overcome this problem, we start from region-based methylation data which aggregate across multiple CpG sites within a particular size of window, which helps increase the total counts in a region and makes the power calculation feasible.

To the best of our knowledge, no existing statistically rigorous power calculation methods have been developed for Methyl-Seq data. Here, we propose a statistical framework "MethylSeqDesign" and provide a useful R package to solve this important scientific question.

The chapter is structured as follows. In Section 3.2, we present the statistical framework of MethylSeqDesign using Wald test from pilot data, model fitting of the resulting p-value distribution, parametric bootstrapping and two-dimensional smoothing for fast N and R optimization. In Section 3.3, we present comprehensive simulations and real data applications. Section 3.4 provides final conclusion and discussion.

3.2 GENOME-WIDE POWER CALCULATION IN METHYL-SEQ

3.2.1 Notations and terminology

Consider D_0=$\{Y = (y_{gj})_{G\times(n_0+n_1)}, M = (m_{gj})_{G\times(n_0+n_1)}, X = (x_{jp})_{(n_0+n_1)\times P}\}$ ($1 \leq g \leq G$, $1 \leq j \leq n_0 + n_1$) 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 $G \times P$, which contains group information and other continuous or discrete covariates. When the design only has a case-control indicator, e.g., $P = 1$ and x_j=0 for controls and x_j=1 for cases respectively. n_0 and n_1 are the number of controls and cases in the pilot data. Denote by $\theta_p = n_1/n_0$ the sample size ratio between the number of cases (n_1) and controls (n_0). Let $R_j = \sum_{g=1}^{G} m_{gj}$ be the total number of reads observed in subject j (a.k.a. library size). For simplicity, we assume equal library size R_0 for all pilot subjects. As discussed in the previous section, 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) (i.e. FDR = $\mathbf{V/R}$ in Figure 2C). Following Gadbury et al. (2004), we define expected discovery rate, EDR=E(number of claimed true positives/number of total true positives) (i.e. EDR = $\mathbf{S}/(m - m_0)$ in Figure 2C), as the genome-wide average power that we aim to estimate. The basic statistical framework of RNASeqDesign is to estimate the genome-wide power $\widehat{\text{EDR}}(N_0, N_1, R|D_0)$ (equivalent to the notation $Pow(N, R)$ in Section 1) based on the pilot data D_0 for designing a future experiment with targeted sample sizes in control and case groups (N_0 and N_1; denote $\theta = N_1/N_0$ as the case-control ratio in targeted samples) and targeted sequencing depth R, under certain FDR control (e.g. FDR=5%). We assume equal sequencing depth R for all subjects in the planned experiment.

3.2.2 Four sequential steps for genome-wide Methyl-Seq Power calculation

Park and Wu (2016) proposed a decent method for detecting differentially methylated loci (DML) or regions (DMRs) based on beta-binomial generalized linear model (GLM) with arcsine link function. The estimation procedure is then based on generalized least square approach without iterative steps, which helps reduce the computation demands dramatically compared to other beta-binomial based methods (Dolzhenko and Smith, 2014; Feng et al., 2014). Considering the computing efficiency and the ability to predict EDR, we decide to embed Park and Wu (2016)'s method in our power calculation tool.

We propose four sequential steps in MethylSeqDesign to estimate EDR as the desired genome-wide power. In Step I, p-values and effect size distribution of all methylated regions

from pilot data are obtained using a beta-binomial generalized linear model with arcsine link function and Wald test. 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 DE posterior probability is used to simulate and transform the genome-wide p-value distribution towards the targeted sample size and sequencing depth. In the final step, two-dimensional smoothing and hypersurface fitting is applied to stabilize the estimation of $\widehat{\mathrm{EDR}}(N, R|D_0)$ for any N and R. Below, we describe the details of each step.

Step I. Differential expression analysis on pilot data To account for both sampling and biological variation, Y_{gj} is modeled using a beta-binomial distribution, noted as $Y_{gj} \sim beta - bin\left(m_{gj}, \pi_{gj}, \phi_g\right)$, where π_{gj} and ϕ_g are the mean and dispersion parameter of beta distribution. Under generalized linear model framework, one can associate π_{gj} and covariates through arcsine link function, i.e.,

$$\arcsin\left(2\pi_{gj} - 1\right) = x_j\beta_g, \tag{3.1}$$

where x_j is j^{th} subject's covariate, which is essentially the j^{th} row of the design matrix X. And β_g is a vector of p covariate coefficients for g^{th} CpG region.

Let $Z_{gj} = \arcsin\left(2Y_{gj}/m_{gj} - 1\right)$, and the expectation of Z_{gj} can be approximated as $E\left(Z_{gj}\right) \approx \arcsin\left[2E\left(Y_{gj}\right)/m_{gj} - 1\right] = \arcsin\left(2\pi_{gj} - 1\right) = x_j\beta_g$. Furthermore, the variance of Z_{gj} can be also approximated as $Var\left(Z_{gj}\right) \approx \frac{1+(m_{gj}-1)\hat{\phi}_g}{m_{gj}}$, which is approximately independent of mean structure. Hence, given dispersion parameter ϕ_g, the regression coefficients β_g can be estimated using generalized least square method. That is, $\hat{\beta}_g = \left(X^T V_g^{-1} X\right)^{-1} X^T V_g^{-1} Z$, where $V_g = diag\left(\frac{1+(m_{gj}-1)\phi_g}{m_{gj}}\right)$ is the covariance matrix. Given the estimator of ϕ_g as $\hat{\phi}_g = \frac{D\left(\hat{\sigma}_g^2 - 1\right)}{\Sigma_j (m_{gj}-1)}$, $\hat{V}_g = diag\left(\frac{1+(m_{gj}-1)\hat{\phi}_g}{m_{gj}}\right)$. And then the estimator of variance of β_g is $\hat{\Sigma}_g \equiv \widehat{\mathrm{var}\left(\hat{\beta}_g\right)} = \left(X^T \hat{V}_g^{-1} X\right)^{-1}$.

Hypothesis testing is tested through the Wald test. The Wald statistic is calculated as

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

where C is a vector representing any linear combination of the covariate effects. The statistics approximately follow a standard normal distribution.

For simplicity, here we consider only have class label (case or control) as our covariate (can be extended to general case). Let $n_0 = n_1 = n$ be the number of subjects in each group. Then the variance of $\widehat{Var\left(\hat{\beta}_g\right)}$ can be written as

$$\begin{aligned}
\widehat{Var\left(\hat{\beta}_g\right)} &= \frac{\sum_{j=1}^{n_0+n_1} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}_g}}{\sum_{j_1=1}^{n_0} \frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}_g} \times \sum_{j_2=1}^{n_1} \frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}_g}} \\
&= \frac{\sum_{j=1}^{2n} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}_g}}{\sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}_g} \times \sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}_g}} \\
&= \frac{n \times \frac{1}{n} \sum_{j=1}^{2n} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}_g}}{n^2 \times \frac{1}{n} \sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}_g} \times \frac{1}{n} \sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}_g}} \\
&= \frac{1}{n} \frac{\bar{A}+\bar{B}}{\bar{A} \times \bar{B}},
\end{aligned} \tag{3.2}$$

where $\bar{A} = \frac{1}{n} \sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}_g}$ and $\bar{B} = \frac{1}{n} \sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}_g}$. Here, we assume a common over-dispersion parameter shared by all CpG regions, and it is the median of all tag-wise dispersion parameters estimated from the procedure proposed by Park and Wu (2016). The reason for using a common dispersion parameter is that when sample size is small (which is usually the case in pilot studies), estimation of tag-wise dispersion parameter is not precise. Denote by p_g the p-value of region g from the aforementioned Wald test. As we will see in Step III, the format of variance estimator from Wald test statistic (Z-statistics) in Equation (3.2) has a convenient form to project the observed Z-statistics distribution from pilot data to the targeted sample size N_0 and N_1 and sequencing depth R. In Section 3.3, we will compare performance of this approach (arcsine transformation + Wald test) with other conventional options to justify that this approach not only provides a convenient mathematical form for power calculation but also generates comparable hypothesis testing performance.

Step II. Mixture model fitting for p-value distribution Traditionally, a beta-uniform mixture (BUM) model (Allison et al., 2002) was used to fit the p-value distribution. To be specific, using a beta distribution f_1 with shape parameter r and s ($0 < r < 1 \leq s$) for p-values of DMRs and a uniform distribution f_0 for p-values of non-DMRs. The density of 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. Note that the constraint for r and s is necessary to guarantee a proper shape for the p-value distribution of DMRs. Proper estimation of λ is a critical component in fitting a BUM model. To robustly estimate λ, we use a method called censored BUM (CBUM) proposed by Markitsis and Lai (2010), which alleviates the impact of extremely small p-values by treating those as censored. The shape parameters r and s can then be estimated using maximum likelihood estimator after λ is estimated from CBUM method.

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

Conceptually, the p-value distribution for non-DMRs with zero effect size follows a uniform distribution and does not change when the sample size and sequencing depth change. On the other hand, the p-values for those DMRs become more significant as sample sizes and/or sequencing depth increase. Equation (3.2) is the key formula to allow transformation of Z-statistics of DMRs to the targeted sample size N_0 and sequencing depth R. Let I_g be the latent variable representing region g to be DE (I_g=1) or non-DE (I_g=0). We compute the posterior probability of I_g based on the estimated two beta mixture model from Step II. Then p-values are drawn from the posterior probability of I_g to transform the Z-statistics distribution to a new Z distribution at targeted N_0 and R. Note that only p-values of DMRs should be transformed, while p-values of non-DMRs stay unchanged. Parametric bootstrapping procedures are described as below.

1. The posterior probability of the DE indicator I_g is calculated as

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

   where $\hat{\lambda}$, $\hat{r}$ and $\hat{s}$ are estimated from Step II. In the b-th simulation ($1 \leq b \leq B$), we randomly simulate $I_g^{(b)}$ from $P(I_g|\hat{\lambda}, \hat{r}, \hat{s}, p_g)$ for $1 \leq g \leq G$.

2. Only Z-statistics from DMRs are transformed. Specifically we derive

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \sqrt{\frac{N_0}{n_0}} + (1 - I_g^{(b)}) \times Z_g.$$

In Equation (3.2), $\text{Var}(\hat{\beta}_1)$ can be considered as a function of n_0, θ_p and R_0. If we assume the effect size and the quantity $\frac{\bar{A}+\bar{B}}{\bar{A}\times\bar{B}}$ of a DMR remain the same as n_0 and R_0 change, the above formula can transform the test statistics of DMRs to targeted N_0, θ and R.

3. Compute p-value based on the 2-sided test: $p_g^{(b)} = 2 \times (1 - \Phi(|Z_g^{(b)}|))$ if region g with $I_g^{(b)} = 1$, where Φ is a CDF of a standard normal distribution. When $I_g^{(b)} = 0$, $p_g^{(b)} = p_g$.
4. Control FDR at level α:
   a. In the b^{th} simulation, calculate $\text{FDR}^{(b)}(u) = \frac{\sum_{g=1}^{G}(1-I_g^{(b)})\cdot\chi(p_g^{(b)}\leq u)}{\sum_{g=1}^{G}\chi(p_g^{(b)}\leq u)}$ 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.
   b. Let $u^{(b)} = \underset{u}{\text{argmax}}(\text{FDR}^{(b)}(u)) \leq \alpha)$, where $u^{(b)}$ is the p-value threshold that controls FDR at α level for the b^{th} simulation.
5. The estimated EDR for the b^{th} simulation can be calculated as $\widehat{\text{EDR}}^{(b)} = \frac{\sum_{g=1}^{G} I_g^{(b)}\cdot\chi(p_g^{(b)}<u^{(b)})}{\sum_{g=1}^{G} I_g^{(b)}}$. Finally, the robust estimated EDR for all B simulations is: $\widehat{\text{EDR}}(N_0, N_1, R|D_0) = median_b(\widehat{\text{EDR}}^{(b)})$. The first and third quantile of estimated EDR can be also derived and used to account for the variability of EDR estimation. For simplicity of presentation, we assume $N_0 = N_1 = N$ hereafter although the restriction can be relaxed easily and use $\text{EDR}(N, R|D_0)$ to represent $\text{EDR}(N_0, N_1, R|D_0)$.

Step IV. Two-dimensional smoothing and hypersurface fitting The inverse power law model has been widely applied in the machine learning field to model learning accuracy curves with increasing sample size (Mukherjee et al., 2003; Ding et al., 2014). Here we propose a two-way inverse power law hypersurface model to fit the EDR hypersurface:

$$\text{EDR}(N, R|D_0) = 1 - b \times N^{-c} - d \times R^{-e}. \tag{3.3}$$

We first calculate $\hat{\text{EDR}}(N, R|D_0)$ from Step I-III for grid selections of N and R. The inverse power law hypersurface is then fitted by minimizing sum of squared errors using BFGS quasi-Newton method (Lewis and Overton, 2009) in R with "optim" function to estimate parameters b, c, d and e. With smoothing and hypersurface fitting, the EDR estimation is

more stable and can be calculated for any N and R. In Step III, we use a small $B = 10$ for faster computing and rely on hypersurface smoothing to reduce variability.

3.3 SIMULATION ANALYSIS AND REAL DATA ANALYSIS

3.3.1 Simulation

The motivating dataset is from Katz et al. (2015), which investigated the protective effect of pregnancy toward breast cancer in mice. The DNA methylation data is from the mammary gland tissue. The sample library is prepared using Agilent SureSelectXT Mouse Methyl-Seq Kit. The Kit design covers 109 Mb of Ensemble regulatory features (CpG shores and shelves, DNAse I hypersensitive sites, transcription factor-binding sites, etc.), CpG islands, known tissue-specific DMR, and open regulatory elements. The alignment reference genome used was mm9 assembly (Kent et al., 2002).The mm9 DNA reference genome was converted to a DNA methylation reference genome. Genome indexing was performed using Bismark genome preparation tools. Aligned reads outside of the targeted regions (provided from Agilent SureSelectXT Mouse Kit) were removed. Data preprocessing is performed by R package "MethyKit". We only use samples of time point one which included 6 vs. 6 mice for the analysis.

We simulated data based on parameters that we estimated directly from above mouse pregnancy dataset. We empirically draw mean counts, baseline methylation proportion in control group, and effect size (methylation level difference between different groups) from the data. In total 2000 regions were simulated and we assume 10% of them are DMRs. The common dispersion parameter is set to 0.048, which was estimated by the median of tag-wise dispersion parameters.

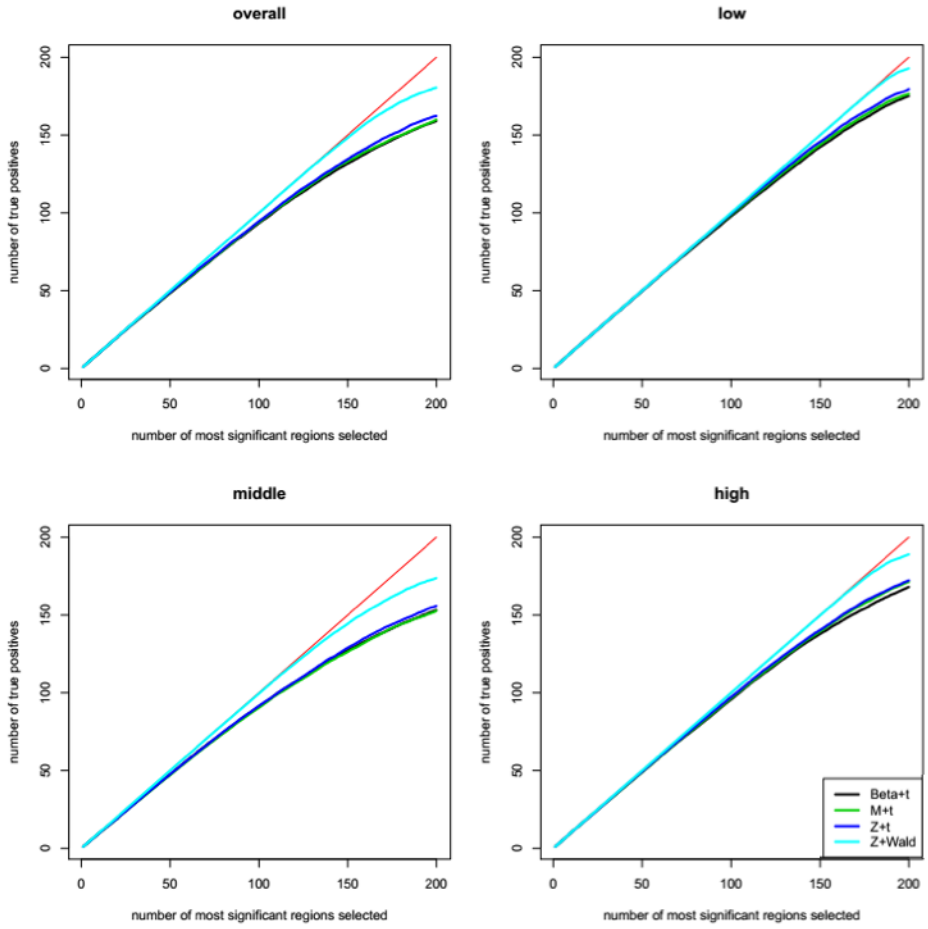
The detailed steps to simulate pilot data with (n_0, R_0) and targeted data with (N, R) are shown below.

1. Mean counts: Randomly sample mean counts μ_g for each CpG region from the empirical distribution as estimated by the mouse pregnancy data.

2. Baseline methylated proportion: Randomly baseline methylated proportion p_g for each CpG region from the empirical distribution as estimated by the mouse pregnancy data.
3. DE index: Generate random number r_g from $Uniform(0,1)$ for each gene, if $r_g \leq 0.1$ then g-th region is DMR, otherwise, it is non-DMR.
4. Effect size $\triangle$: Generate effect size from $Uniform(0.1, 0.2)$ for each DMR. The effect size for non-DMRs is set to 0.
5. Generate count data for pilot data for each sample: If the g-th region is DMR and the j-th subject belongs to control group, then the methylated counts $y_{gj} \sim beta-bin(\mu_g, \alpha_g, \beta_g)$, where α_g and β_g are estimated from given ϕ and p_g. If the subject belongs to case group, then $y_{gj} \sim beta-bin(\mu_g, \alpha'_g, \beta'_g)$, where α'_g and β'_g are estimated from ϕ and $p'_g = p_g + \triangle_g$. If the g-th region is non-DMR, then $y_{gj} \sim beta - bin(\mu_g, \alpha_g, \beta_g)$.
6. Generate count data for true data for each sample: If the g-th region is a DMR and the j-th subject belongs to the control group, then the methylated counts $y_{gj} \sim beta - bin(\mu_g \times \frac{R}{R_0}, \alpha_g, \beta_g)$, where α_g and β_g are estimated from ϕ and p_g. If the subject belongs to the case group, then $y_{gj} \sim beta - bin(\mu_g \times \frac{R}{R_0}, \alpha'_g, \beta'_g)$, where α'_g and β'_g are estimated from ϕ and $p'_g = p_g + \triangle_g$. If the g-th region is a non-DMR, then $y_{gj} \sim beta - bin(\mu_g \times \frac{R}{R_0}, \alpha_g, \beta_g)$.

Hypothesis testing performance Here we first compared the power performance of our proposed test statistic with other three naive methods: Beta value with t-test, M value with t-test, and our Z value with t-test.

To compare the performance, we conduct the analysis by stratifying the baseline methylation proportion in control group into three categories: low ($0 < p < 0.2$), middle ($0.2 < p < 0.8$), and high ($0.8 < p < 1$). In each baseline group, we simulate 20 times independent analysis, in which pilot data have 10 subjects in each group (e.g., $n_0 = n_1 = 10$), and 2000 regions (200 are DMRs). As shown in figure 14, under FDR control 0.05, we compare the power based on how many true DMR can be declared among top declared DMR. As a result, our proposed method outperforms all other methods in all baseline groups. Furthermore, we observe that the power of each method is stronger in either low or high baseline group and relatively weaker in middle baseline group, which is reasonable.



Stratified power comparisons with naive methods in three different signal level, low, middle and high. Different color represents different methods as shown in the legend (Beta values with t-tests in black, M values with t-tests in green, arcsine transformed Z statistics with t-tests in blue, and arcsine transformed Z statistics with Wald tests in cyan). X-axis is the number of top declared DMRs and Y-axis is the number of true DMRs among selected. Over all conditions, the Wald tests with arcsine transformed Z statistics perform the best.

Figure 14 Hypothesis testing performance comparisons based on stratified baseline methylation level

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 with R_0=2M reads. For each pilot dataset with (n_0, R_0), the projected power for target sample size $N_j = 5, 10, 15, 20, 30, 50$ ($j = 1, 2, ..., 6$) and R from a power calculation method is denoted as $\widehat{\text{EDR}}(N_j, R; n_0, R_0)$. Since the underlying truth is known, the true EDR for each (N_j, R) can be estimated as $\widehat{EDR}\,(N_j, R) = \frac{\sum_{b=1}^{B} \widehat{EDR}^{(b)}(N_j,R)}{B}$

where $\widehat{EDR}^{(b)}(N_j, R)$ is the actual EDR in the b-th simulation when sample size N_j and R are simulated. We propose the following benchmarks based on root mean squared error (RMSE) to evaluate performance of different power calculation methods:

1. Benchmark 1: Consider $R = R_0$ for one-dimensional power calculation from n_0 to N_j ($j = 1, 2, ..., 6$). The RMSE of estimated EDR from power calculation is

$$\sqrt{\frac{\sum_{b=1}^{B}\sum_{j=1}^{6}\left[\widehat{EDR_j}^{(b)}(N_j, R; n_0, R_0) - \widehat{EDR}(N_j, R)\right]^2}{B \cdot 7}}$$

We first perform a stratified analysis based on different level of effect size, as we already know it will impact the EDR. $\triangle$ are set as 0.1, 0.14, and 0.18. In each setting, we generate the same number of regions to compare the performace (Figure 15). Table 5 shows the RMSEs of Figure 15 based on Benchmark 1 and computing time. The predictive EDR curves from MethylSeqDesign was close to the true EDR curve and the performance improved when sample size of pilot data (n_0) increased, as expected. The result also showed affordable computing time (6-7 minutes using a regular laptop) for this simulation setting. Secondly, to mimic real situation, we generate $\triangle$ from $Uniform(0.1, 0.2)$ and compare the performance (Figure 16).

3.3.2 Real data application

In this section, we demonstrate the performance of MethylSeqDesign using a real dataset which studied the epigenetic changes in chronic lymphocytic leukemia (CLL). It contains 43 tumors and 8 controls subjects in total. The GEO accession number is GSE66167.

In real applications, the true underlying EDR is unknown. We instead showed how our method performed by comparing the predicted EDR from smaller sample size (pretended as pilot data) to full sample size ($N_0 = 8$, $N_1 = 43$). Since the sample size in control and tumor groups are unbalanced (tumor samples is roughly 5 times larger than control groups), we keep this ratio and randomly subsampled n_0=2, 4 and 6 from full data to treat as pilot data and repeated independent subsampling for 10 times for each n_0. For full data, we also derived predicted EDR and treated it as a reference to compare with predicted EDR from

Table 5 Performance evaluation in simulation study stratified by different effect sizes

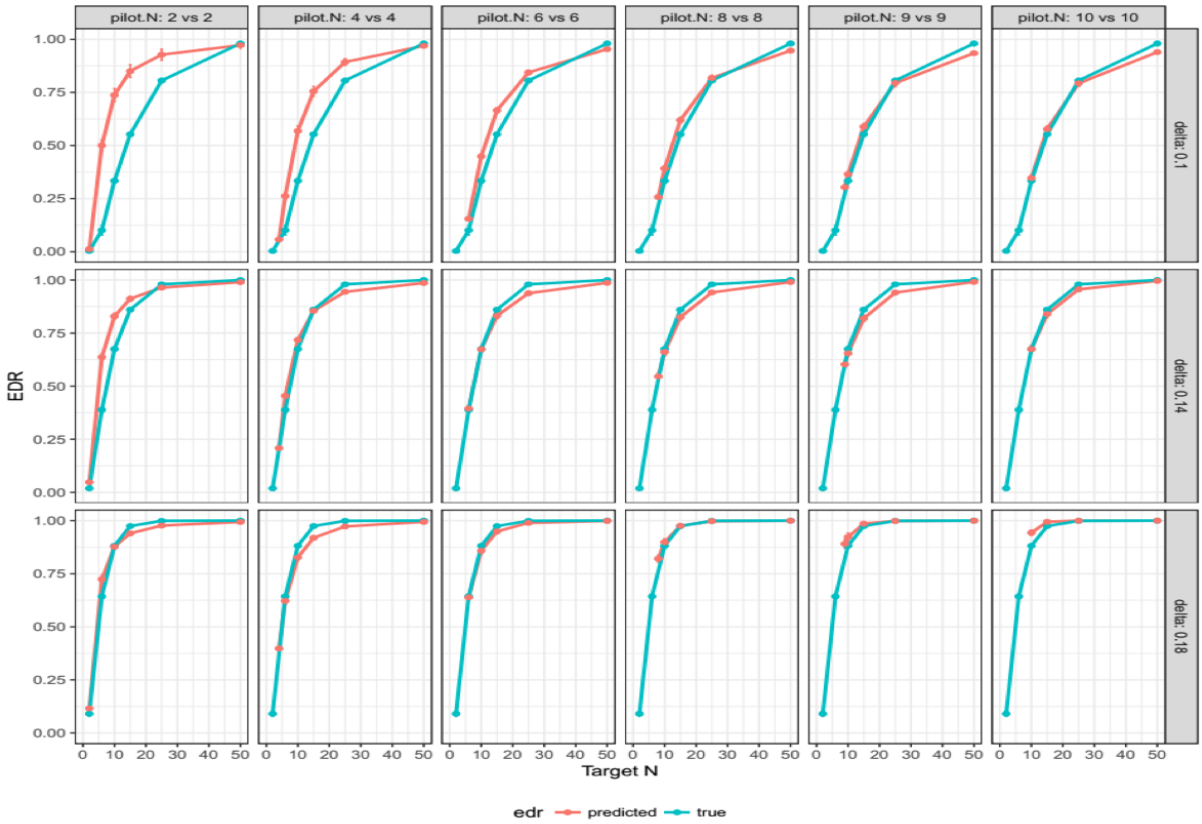
	Benchmark 1			
Pilot n_0	$\triangle = 0.1$	$\triangle = 0.14$	$\triangle = 0.18$	Overall
2	0.27	0.12	0.04	0.2
4	0.16	0.04	0.04	0.1
6	0.08	0.02	0.02	0.05
8	0.05	0.02	0.01	0.01
9	0.03	0.02	0.02	0.01
10	0.2	0.01	0.03	0.01

Performance evaluation based on RMSE of $\widehat{\text{EDR}}(D; D_0)$ (Benchmark 1) 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, "Overall" refers as analysis generating $\triangle$ from $Uniform(0.1, 0.2)$.

pilot data (shown in the Figure 17). Although no underlying truth was available for this application, predicted EDR from our method seemed to give reasonable results.

3.4 DISCUSSION AND CONCLUSION

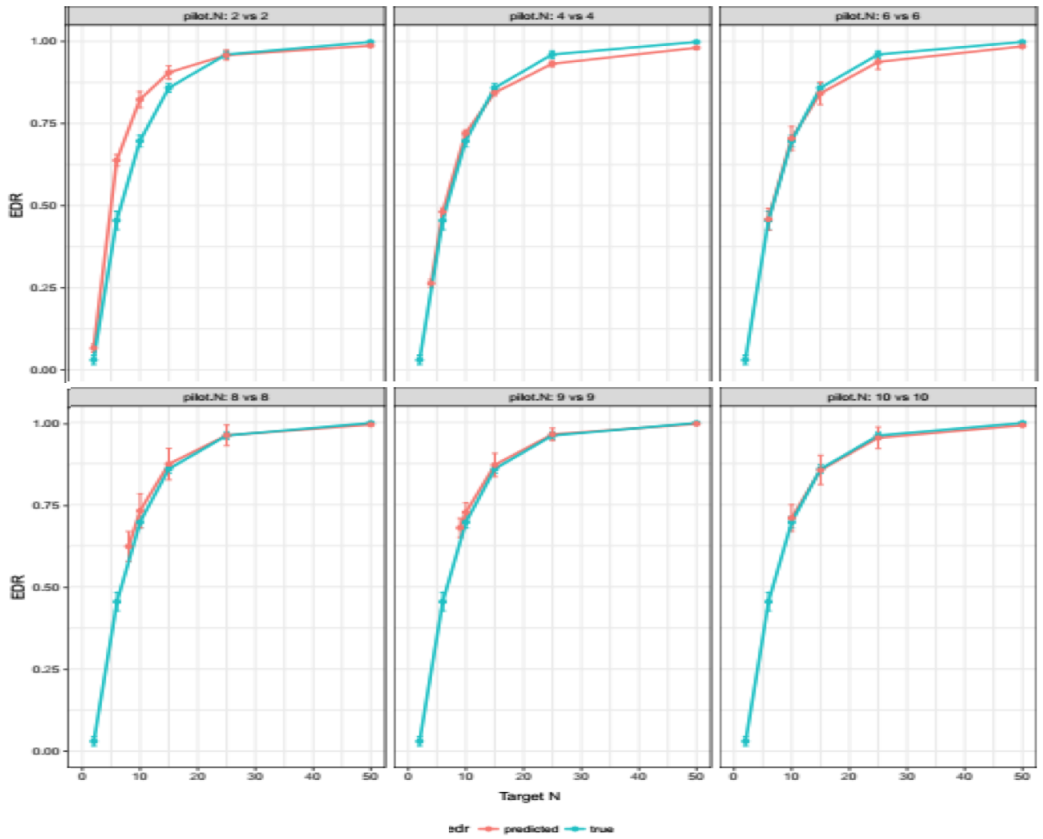
Careful power calculation and study design is critical in high-throughput experiments to save cost and maximize the yield of experimental effort. Due to simultaneous testing of thousands of CpG sites/regions, power calculation of high-throughput experiments needs to consider multiple comparison control, genome-wide statistical power and distribution of effect sizes in DMRs. For Methyl-Seq, we need to further consider the nature of count data and the balance between sample size and sequencing depth, which leads to a constrained optimization problem in the study design. In this paper, we propose a MethylSeqDesign statistical framework to accommodate all the features using information from a pilot dataset. MethylSeqDesign provides: (1) better model fitting: Our method is based on beta-binomial model for methy-



EDR prediction from MethylSeqDesign compared to true EDR using different pilot data sample sizes from 2 to 10. Effect sizes are fixed at 0.1, 0.14, and 0.18 respectively.

Figure 15 Stratified simulation study for MethylSeqDesign

lation count data, instead of Gaussian or Binomial assumption; (2) genome-wide type I error control and power: We use genome-wide type I error control (FDR) and genome-wide power (EDR), which consider DMR detection in the realm of whole genome, instead of a single site/region level; (3) better accuracy: Simulation and real data analysis demonstrate high accuracy of MethylSeqDesign; (4) optimal study design guidance: We consider the influence of both sample size and read depth on genome-wide power; (5) better utilization of pilot data: Our method performs self-learning and utilizes the pilot data without the needs to specify arbitrary fold change for DMR detection or proportion of true DMRs. To our knowledge, MethylSeqDesign is the first statistical tool that comprehensively addresses the power calculation and study design issues for Methyl-Seq data. As the sequencing cost keeps

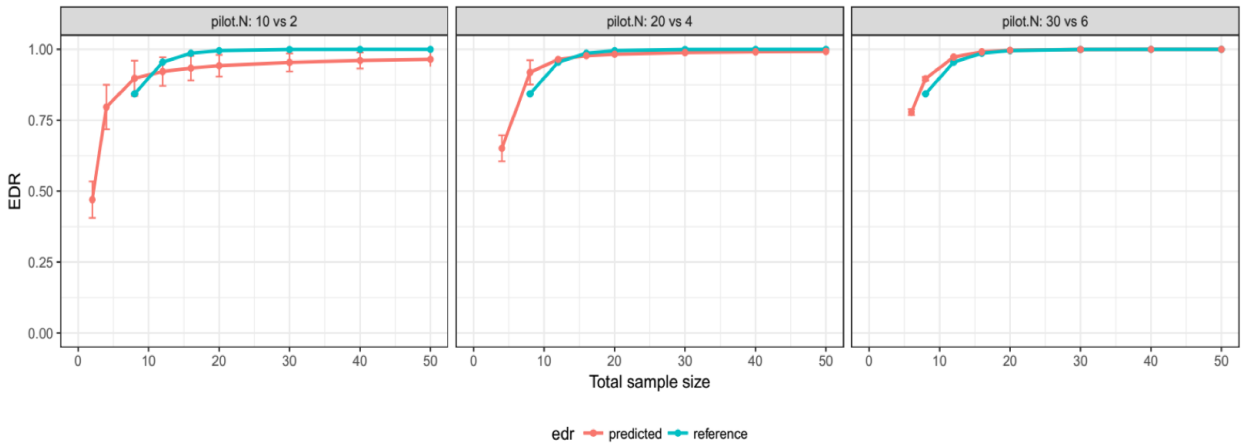


EDR prediction from MethylSeqDesign compared to true EDR. Effect sizes are sampled from $Uniform(0.1, 0.2)$.

Figure 16 Simulation study for MethylSeqDesign with variable effect size

dropping, Methyl-Seq experiments will become more and more prevalent and projects of large sample sizes will be expected. Thoughtful study planning will be essential. We believe MethylSeqDesign will provide guidance for an economical and effective study design under realistic settings.

The current MethylSeqDesign framework needs a pilot dataset as an input. If no pilot data exists from the same lab, one can seek existing public datasets with similar biological setting (e.g. similar tissue, disease or treatment). If possible, a recommended strategy is to perform a two-stage design by first generating suitable pilot data (e.g. n_0=n_1=6 with adequately deep sequencing). MethylSeqDesign can then help determine the optimal sample size and sequencing depth needed to achieve the optimal power under a certain budget.



As pilot data has larger sample size, the performance is closer to the full dataset.

Figure 17 Real data application using CLL dataset

There are a few technical considerations and limitations in our model and performance evaluation. The Wald test in MethylSeqDesign is based on Gaussian approximation of the Z-statistics. Our simulation result has shown superior performance of the Wald test over other tests, e.g., Beta values with t-tests, M values with t-tests, and arcsine transformed Z statistics with t-tests. One possible reason of the good approximation and performance of the Wald test could be the nature of sequencing, i.e., most CpG regions have large enough number of counts. Hence although the pilot sample size could be small, rich counts sufficed the normality approximation in Wald test. Secondly, MethylSeqDesign adopts the mixture model fitting of the p-value distribution. It is possible that some applications may generate p-value distributions that deviate from the parametric mixture model. More sophisticated semi-parametric model fitting will be necessary for realistic power calculation.

An R package "MethySeqDesign" is under preparation and will be released publicly soon.

4.0 DISCUSSION AND FUTURE DIRECTION

Careful power calculation and study design is critical in high-throughput experiments to save cost and maximize the yield of experimental effort. Due to simultaneous testing of thousands of genes, power calculation of high-throughput experiments needs to consider multiple comparison control, genome-wide statistical power and distribution of effect sizes in DML/DMRs. For NGS data, we need to further consider the nature of count data and the balance between sample size and sequencing depth, which leads to a constrained optimization problem in the study design. For RNA-Seq, although several methods have been proposed for RNA-Seq power calculation, these methods miss many of the necessary elements described in Table 2. For Methyl-Seq, unfortunately, to the best of our knowledge, no existing methods had been proposed. In this dissertation, we propose two statistical frameworks to accommodate all the features using information from a pilot dataset, RNASeqDesign and MethylSeqDesign. These two methods have several unique advantages over existing methods: (1) better model fitting: Our methods are based on widely accepted negative binomial (for RNA-Seq) model and Beta-binomial (for Methyl-Seq) for count data to better account for both biological and sampling variation, instead of Poisson (for RNA-Seq), Binomial (for Methyl-Seq) or Gaussian assumption (for both!); (2) genome-wide type I error control and power: We use genome-wide type I error control (FDR) and genome-wide power (EDR), which consider DE gene/DMR detection in the realm of whole genome, instead of a single gene/CpG region level; (3) better accuracy: Simulation and real data analysis demonstrate high accuracy of our methods; (4) optimal study design guidance: We consider cost-benefit analysis and the influence of both sample size and read depth on genome-wide power, which provides guidance for scientists decision making in five commonly encountered study design tasks(i.e. T1 to T5 in Section 2.3.2); (5) better utilization of pilot data: Our methods

perform self-learning and better utilizes the pilot data, compared to existing methods and there is no need to specify arbitrary fold change for DE gene/DMR detection or proportion of true DE genes/DMRs. To our knowledge, RNASeqDesign and MethylSeqDesign are the first statistical tools that comprehensively address the power calculation and study design issues for RNA-Seq and Methyl-Seq data respectively, with the key elements mentioned in Table 1. As the sequencing cost keeps dropping, RNA-Seq and Methyl-Seq experiments will become more and more prevalent and projects of large sample sizes will be expected. Thoughtful study design planning, including the five tasks included in this dissertation, will be essential. We believe these two methods will provide guidance for an economical and effective study design under a realistic setting.

The current framework needs a pilot dataset as an input for inference. If no pilot data exists from the same lab, one can seek existing public datasets with similar biological setting (e.g. similar tissue, disease or treatment). If possible, a recommended strategy is to perform a two-stage design by first generating suitable pilot data (e.g. n_0=n_1=6 with adequately deep sequencing). Then our methods can then help determine the optimal sample size and sequencing depth needed to achieve the optimal power under a certain budget. Alternatively, as a future work, when pilot data is not available, we can use a full parametric model by imposing distributions of parameters for DE genes/DMRs and non-DE genes/DMRs, e.g., mean counts, effect sizes, and dispersion parameters. Afterwards, our methods can work given these simulated parameters.

There are a few technical considerations and limitations in our model and performance evaluation. In RNASeqDesign, we consider problem of thousands of simultaneous hypothesis tests and we use EDR (defined as the proportion of true detected positives among all true DE genes) as the genome-wide power. While in RNASeqPower and other methods, they usually pursue statistical power or type I error control for a single test setting. Consequently, the power from different methods are not directly comparable although we have made best effort to match them. Secondly, the Wald test in RNASeqDesign is based on Gaussian approximation of the Z-statistics. Our simulation result has shown comparable performance of Wald test with the exact test used in edgeR. One possible reason of the good approximation and performance in Wald test could be the nature of sequencing, i.e., most genes have

large enough number of counts. Hence although the pilot sample size could be small, rich counts sufficed the normality approximation in Wald test. Finally, RNASeqDesign adopts the mixture model fitting on p-value distribution. It is possible that some applications may generate p-value distributions that deviate from the parametric mixture model. More sophisticated semi-parametric model fitting will be necessary for realistic power calculation.

There are several future directions from this dissertation. One is the case when there is no pilot data, which we mentioned above. Secondly, I would like to explore the possibility to apply this framework to other omics data. One example is microRNA, which is a small non-coding RNA molecule (only containing about 22 nucleotides). The function of microRNA is mostly in RNA silencing and post-transcriptional regulation of gene expression. Also, the aberrant expression of microRNA has been found associated with human diseases (e.g., chronic lymphocytic leukemia) (Musilova and Mraz, 2015). The second direction is for RNA isoforms (as a result of alternative splicing). RNA-Seq has the advantages over microarray to detect isoform-specifc expression levels. Many studies have shown that aberrant expression of some RNA isoforms are associated with certain diseases (Cooper et al., 2009; Scotti and Swanson, 2016). We expect in RNA-isoform data, sequencing depth will play even more important role since we need deep enough reads to discover the existent of specific isoforms. The challenges of this direction include the detection of RNA-isoforms and the following DE analysis.

We believe in the near future these experiments will become more and more prevalent and projects of large sample sizes will be expected. In other words, the demands of good sample size and power calculation tools with rigorous statistical framework will keep growing and we believe our approach has the potential to generalize to various types of data.

BIBLIOGRAPHY

Peter AC't Hoen, Yavuz Ariyurek, Helene H Thygesen, Erno Vreugdenhil, Rolf HAM Vossen, Renee X de Menezes, Judith M Boer, Gert-Jan B van Ommen, and Johan T den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, 36(21), 2008. ISSN 03051048. 10.1093/nar/gkn705.

Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine Garrett-Bakelman, Maria Figueroa, Ari Melnick, and Christopher Mason. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, 13(10): R87, 2012. ISSN 1465-6906. 10.1186/gb-2012-13-10-r87.

Allison, Gadbury, Heo, Fernandez, Lee, Prolla, Weindruch, Allison DB, Gadbury GL, Heo M, Fernandez JR, Lee C-K, Prolla TA, and Weindruch R. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39:1–20, 2002. ISSN 01679473. 10.1016/S0167-9473(01)00046-9.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. ISSN 1474-760X. 10.1186/gb-2010-11-10-r106.

Martin J. Aryee, Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014. ISSN 14602059. 10.1093/bioinformatics/btu049.

Stephen B Baylin. Dna methylation and gene silencing in cancer. *Nature Reviews. Clinical Oncology*, 2(S1):S4, 2005.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1): 289–300, 1995. ISSN 00359246. 10.2307/2346101.

Joshua S Bloom, Zia Khan, Leonid Kruglyak, Mona Singh, and Amy A Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC genomics*, 10(221):221, 2009. ISSN 1471-2164. 10.1186/1471-2164-10-221.

James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *U.C. Berkeley Div. Biostat. Pap. Ser.*, 11(1):94, 2009. ISSN 1471-2105. 10.1186/1471-2105-11-94.

Michele A. Busby, Chip Stewart, Chase A. Miller, Krzysztof R. Grzeda, and Gabor T. Marth. Scotty: A web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, 29(5):656–657, 2013. ISSN 13674803. 10.1093/bioinformatics/btt015.

Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13, 2016. ISSN 1474-760X. 10.1186/s13059-016-0881-8.

Thomas A. Cooper, Lili Wan, and Gideon Dreyfuss. RNA and Disease. *Cell*, 136(4):777–793, 2009. ISSN 00928674. 10.1016/j.cell.2009.02.011.

Xiangqin Cui and Gary A Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome biology*, 4(4):210, 2003. ISSN 1465-6914. 10.1186/gb-2003-4-4-210.

Yannick Delpu, Pierre Cordelier, William C. Cho, and Jérôme Torrisani. DNA methylation and cancer diagnosis, 2013. ISSN 16616596.

Ying Ding, Shaowu Tang, Serena G Liao, Jia Jia, Steffi Oesterreich, Yan Lin, and George C Tseng. Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics (Oxford, England)*, 30(22):3152–8, 2014. ISSN 1367-4811. 10.1093/bioinformatics/btu520.

Kevin Dobbin and Richard Simon. Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, 6(1):27–38, 2005. ISSN 1465-4644. 10.1093/biostatistics/kxh015.

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(1):215, 2014. ISSN 1471-2105. 10.1186/1471-2105-15-215.

Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010. ISSN 1471-2105. 10.1186/1471-2105-11-587.

Bradley Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004. ISSN 01621459. Doi10.1198/016214504000000089.

Bradley Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, 2007. ISSN 00905364. 10.1214/009053606000001460.

Manel Esteller. Aberrant DNA methylation as a cancer-inducing mechanism. *Annual review of pharmacology and toxicology*, 45:629–56, 2005. ISSN 0362-1642. 10.1146/annurev.pharmtox.45.120403.095832.

Zhide Fang and Xiangqin Cui. Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics*, 12(3):280–287, 2011. ISSN 14675463. 10.1093/bib/bbr004.

Hao Feng, Karen N Conneely, and Hao Wu. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69, 2014.

José A. Ferreira and Aeilko Zwinderman. Approximate Sample Size Calculations with Microarray Data: An Illustration. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006. ISSN 1544-6115. 10.2202/1544-6115.1227.

Gary L Gadbury, Grier P Page, Jode Edwards, Tsuyoshi Kayo, Tomas A Prolla, Richard Weindruch, Paska A Permana, John D Mountz, and David B Allison. Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research*, 13 (4):325–338, 2004. ISSN 09622802. 10.1191/0962280204sm369ra.

Yongchao Ge, Stuart C Sealfon, and Terence P Speed. Multiple testing and its applications to microarrays. *Statistical methods in medical research*, 18(6):543–563, 2009. ISSN 1477-0334. 10.1177/0962280209351899.

Kangxia Gu, Hon Keung Tony Ng, Lai Tang Man, and William R. Schucany. Testing the ratio of two poisson rates. *Biometrical Journal*, 50(2):283–298, 2008. ISSN 03233847. 10.1002/bimj.200710403.

Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13 (10):R83, 2012. ISSN 1474-760X. 10.1186/gb-2012-13-10-r83.

Thomas J. Hardcastle and Krystyna A. Kelly. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11:422, 2010. ISSN 1471-2105. 10.1186/1471-2105-11-422.

Steven N Hart, Terry M Therneau, Yuji Zhang, Gregory A Poland, and Jean-Pierre Kocher. Calculating sample size estimates for RNA sequencing data. *J Comput Biol*, 20(12):970–8, 2013. ISSN 1557-8666. 10.1089/cmb.2012.0283.

Paul J. Hurd and Christopher J. Nelson. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics and Proteomics*, 8(3):174–183, 2009. ISSN 14739550. 10.1093/bfgp/elp013.

Sin-Ho Jung and S Stanley Young. Power and sample size calculation for microarray studies. *Journal of biopharmaceutical statistics*, 22(1):30–42, 2012. ISSN 1520-5711. 10.1080/10543406.2010.500066.

Sin-Ho Jung, Heejung Bang, and Stanley Young. Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6(1):157–169, 2005. ISSN 1465-4644. 10.1093/biostatistics/kxh026.

Tiffany A. Katz, Serena G. Liao, Vincent J. Palmieri, Robert K. Dearth, Thushangi N. Pathiraja, Zhiguang Huo, Patricia Shaw, Sarah Small, Nancy E. Davidson, David G. Peters, George C. Tseng, Steffi Oesterreich, and Adrian V. Lee. 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(10):1000–1009, 2015. ISSN 19406215. 10.1158/1940-6207.CAPR-15-0178.

W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002. ISSN 1088-9051. 10.1101/gr.229102.

M Kathleen Kerr and Gary A Churchill. Statistical design and the analysis of gene expression microarray data. *Genetical research*, 89(5-6):509–514, 2007.

Kalimuthu Krishnamoorthy and Jessica Thomson. A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, 119(1):23–35, 2004. ISSN 03783758. 10.1016/S0378-3758(02)00408-1.

Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(4):555–572, 2005. ISSN 13697412. 10.1111/j.1467-9868.2005.00515.x.

Mei-Ling Ting Lee and George Alex Whitmore. Power and sample size for DNA microarray studies. *Statistics in Medicine*, 21(23):3543–3570, 2002. ISSN 02776715. 10.1002/sim.1335.

Adrian S Lewis and Michael L Overton. Nonsmooth optimization via BFGS. *Submitted to SIAM J. Optimiz*, pages 1–35, 2009. ISSN 0025-5610. 10.1007/s10107-012-0514-2.

Chung-I Li, Pei-Fang Su, Yan Guo, and Yu Shyr. Sample size calculation for differential expression analysis of rna-seq data under poisson distribution. *International journal of computational biology and drug design*, 6(4):358–375, 2013a.

Chung-I Li, Pei-Fang Su, and Yu Shyr. Sample size calculation based on exact test for assessing differential expression analysis in rna-seq data. *BMC bioinformatics*, 14(1):357, 2013b.

En Li, Caroline Beard, and Rudolf Jaenisch. Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–5, 1993. ISSN 0028-0836. 10.1038/366362a0.

Jun Li, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3): 523–538, 2012. ISSN 14654644. 10.1093/biostatistics/kxr031.

Ming D Li, Junran Cao, Shaolin Wang, Ju Wang, Sraboni Sarkar, Michael Vigorito, Jennie Z Ma, and Sulie L Chang. Transcriptome sequencing of gene expression in the brain of the hiv-1 transgenic rat. *PLoS One*, 8(3):e59582, 2013c.

Jonathan D. Licht. DNA Methylation Inhibitors in Cancer Therapy: The Immunity Dimension. *Cell*, 162(5):938–939, 2015. ISSN 10974172. 10.1016/j.cell.2015.08.005.

Peng Liu and JT Gene Hwang. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6):739–746, 2007. ISSN 1367-4803. 10.1093/bioinformatics/btl664.

John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008. ISSN 10889051. 10.1101/gr.079558.108.

Anastasios Markitsis and Yinglei Lai. A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, 26(5):640–646, 2010. ISSN 13674803. 10.1093/bioinformatics/btq001.

Alexander Meissner, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877, 2005. ISSN 03051048. 10.1093/nar/gki901.

Sayan Mukherjee, Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R Golub, and Jill P Mesirov. Estimating dataset size requirements for classifying DNA microarray data. *Journal of computational biology : a journal of computational molecular cell biology*, 10(2):119–142, 2003. ISSN 1066-5277. 10.1089/106652703321825928.

K Musilova and M Mraz. MicroRNAs in B-cell lymphomas: how a complex biology gets more complex. *Leukemia*, 29(5):1004–17, 2015. ISSN 1476-5551. 10.1038/leu.2014.351.

Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with

microarrays: A case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, 40(20): 10084–10097, 2012. ISSN 03051048. 10.1093/nar/gks804.

Marlies Noordzij, Giovanni Tripepi, Friedo W. Dekker, Carmine Zoccali, Michael W. Tanck, and Kitty J. Jager. Sample size calculations: Basic principles and common pitfalls. *Nephrology Dialysis Transplantation*, 25(5):1388–1393, 2010. ISSN 09310509. 10.1093/ndt/gfp732.

Fatih Ozsolak and Patrice M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, 2011. ISSN 1471-0056. 10.1038/nrg2934.

Grier P Page, Jode W Edwards, Gary L Gadbury, Prashanth Yelisetti, Jelai Wang, Prinal Trivedi, and David B Allison. The PowerAtlas: a power and sample size atlas for microarray experimental design and research. *BMC bioinformatics*, 7(1):84, 2006. ISSN 1471-2105. 10.1186/1471-2105-7-84.

Yongseok Park and Hao Wu. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453, 2016. ISSN 14602059. 10.1093/bioinformatics/btw026.

Yongseok Park, Maria E. Figueroa, Laura S. Rozek, and Maureen A. Sartor. MethylSig: A whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422, 2014. ISSN 14602059. 10.1093/bioinformatics/btu339.

Martina Paulsen and Anne C Ferguson-Smith. DNA methylation in genomic imprinting, development, and disease. *The Journal of pathology*, 195(1):97–110, 2001. ISSN 0022-3417. 10.1002/path.890.

Elena Perelman, Alexander Ploner, Stefano Calza, and Yudi Pawitan. Detecting differential expression in microarray data: comparison of optimal procedures. *BMC bioinformatics*, 8 (28):28, 2007. ISSN 1471-2105. 10.1186/1471-2105-8-28.

Magda E Price, Allison M Cotton, Lucia L Lam, Pau Farré, Eldon Emberly, Carolyn J Brown, Wendy P Robinson, and Michael S Kobor. Additional annotation enhances po-

tential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & chromatin*, 6(1):4, 2013. ISSN 1756-8935. 10.1186/1756-8935-6-4.

Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9):R95, 2013. ISSN 1465-6914. 10.1186/gb-2013-14-9-r95.

Mark Reimers. Making informed choices about microarray data analysis. *PLoS Computational Biology*, 6(5):1–7, 2010. ISSN 1553734X. 10.1371/journal.pcbi.1000786.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, 12(1):480, 2011. ISSN 1471-2105. 10.1186/1471-2105-12-480.

Keith D Robertson. Dna methylation and human disease. *Nature Reviews Genetics*, 6(8): 597–610, 2005.

Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321–332, 2008. ISSN 14654644. 10.1093/biostatistics/kxm030.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2009. ISSN 13674803. 10.1093/bioinformatics/btp616.

Mark D Robinson, Abdullah Kahraman, Charity W Law, Helen Lindsay, Malgorzata Nowicka, Lukas M Weber, and Xiaobei Zhou. Statistical methods for detecting differentially methylated loci and regions. *Frontiers in genetics*, 5:324, 2014.

Almut Schulze and Julian Downward. Navigating gene expression using microarrays a technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001. ISSN 1465-7392. 10.1038/35087138.

Axel Schumacher, Philipp Kapranov, Zachary Kaminsky, James Flanagan, Abbas Assadzadeh, Patrick Yau, Carl Virtanen, Neil Winegarden, Jill Cheng, Thomas Gingeras, and Arturas Petronis. Microarray-based DNA methylation profiling: Technology and applications. *Nucleic Acids Research*, 34(2):528–542, 2006. ISSN 03051048. 10.1093/nar/gkj461.

Marina M. Scotti and Maurice S. Swanson. RNA mis-splicing in disease. *Nature reviews. Genetics*, 17(1):19–32, 2016. ISSN 1471-0064. 10.1038/nrg.2015.3.

Jay Shendure. The beginning of the end for microarrays? *Nature Methods*, 5(7):585–587, 2008. ISSN 1548-7091. 10.1038/nmeth0708-585.

David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics*, 15 (2):121–32, 2014. ISSN 1471-0064. 10.1038/nrg3642.

Donna K. Slonim and Itai Yanai. Getting started in gene expression microarray analysis. *PLoS Computational Biology*, 5(10):e1000543, 2009. ISSN 1553734X. 10.1371/journal.pcbi .1000543.

Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–26, 2004. ISSN 1544-6115. 10.2202/1544-6115.1027.

John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(3):479–498, 2002. ISSN 13697412. 10.1111/ 1467-9868.00346.

John D. Storey and Robert Tibshirani. Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical report, Department of Statistics, Stanford University*, pages 9440–9445, 2001.

Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12):2213–2223, 2011. ISSN 10889051. 10.1101/gr.124321.111.

M Van Iterson, P Pedotti, GJEJ Hooiveld, JT Den Dunnen, GJB van Ommen, JM Boer, RX Menezes, et al. Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*, 10(1):439, 2009. ISSN 1471-2164. 10.1186/1471-2164-10-439.

Dan Wang, Li Yan, Qiang Hu, Lara E. Sucheston, Michael J. Higgins, Christine B. Ambrosone, Candace S. Johnson, Dominic J. Smiraglia, and Song Liu. IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, 28(5):729–730, 2012. ISSN 13674803. 10.1093/bioinformatics/bts013.

Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138, 2009. ISSN 13674803. 10.1093/bioinformatics/btp612.

Hao Wu, Chi Wang, and Zhijin Wu. PROPER: Comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*, 31(2):233–241, 2015. ISSN 14602059. 10.1093/bioinformatics/btu640.

Haiyuan Zhu and Hassan Lakkis. Sample size calculation for comparing two negative binomial rates. *Statistics in Medicine*, 33(3):376–387, 2013. ISSN 02776715. 10.1002/sim.5947.