GENOME-WIDE POWER CALCULATION AND EXPERIMENTAL DESIGN IN RNA-SEQ EXPERIMENT

by

Ge Liao

BS., Fudan University, China, 2010

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

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Ge Liao

It was defended on

June 27th 2014

and approved by

George C. Tseng, Sc.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Steffi Oesterreich, Ph.D., Professor, Department of Pharmacology Chemical Biology, School of Medicine, University of Pittsburgh

Eleanor Feingold, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Yan Lin, Ph.D., Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Yong Seok Park, Ph.D., Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: George C. Tseng, Sc.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

George C. Tseng, ScD

GENOME-WIDE POWER CALCULATION AND EXPERIMENTAL DESIGN IN RNA-SEQ EXPERIMENT

Ge Liao, PhD

University of Pittsburgh, 2014

ABSTRACT

Next Generation Sequencing (NGS) technology is emerging as an appealing tool in characterizing genomic profiles of target population. However, the high sequencing expense and bioinformatic complexity will continue to be obstacles for many biomedical projects in the foreseeable future. Modelling of NGS data not only involves sample size and genome-wide power inference, but also includes consideration of sequencing depth and count data property. Given total budget and pre-specified cost parameters such as unit sequencing and sample collection, researchers usually seek for a two-dimensional optimal decision.

In this dissertation, I will introduce a novel method SeqDEsign, which is developed to predict genome-wide power (EDR) of detecting differential expression (DE) genes in RNA-Seq experiment under targeted sample size (N') and read depth (R') given a pilot data (N,R). We aimed at providing advice for researchers regarding the design of RNA-Seq experiment with a limited budget.

The first part of this dissertation is about predicting genome-wide power at N' with R being fixed. The pipeline started with hypothesis test for differential expressed gene detection based on Wald test and negative binomial assumption. We proposed ways to directly model p-value distribution by both parametric and semi-parametric mixture model. To predict the genome-wide power of DE gene detection at N, posterior approaches based on either parametric or non-parametric model were implemented.

In the second part, we discussed ways to extend power prediction to N' and R' simultaneously. Both nested down-sampling (NDS) scheme and model-based (MB) method were proposed and compared. The three-dimensional EDR surface (Pow(N',R')) was constructed by two-way inverse power law model.

Finally, we discussed the cost-benefit analysis of RNA-Seq experiment with specification of a cost function. We also explored answers to other practical questions for experimental design. This framework was illustrated in both simulations and a real data application of rat RNA-Seq data.

The public health relevance of this work lies in the development of a novel methodology for genome-wide power calculation of RNA-Seq experiment. By accurately predicting genome-wide power, researchers can detect more biologically meaningful bio-markers, which will promote better understanding of human disease.

Keywords: Next Generation Sequencing(NGS), RNA-Seq data, Power calculation, Sample size, Mixture model, Cost-benefit analysis, Experiment design.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I want to thank my primary advisor Sc.D. George C. Tseng for his leadership and guidance. He introduced me to the field of genomic data analysis and inspired me to develop statistical methods to address important biomedical problems. As a mentor, he shared his rich academic experience with me and gave insightful suggestions for my career development.

My committee members, Dr. Steffi Oesterreich, Dr. Eleanor Feingold, Dr. Yan Lin and Dr. Yong Seok Park, were amazingly constructive in advice and accommodating in schedule. They provided insightful and critical suggestions that improved many aspects of this work. I especially want to thank Dr. Steffi Oesterreich, who have led me into breast cancer research. I also want to address the many practical suggestions Dr. Yong Seok Park had given to me for this work.

Sincere thanks also to all the peer students in our lab during the past four years. We have learned a lot from each other and collaborated to develop various statistical methods. And thanks to all the statistic staff and students in our department for the encouragement and sharing.

Finally, I want to thank my family for the long lasting support, especially my husband, Dr. Yang Feng. No matter what, he is always the one who stands by me and encourages me through the happiness and suffering.

1.0 INTRODUCTION

With the advances in robotics and the vastly available genomic information on public databases, microarray technology gained tremendous popularity for its high-throughput quantitative representation and cost-effectiveness since last decade[Reimers, 2010]. While microarray experiment provides access for biologist to a range of applications, statistical analysis has played an active and significant role in the whole process. Statisticians correspondingly, have taken enthusiastic interests in developing statistical tools that led to more profound biological interpretations for certain research questions[Slonim and Yanai, 2009, Keer and Churchill, 2001]. The next-generation sequencing, which is based on random amplification and shotgun sequencing, is another revolutionary technology first came to market in 2004, making genomic profiles available in much higher resolution and in extremely high parallel[Fang and Cui, 2011]. Although errors and biases might be involved in major steps of experimental preparation processes, next-generation sequencing has been hailed as the future of genomic research because of its higher sensitivity and potential of generating unlimited dynamic ranges. In this sense, research is gradually transiting from microarray technologies to next-generation sequencing[Shendure, 2008]. From statistical point of view, some methodologies developed under the microarray context may still be extended to NGS, however we are facing many new challenges in data analysis.

In a biological study, the procedure of exploring a research topic usually starts from an experimental 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 literature[Lee and Whitemore, 2002, Gadbury et al., 2004]. But methods for sequencing data are very limited. As sequencing technology is still not quite affordable to the majority of researchers, it's significant to ensure desirable power of bio-marker detection (usually a Differential expression(DE) analysis) in the earlier phase of study (herein called "pilot study").

In this introduction section, we will first clarify the significance of quantifying gene expression and go over both traditional and novel biotechnology. After that, we will introduce the structure of a typical gene expression data and review methods of differential expression analysis. Furthermore, we will distinguish between traditional power and genome-wide power definition and review the existing methods for microarray and RNA-Seq data power calculation. Finally, the major motivation of developing new methods will be addressed.

1.1 QUANTIFICATION OF GENE EXPRESSION

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. There are more than 20,000 genes in human genome, and only a small fraction of them are actually expressed in certain cell types and at certain times. Understanding the dynamic changes of gene expression of a given subject is important for us to study biological process ranging from inflammation to human aging. By comparing gene expression data between different groups of subjects, we can explain the activation or deactivation of pathways and the heterogeneity of diseases. Now, the question comes that how we can quantify the gene expression. The techniques for quantification of gene expression could be categorized into two types: (1) Candidate gene transcriptome profiling; (2) High-throughput transcriptome profiling.

1.1.1 Candidate gene transcriptome profiling

The candidate gene approach starts with a given gene list of interest, ranging from hundreds to thousands. Comparing with high-throughput transcriptome profiling, it requires additional expert knowledge and consequently creates bias in the study.

Conventional method for mRNA quantification by this approach is electrophoretic techniques, including northern blot(1977), which is still the benchmark against other techniques. It has advantages that the size of transcript is obtained by gel electrophoresis and it allows for the identification of splicing variant which may be present. However, northern blots is tedious and its sensitivity is limited by the capacity of the gel[Roth, 2002].

Later in 1989, reports of real-time PCR (rt-PCR) experiments for transcriptome analysis were published[Burg et al., 1989]. It is generally accepted that rt-PCR produces the most superior quantitative data due to the exquisite sensitivity and specificity of the PCR. For many cDNA microarray data, rt-PCR could serve as the validation instead of the other way around. While the advantage of routine microchip array is the large number of genes that could be profiled simultaneously, they also suffers from high batch effect, poor sensitivity and specificity. In comparison, though rt-PCR can only profile a smaller number of genes ($100 \sim 400$), the quality of quantification is much more desirable[Schmittgen et al., 2008]. Comparing with northern blot, the assay is far more quantitative, allowing more accurate measurements of mRNA amounts.

1.1.2 High-throughput transcriptome profiling

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

1.1.2.1 Hybridization based approaches(Microarray) With the increased popularity of high throughput technology in mid 90's, microarray became a novel tool in quantifying genomic changes. The ability of these arrays to simultaneously interrogate thousands of transcripts had led to important advances in a wide range of biological problems. These advances include the identification of gene expression differences between disease 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 sequence, 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 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) assay is limited to transcript with relevant probes[Marioni et al., 2008].

1.1.2.2 Sanger sequencing of cDNA or EST libraries Sanger sequencing, developed by Frederich Sanger and colleagues in 1977 [Sanger and Coulson, 1975], is a method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro replication. It is more desirable to be used for projects with smaller-scale requirement for gene expression quantification.

1.1.2.3 Serial Analysis of Gene Expression (SAGE) Serial Analysis of Gene Expression(SAGE) was used to overcome the disadvantage of low throughput, expensive and non-quantitative in Sanger sequencing. This tag-based sequencing method can provide precise and digital gene expression levels. They have, however, shortcomings, such as that, a significant portion of the short tags cannot be uniquely mapped to the reference genome and only partial transcripts are covered, etc. These disadvantages have largely limited the application of traditional sequencing technologies[Wang et al., 2009b].

1.1.2.4 RNA-Seq With the advent of rapid Next Generation Sequencing(NGS) technology with reduced cost, RNA-Seq was recently developed to characterize the transcriptomic profiling, impacting almost every field in life science and is being applied for clinical use. In general, RNA sample is converted to a library of cDNA fragments with adaptors attached to one or both ends. Then, each RNA molecule is sequenced in a high-throughput manner to obtain short sequences for one or both end. Once reads with high quality have been retrieved after preprocessing, the next step for quantification of gene expression is the alignment to reference genome. Consequently, we can compute the number of reads that have been aligned to each gene region[Wang et al., 2009b]. Comparing with microarray platforms, RNA-Seq technology has many advantages. First, RNA-Seq can cover the transcription of whole genomic region unbiasedly comparing with hybridization-based approaches. Secondly, it renders single-base resolution in quantification of gene expression and therefore reveals the precise location of transcription boundaries. Thirdly, it provides more information regarding the alternative splicing to improve our understanding of genomic transcription. Furthermore, since there's no control probe to compared with as it is in microarray, background signal will not be an issue to reduce accuracy. Different from the response variable in microarray, which is continuous intensity, RNA-Seq data is consist of aligned read count for each gene. Since RNA-Seq does not have an upper limit for quantification, there will be a larger dynamic range it could cover. RNA-Seq has also been shown to have high accuracy and reproducibility from previous studies[Marioni et al., 2008]. Table 1 compared the advantages and disadvantages of major biotechnologies in the quantification of gene expression.

Table 1: Comparison between high-throughput technology in transcriptome profiling

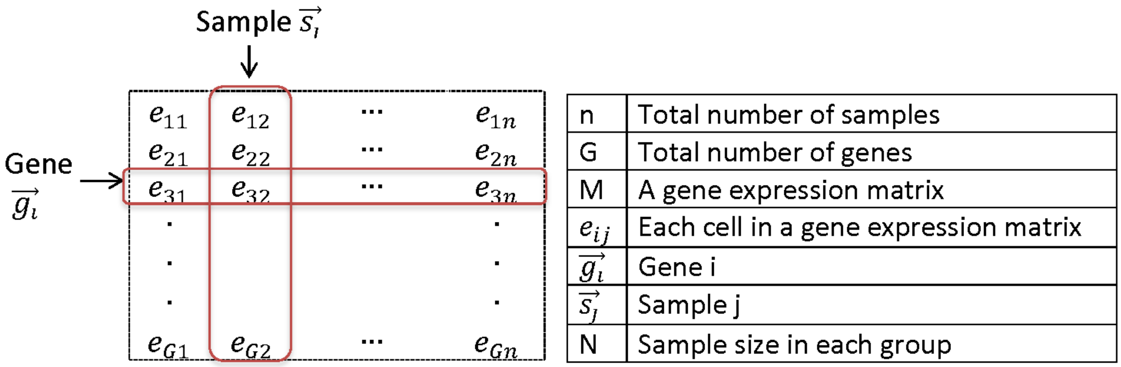
Technology	Microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	$\geq$ 8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

1.2 DATA STRUCTURE OF MICROARRAY AND RNA-SEQ EXPERIMENT

A genomic study typically assesses a large number of DNA sequences under multiple conditions, e.g., a collection of different tissue samples. The resulting data after proper preprocessing 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 gens, the columns ($S = \{\overrightarrow{s_{11}}, ..., \overrightarrow{s_{1n_1}}, ..., \overrightarrow{s_{kn_k}}\}$) represent the expression profiles of samples, and each cell e_{gij} is the measured expression level of gene g in sample j of group i. We assume that the genomic study has a balanced design. If there are two groups of interest, there are $n_1 = n_2 = N$ samples in each experiment condition. In other words, subjects $\{\overrightarrow{s_{11}}, ..., \overrightarrow{s_{1n_1}}\}$ are in group A with $x_j = 1$, while subjects $\{\overrightarrow{s_{21}}, ..., \overrightarrow{s_{2n_2}}\}$ are in group B with $x_j = 0$. See Figure 1 for the detailed illustration.

In microarray study, e_{gij} is log2 of raw intensity or intensity ratio, which is typically modeled as a continuous variable. For RNA-Seq data, e_{gij} is read count of gene g of subject j in group i. And e_{gij} is frequently observed to follow over-dispersed poisson distribution. [McCarthy et al., 2012]

In both technologies, sample size is the most influential factor in the determination of power. Assuming we are primarily interested in detecting differentially expressed genes(DE gene) between two groups, here we formally define the sample size (N) as the number of biological replicates in each group. For RNA-Seq data, read(or sequencing) depth, which is proportional to total aligned reads(R), is another important factor that impact power calculation[Rapaport et al., 2013]. It is clear that higher read depth generates more informational reads, which will increase statistical power to detect DE genes. Throughout this thesis, we refer to read depth and coverage inter-changeably both meaning how many reads are assigned to a particular gene.



$$\begin{array}{cccc} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ e_{31} & e_{32} & \cdots & e_{3n} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ e_{G1} & e_{G2} & \cdots & e_{Gn} \end{array}$$

n	Total number of samples
G	Total number of genes
M	A gene expression matrix
e_{ij}	Each cell in a gene expression matrix
$\vec{g_i}$	Gene i
$\vec{s_j}$	Sample j
N	Sample size in each group

(a) A gene expression matrix; (b) Notations in this paper.

Figure 1: Data structure of genomic study.

1.3 BIOMARKER DETECTION IN MICROARRAY AND NGS DATA

The most important reason to generate gene expression data regardless of specific platform is to identify the DE genes in two or more conditions. Such genes are usually detected based on a combination of expression change threshold and p-value cutoff[Rapaport et al., 2013].

Sandrine and others [Dudoit et al., 2003] provided a comprehensive review for the statistical issues that are addressed in microarray gene expression data. Elena and others [Perelman et al., 2007] compared several alternative methods including t-test, modification of t-test(significance analysis model, SAM) for differential expression analysis. In the limma package, an empirical Bayes approach is implemented that employs a global variance estimator s_0^2 computed on the basis of all genes' variances[Smyth, 2004]. These methods are all based on Gaussian assumption for log transformed gene expression e_{ij}.

Due to different characteristics, methods for detecting DE gene from RNA-Seq data is more complicated and diverse. The methods can be splitted into three major categories:

(1) Method based on Gaussian assumption: Bloom and others[Bloom et al., 2009] applied the t-test to the total-count normalized data. Peter and others['t Hoen et al., 2008] performed

a square-root transformation for the total-count normalized data to stabilize the variance and applied a t-test afterwards. In DEGseq proposed by Wang and others[Wang et al., 2009a], it is assumed that log ratios of the counts have a normal distribution and a z-score is calculated.

(2) Methods based on Poisson assumption: Marioni and others[Marioni et al., 2008] proposed a Poisson log-linear model and performed likelihood ratio test (LRT) for differential expression gene detection. They applied normalization based on total-count implicitly. Bullard and others [Bullard et al., 2010] took an external quantile normalization step rather than doing total-count normalization. "Poissonseq" method [Li et al., 2012] is based on Poisson log-linear model, and can be used to not only two-class outcome but also multiple-class and even quantitative outcome.

(3) Methods based on negative binomial assumption: Generalized linear model based on a negative binomial distribution has also been developed in order to deal with over-dispersed counts in RNA-Seq data. Robinson and others[Robinson et al., 2010] developed edgeR by extending from previous methods for SAGE data. In their method, the dispersion parameter can be estimated for each gene or can be assumed to be common across all genes, making this method quite flexible. DESeq, developed by Anders and Huber[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] was proposed based on empirical Bayes theory. NOISeq[Tarazona et al., 2011] differs from previous methods in that it is data-adaptive and nonparametric, and consequently better adapts to the size of the data set.

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, sensitivities as well as good control of false positive errors. Intawat and others [Nookaew et al., 2011] found that edgeR could uniquely identify more differential gene expression(DGE) than Cuffdiff, baySeq, DESeq and NOISeq.

1.4 SAMPLE SIZE, POWER, GENOME-WIDE POWER

The design issues in microarray experiment had been broadly discussed in [Kerr and Churchill, 2001, Simon and Dobbin, 2003]. Generally, design problems cover level of replication, reference design, the balanced block design, the loop design and so on, which are proposed to address different research questions. One of the most common tasks of statistician requested by investigators is the sample size and power calculation. In general, sample size is the number of patients or other experimental units enrolled in a study, and is usually referred as biological replicates.

In order to calculate the sample size, it is required to have some idea of the results expected in a study. In general, the greater the variability in the outcome variable, the larger the sample size is required to assess whether an observed effect is a true effect. On the other hand, the more effective a tested treatment is, the smaller the sample size is needed to detect this positive or negative effectNoordzij et al. [2009].

Traditional definition of power is based on the framework of one hypothesis testing. Assume we are interested in testing $H_0 : \mu_A - \mu_B = 0$ against $H_1 : \mu_A - \mu_B = 2$, where A and B are two different treatment groups, both of which have the same number of subjects. To achieve a statistical power of $1-\beta$, the sample size needs to be $n = \frac{(s^2_{\bar{Y}_A - \bar{Y}_B})(z_\alpha + z_\beta)^2}{(\bar{Y}_A - \bar{Y}_B)^2}$, where α and β denotes for type I and type II error respectively. $\bar{Y}_A - \bar{Y}_B$ is usually defined as effect size, indicating the difference between two groups of interest. $s^2_{\bar{Y}_A - \bar{Y}_B}$ is the variability of group difference.

Data generated from genomic study are constitute of more than 20,000 probes or genes. In this large-scale simultaneous hypothesis testing problem, with hundreds of cases considered together, we can quantify the power of detecting genomic changes by "genome-wide power". It was also referred as expected discovery rate(EDR) in Gary and others' paper [Gadbury et al., 2004]. Under this framework, an important question is how we can maintain type I error since there are multiple comparisons. Family-wise error rate and False discovery rate(FDR)[Benjamini and Hochberg, 1995] are widely used to address this problem.

Suppose we construct the following two by two contingency table with specific G hypotheses to be known in advance. The numbers G_0 and G_1 of false and true null hypotheses are unknown parameters, R is an observable random variable and A_0, A_1, R_0, R_1 are unobservable random variables. In the context of microarray experiment, we would like to minimize the number R_0 of false positives[Efron, 2007, Ge et al., 2009]. In this case, genome-wide power is defined as $EDR = E(\frac{R_1}{G_1}) = 1 - \beta_1$. FDR is defined as $FDR = E(\frac{R_0}{R})$. In most genomic applications, one controls FDR under a certain pre-specified threshold (e.g., FDR=0.05) to obtain the DE gene list. In the power calculation method throughout this thesis, we pursue genome-wide power (EDR)under pre-specified FDR control.

Table 2: 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

1.5 COST OF RNA-SEQ EXPERIMENT

The cost of RNA-seq experiments often limits RNA-seq studies to only a small number of replicate libraries. Many methods developed have also limitation when being applied to small sample size. On May.28 2014, we checked the pricing of RNA-Seq experiment in the NGS services at MD Anderson center. Sample preparation costs 572.30 dollars per sample. Sequencing cost for HiSeq 2000 100bp pair-end reads is 2771.55 dollars per lane for external user. For each lane, usually 300 - 400M paired-end reads could be generated which takes 11 days according to Duke institute for genome science and policy. That means, assuming the alignment rate is 50%, One can generate $\sim$ 650X coverage of RNA-Seq data if running a sample per lane. The cost will be (572.30+2771.55)$\times$ 5 $\approx$ \$ 16700 for five samples at 650X. Alternatively, one can tag two samples per lane and run ten samples at $\sim$ 325X for 572.30 $\times$ 10 +277.1$\times$5$\approx$ \$ 19600. In our method, we will consider both sequencing cost and sample preparation cost in the cost function.

1.6 EXISTING SAMPLE SIZE AND POWER CALCULATION METHOD

1.6.1 The use of pilot study in power calculation

In general, power calculation method could be based on information from a pilot study or purely model-based. Model-based methods is straight-forward and more economical, while in most cases, it will give unrealistic estimation of sample size and power. By conducting pilot study, we can estimate variability and effect size from pilot data to infer proper sample size and power. Especially for genomic data, pilot study is of greater importance, since there are variability coming from biological replicates, technical replicates, experiment and batch effects.

In this thesis, we assume a pilot study with sample size N and total read R 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 part.

1.6.2 Existing Methods for microarray data Sample Size Calculation

The significance of performing power and sample size calculation for genomic data was firstly addressed by Mei-Ling Ting Lee.[Lee and Whitemore, 2002] They started from a common setting of ANOVA model in microarray data analysis:

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

where $l = 1, ..., L$ denotes a set of L experimental factors. Parameter $\gamma_l(b_l)$ denotes a main effect for factor l when it has level b_l, for $l = 1, ..., L$, respectively. With the purposed hypothesis test $H_0 : I_g = 0$ against $H_1 : I_g = I^d$ for main effects, a t-statistics, F-statistics, χ^2 or z-statistics could be constructed for single gene under different study designs. Here $I_g(= I_{gc})$ is denoted as the effect of a covariate/condition c for gene g, and the non-zero vector I^d in H_1 is a target vector of differential expression levels that is expected to detect. In their paper, ways to control multiple comparison were discussed when genes are correlated and not correlated. Microarray studies usually involve simultaneous test of thousands of genes, therefore the probability of producing incorrect conclusions must be controlled.

Family-wise error rate(FWER), $(\alpha_F = P(R_0 > 0))$, is discussed in details for application in multiple comparison issues in Mei-Ling Lee's paper. Both(1)Sidak approach: assuming independent estimation errors; and (2)Bonferroni procedure: assuming dependent estimation errors are considered. It is obvious that the approach proposed there doesn't consider the heterogeneity across genes, since they assumed all genes have same variance and same effect size for alternative hypothesis. They also mentioned the possibility to solve 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. But this approach is not investigated enough until the methodological paper for PowerAtlas[Gadbury et al., 2004].

PowerAtlas is a popular web tool for power and sample size calculation proposed by Gary L Gadbury and others [Page et al., 2006]. They considered the variability for mean expression and effect size across genes by directly modeling p-value distribution of test statistics. They introduced the concept of expected discovery rate(EDR), which could be viewed as the average power for all genes with null hypothesis being false. Since microarray studies are influenced by multiple comparison issues, they also defined true positive rate(TP), true negative rate(TN) as shown below. The notations are consistent with Table 2.

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

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

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

Assuming we have c being a binary covariate, with 0 as control and 1 as case. If μ_{iA} and μ_{iB} are underlying true expression for group A and B, we want to test whether the expression of group A and B are different for the ith gene with $H_0 : \mu_{iA} - \mu_{iB} = 0$ against $H_0 : \mu_{iA} - \mu_{iB} \neq 0$. In their power calculation procedure, they started from a set of p-values of t-statistics to test for differential expression based on a pilot data. This pilot data should have similar experimental characteristics as the future data. It could either be generated in a pilot study or from a public database. t-statistics could 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 $S_{e_{iA}e_{iB}} = \frac{(n_A-1)S^2_{e_{iA}}+(n_B-1)S^2_{e_{iB}}}{n_A+n_B-2}$ and $n = n_A + n_B$., assuming equal variance. When the two groups have equal sample size $n_A = n_B = N$, test statistics reduces to:

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

With the assumption that p-value distribution by DE analysis from microarray experiment is a mixture of beta component and uniform component, a mixture model is fitted with p-values from $t_i (i = 1, ..., G)$. The fitted model is $f^*(p)$. Then a parametric bootstrap procedure is performed to obtain new set of p-values with a targeted sample size N'. Their key step is the transformation of t-statistics by:

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

The underlying assumption is that the group difference of each gene remains the same under different sample size. By directly modeling p-value distribution, the heterogeneity across genes could be maintained. Besides, it is more suited for this task since thousands of hypotheses are tested in a discovery oriented research[Gadbury et al., 2004]. However, their method cannot be directly apply to RNA-Seq data and they didn't control FDR at fixed level by imposing arbitrary p-value cut-off instead.

1.6.3 Existing Methods for RNA-Seq data Sample Size Calculation

The greatest distinction between RNA-Seq and microarray gene expression data is the types of expression values. Microarray has continuous intensity, while RNA-Seq data is in read count for each gene. As a consequence, their distribution assumption differs: Gaussian assumption for microarray(usually at log-intensity level), and poisson/negative binomial distribution for NGS data.

In the typical DE gene analysis, we want to compare expression level of two experimental groups, e.g., tumor and normal groups. it is equivalent to test $H_0 : \mu_{iA} = \mu_{iB}$ or $H_0 : \frac{\mu_{iA}}{\mu_{iB}} = 1$ against $H_1 : \mu_{iA} \neq \mu_{iB}$ or $H_1 : \frac{\mu_{iA}}{\mu_{iB}} \neq 1$.

1.6.3.1 Method based on Poisson assumption Many literature has discussed various Poisson tests for this hypothesis testing: (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, and corresponding power were calculated. See [Gu et al., 2008] for a comprehensive review for poisson rate tests. Chung-I Li and others [Li et al., 2013a] developed methods(we will call it "Poisson model" for later reference) for sample size and power calculation, based on previous introduced poisson tests. They used false discovery rate(FDR) for multiple comparison, which was originally proposed by [Storey and Tibshirani, 2001, Storey, 2002].

1.6.3.2 Methods based on negative binomial assumption

Poisson tests are widely used, while it ignores the nature of over-dispersion in real sequencing data. We have reviewed methods to detect DE genes based on over-dispersed poisson model. Among them, edgeR[Robinson and Smyth, 2008] and DEseq[Anders and Huber, 2010] have been two most popular packages to perform DE analysis. Extensive comparative studies have shown the superiority of these two tests in detecting biomarkers over other tests. But it is obvious that the two exact tests don't have a closed form for sample size and power calculation.

Until now, there are two methods for RNA-Seq power calculation proposed: (1)RNASeqPower[Hart et al., 2013]; (2)Method based on exact test.[Li et al., 2013b] The two method are similar in that they both require estimation or pre-specification of fold change, mean counts, coefficient of variation and 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 testing target(typically fold change or effect size). These three parameters are required to be fixed across genes or a given study, and are often determined by external requirements. μ and σ, which are coverage and coefficient of variation(CV) between biological replicates are gene specific. The derivation of this formula is based on a generalized linear model framework and is presented in their paper. In their paper, the test is only limited to single gene level. CV is estimated by edgeR. ($\sigma = \frac{1}{\sqrt{\delta}}$,where δ is the dispersion parameter) When considering gene collections, they simply take $\sigma_{0.60}$ (60% quantiles of CV as the overall CV) and quantile of depth distribution across gene for sample size calculation. Currently, they have an R package(RNASeqPower) available. Though this method is pretty straightforward, it has several disadvantage: (1) it does not consider multiple comparison issue since the the power is only computed based on one single test; (2) it fails to incorporate the variability across genes and instead uses summary statistics for effect size, dispersion, coverage of each gene, etc.

Chung-I Li and others [Li et al., 2013b] proposed a method for power calculation based on exact test. For single gene case, power could be calculated by:

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

where $\omega = \frac{d_1^*}{d_0^*}$ 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 proabablity mass function of negative binomial model with mean μ and dispersion δ.

Considering collections of genes, they provide two approaches. In the first approach, μ_{iA}, ρ_i, δ_i can be estimated from pilot data for each prognostic gene g that are know. Then we could use numerical method to solve the equation:

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

where α^* is the type I error when FDR is controlled at f. In the second approach, we can specify a desired minimum fold change ρ^*, a minimum average read count μ_{iA} and a minimum dispersion δ_i

This method has advantages to provide ways in account for across genes heterogeneity. However, the parameter setting is also arbitrary and not flexible enough.

1.6.3.3 Method based on Gaussian assumption

"Scotty"[Busby et al., 2013] is another tool recently developed for interactive power calculation. It first assesses number of reads required to measure a specific number of genes, then estimate the within group variance. After that, Scotty will test a matrix of different experimental designs. Finally, the design with highest power under a user-specified parameter will be selected. The parameters include number of biological replicate, read depth and cost. While Scotty provides novel ways in the experimental design for RNA-seq experiment, the framework is established based on Gaussian assumption. Statistical power is calculated based on a t-test. They argued that by using t-test unbiased calls of differential expression will be produced and power formula for t-test is readily available for the computation.

To validate the prediction accuracy of power in Scotty, the authors compared Scotty with DESeq using simulated data. They found that when sample size is small(N=2), DESeq has more power in detecting DE genes, while sample size increases to greater than 5, t-test have slightly greater power in detection. However, the papers did not evalutate the accuracy of power prediction and selection of optimal experiment configuration since they didn't have a true power surface to compared with in their simulation studies. Furthermore, they didn't take consideration of multiple testing issue when predicting genome-wide power.

We compare the four existing methods with respect to six different perspectives in Table 3. According to our observation, not a single existing method can accommodate requirements for all these criterion and in this thesis, we will develop "SeqDEsign", a method to perform power prediction and provide practical recommendation for the optimal experimental design under a certain constraint of budget.

Table 3: Advantage and disadvantage of existing methods

Method characteristics	Poisson Model by Li (2013)	RNASeqPower by Hart (2013)	Exact test by Li(2013)	Scotty by Busby (2013)	SeqDEsign
Use negative binomial assumption		✓	✓		✓
Use pilot data	✓	✓	✓	✓	✓
Consider sequencing depth		✓		✓	✓
Consider multiple comparison (FDR)	✓		✓		✓
Apply genome wide power calculation				✓	✓
Consider cost function by N and R				✓	✓

1.7 MOTIVATION OF SEQDESIGN

In all, unlike power calculation of traditional microarrays, modeling NGS data not only involves sample size and genome-wide inference, but also includes sequencing depth and count data statistics. The optimal design is beyond one-dimensional dual problem between sample size versus statistical power as in the traditional array scenario. Given a total budget and pre-specified cost factors such as unit sequencing expense (e.g. sequencing cost per million reads), sample collection cost and bioinformatics expenditure, researchers usually seek a two-dimensional optimal decision by balancing between the number of samples and sequencing depth, yet existing methods have intrinsic limitations. We hypothesize that:

Using advanced count-data probabilistic modeling and power calculation, balancing between sample size and sequencing depth under a fixed total budget will provide optimal genome-wide statistical power to detect differentially expressed genes.

In this method, the optimal design involves two-dimensional factors of sample size and sequencing depth under the constraint of a realistic cost schedule, involving per-unit sequencing and sample collection, etc. As a result, the solution and interpretation from optimal design and cost-benefit analysis are readily applicable to a real-world lab setting.

The following chapters will be arranged as described below. In chapter 2, we will first investigate the proposed power prediction method for sample size when sequencing depth is fixed. That means, for pilot data with sample size N and fixed read depth R, what's the genome-wide power in DE gene detection if N is increased to N'? We will introduce our algorithm step by step and demonstrate the accuracy of our method in a simulation study with parameters estimated from real data and compare to four existing methods. Furthermore, we applied the method to real data and evaluate the accuracy of predicted EDR curve. In chapter 3, we develop methods to include the varying read depth in the prediction of EDR. A three dimensional EDR surface can be consequently constructed. In chapter 4, we will discuss the design of cost function, optimization of EDR given cost function and provide a series of case studies for cost-benefit analysis. Furthermore, the predicted optimal design will be compared with underlying true optimal design. Finally, in chapter 5, we will discuss several major issues and limitations of our methods and the future directions.

2.0 SAMPLE SIZE AND GENOME-WIDE POWER(EDR) PREDICTION

As described before, genome-wide power could be modeled as an increasing function of sample size. For RNA-seq data, many power calculation methods[Li et al., 2013a,b] were established based on this assumption. In this chapter, we firstly investigated the influence of sample size on the estimation of genome-wide power, namely EDR, in our proposed methods and then compared with existing methods in simulation studies. We further applied our proposed method in real data applications.

2.1 HYPOTHESIS TESTING FOR DE GENE DETECTION

In chapter 1, we reviewed methods for DE gene detection under different models. There are advantages and disadvantages for each method. Yet, most of them could not be directly applied to predict power of a target sample size N' given a pilot data with sample size N in the case of RNA-Seq data.

Here we proposed the application of Wald test based on generalized linear model for power calculation. The statistical formulation has been discussed in detail in a recent paper[Zhu and Lakkis, 2013]. Following our notation in chapter 1, denote by e_{gij} the read counts of gene $g(g = 1, ..., G)$ for subject $j(j = 1, ..., n_i)$ of group $i(i = 1, 0)$. For example, i=0 represents the control and n_0 is the number of controls; i=1 for the case and n_1 for number of cases. Assume e_{gij} follows a negative binomial distribution with mean μ_{gij} and common dispersion parameter δ, e_{gij} has the following probability mass function:

$$P(e_{gij}) = \frac{\Gamma(\delta + e_{gij})}{\Gamma(\delta)e_{gij}!}(\frac{\delta^{-1}\mu_{gij}}{1+\delta^{-1}\mu_{gij}})^{e_{gij}}(\frac{1}{1+\delta^{-1}\mu_{gij}})^{\delta} \tag{2.1}$$

where $\Gamma(\cdot)$ is the gamma function. Link function for negative binomial regression is:

$$log(\mu_{gij}) = log(R_{ij}) + \beta_{g0} + \beta_{g1}x_{ij} \tag{2.2}$$

where R_{ij} is assumed to be the total mappable reads for subject j in group i and x_{ij} is a indicator variable(subjects x_{1j} comes from case group(i=1) and subjects x_{0j} comes from control group(i=0)). From previous two formula, we can easily get the log-likelihood function:

$$L_g = \sum_{i=0}^{1}\sum_{j=1}^{n_i}[log\frac{\Gamma(\delta + e_{gij})}{\Gamma(\delta)e_{gij}!} + e_{gij}log(\delta^{-1}\mu_{gij}) - (e_{gij}+\delta)log(1+\delta^{-1}\mu_{gij})] \tag{2.3}$$

β_{g0} and β_{g1} could be estimated by standard maximum likelihood(ML) methods with proper initial points. Variance-covariance matrix of the MLE estimates could be approximated by inverse expected Fisher information.

For gene g, our goal is to test null hypothesis $H_0 : \beta_{g1} = 0$ against $H_1 : \beta_{g1} \neq 0$. Based on likelihood theory, asymptotic tests, including likelihood ratio test, score test and Wald test, can be constructed. Among the three tests, only Wald test statistics could be written in a closed form. The advantage of Wald test is the convenience for Z statistic transformation assuming effect size of treatment remains the same for each gene for different sample sizes.

Specifically, Wald test could be constructed by:

$$Z_g = \frac{\hat{\beta}_{g1}}{\sqrt{Var(\hat{\beta}_{g1})}} \sim N(0,1) \tag{2.4}$$

under the null hypothesis for gene g.

$$\begin{pmatrix} \sum_{i=0}^{1}\sum_{j=1}^{n_i}[e_{gij} - (e_{gij}+\delta) \times \frac{q_{ij}}{(1+q_{ij})}] \\ \sum_{i=0}^{1}\sum_{j=1}^{n_i}[x_{ij}e_{gij} - (e_{gij}+\delta) \times \frac{x_{ij}q_{ij}}{(1+q_{ij})}] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{2.5}$$

where $q_{ij} = \delta^{-1}R_{ij}e^{\beta_{g0}+\beta_{g1}x_{ij}}$ and $\binom{\hat{\beta}_{g0}}{\hat{\beta}_{g1}}$ can be estimated by solving equation (2.5) by numerical methods.

For purpose of power calculation, we assume all samples have the same total reads R, $R_{ij} = R \; \forall$ i, j . When samples do not have equal total reads, normalization methods could be applied to meet this requirement. Under this assumption, the expected fisher information can be simplified as:

$$F_g = -E(\frac{\partial^2 L_g}{\partial(\beta_{g0}, \beta_{g1})}) = \begin{pmatrix} \frac{n_0 Re^{\beta_{g0}}}{1+\delta^{-1} Re^{\beta_{g0}}} + \frac{n_1 Re^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1} Re^{\beta_{g0}+\beta_{g1}}} & \frac{n_1 Re^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1} Re^{\beta_{g0}+\beta_{g1}}} \\ \frac{n_1 Re^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1} Re^{\beta_{g0}+\beta_{g1}}} & \frac{n_1 Re^{\beta_{g0}+\beta_{g1}}}{1+\delta^{-1} Re^{\beta_{g0}+\beta_{g1}}} \end{pmatrix} \tag{2.6}$$

$$Cov\begin{pmatrix} \hat{\beta}_{g0} \\ \hat{\beta}_{g1} \end{pmatrix} = F_g^{-1}(\hat{\beta}_{g0}, \hat{\beta}_{g1}) \tag{2.7}$$

And consequently,

$$Var(\hat{\beta}_{g1}) = \frac{1}{n_0} \times (\frac{1 + \theta e^{\hat{\beta}_{g1}}}{\theta Re^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{1+\theta}{\theta\hat{\delta}}) \tag{2.8}$$

where θ is $\frac{n_1}{n_0}$.

Here, the dispersion parameter is assumed to be common across all genes and is estimated by conditional maximum likelihood in "edgeR"[Robinson and Smyth, 2008]. The reason for using common dispersion parameter is that when sample size is small, (which is usually the case in pilot studies) estimating tag-wise dispersion parameter for each gene separately is impractical. We could consequently compute the p-value corresponding to each Z statistics by $p_g = 2 \times \Phi(Z \geq |Z_g|)$ based on the pilot data.

2.2 MIXTURE MODEL FITTING

According to the literature review in chapter 1, we saw that a lot of existing power calculation methods for RNA-seq did not consider the heterogeneity of genes with respect to gene expression variability and effect size. Here, we want to directly model p-value distribution by parametric or semi-parametric model to preserve the heterogeneity. This is argued to provide a reasonable model for distribution of p-value in microarray experiment [Allison et al., 2002], and we assume the same for RNA-seq experiment.

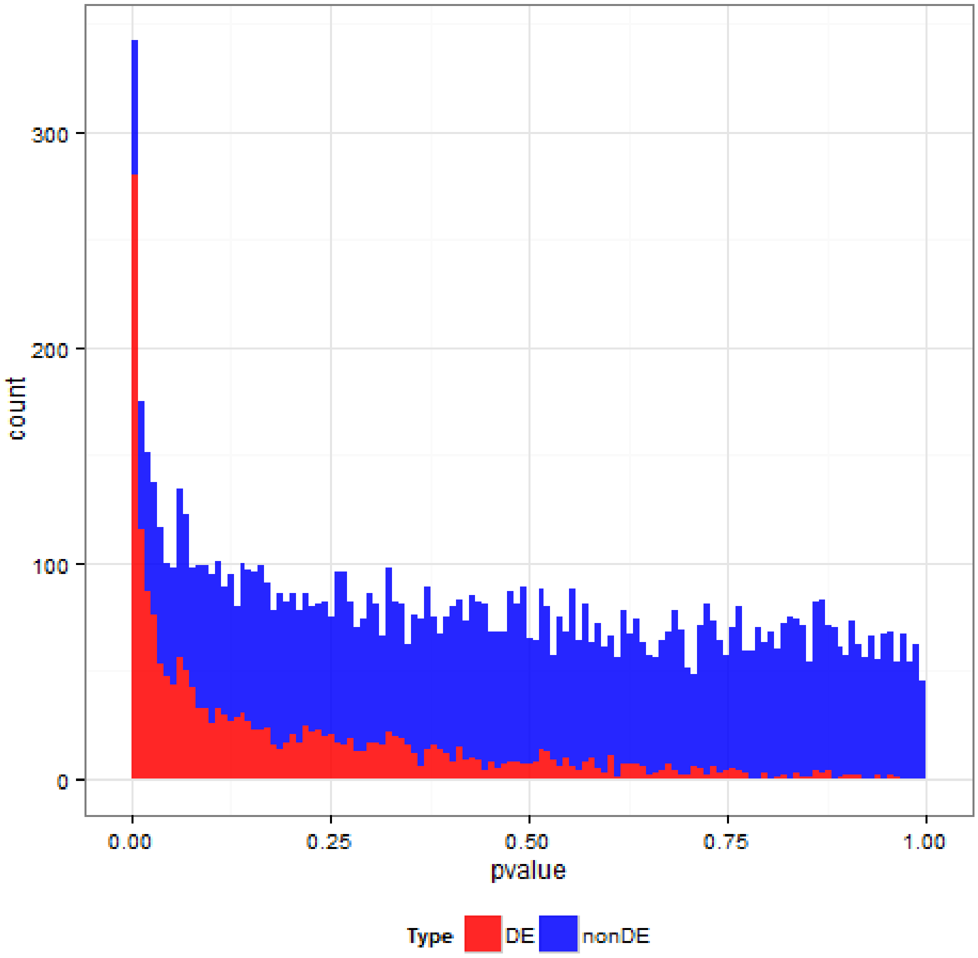
The idea of modeling p-value distribution from microarray study by finite parametric mixture model was proposed by David and others[Allison et al., 2002]. Assuming independence of gene expression levels across genes, under null hypothesis that there's no difference between gene expression of two groups, the distribution of p-value is Uniform(0,1) regardless of the test being used. Under the alternative hypothesis, on the other hand, p-value distribution will tend to concentrate close to 0. Consequently, a useful way in presenting the p-value distribution is by a mixture model. We call the following model as "Beta-Uniform mixture model"(BUM model).

$$f(p|r, s, \lambda) = \lambda f_0(p) + (1 - \lambda) f_1(p|r, s), \tag{2.9}$$

where $f_0(p)$ is a uniform density; $f_1(p|r, s)$ is a beta distribution density with parameter r and s $(0 < r < 1 \leq s)$, $\lambda \in (0, 1)$ is the proportion of non-DE genes. (The constraints are required to guarantee proper shape of beta distribution for DE genes). Figure 2 showed a stacked histogram of p-values with red being the DE gene p-values and blue being non-DE gene p-values from simulated data.

Here, we discussed five alternative approaches in estimating mixture model of p-values distribution. Among them, the first four methods were based on BUM model with parameter λ, r and s. While the fifth method was based on a semi-parametric mixture model, where density of DE component(f_1) was estimated by adaptive kernel smoothing method.

(1)Three parameter BUM model: We used maximum likelihood method to estimate λ, r and s with the above constraints by "L-BFGS-B" method in R function "optim", which allows for box constraints (each variable can be given a lower and/or upper bound). We selected 10 initial points for optimization(9 sets of random initial and 1 set being the estimate of Storey and Tibshirani's method for λ and method of moment estimates for r and s), and choose the solutions corresponds to the largest likelihood.



DE gene p-value follow roughly a beta distribution and non-DE gene p-value follow a uniform distribution.

Figure 2: Stacked histogram: p-value distribution by mixture model.

(2)Storey and Tibshirani's method BUM model: Storey and Tibshirani [Storey and Tibshirani, 2003] proposed an approach to estimate λ though their final goal is to compute q-values given a set of p-values by fitting function of λ to a natural cubic spline.

(3)Two parameter BUM model: This model was introduced in Pounds's work. [Pounds and Morris, 2003] The beta-uniform mixture density was constrained by s=1. So the two parameter BUM model is :

$$f(p|\lambda, r) = \lambda + (1-\lambda)rp^{(r-1)} \tag{2.10}$$

This method was implemented by "Bum" in R package "ClassComparison".

(4)CDD BUM model: Ferkingstad and others [Langaas et al., 2005] proposed another method to estimate λ by nonparametric maximum likelihood. We implemented their method by function "convest" in R package "limma". The shape parameter of beta distribution was then estimated by maximum likelihood method.

(5) Semi-parametric BUM model: In previous work, people had observed the alternative distribution sometimes could not be modeled by beta distribution well. So we now introduce a semi-parametric model for p-value distribution. This is called "semi-parametric" since the non-DE gene p-value distribution is still assume to be uniform, but the DE gene p-value distribution will now be modeled by a non-parametric density. This is usually a decreasing density.

Represent the mixture density as $f(p|\lambda) = \lambda f_0(p) + (1-\lambda)f_1(p)$, our goal is to estimate $\hat{f}_1(p)$. We proposed the following procedure:

1. Estimate λ by Storey and Tibshirani's method;
2. Apply logit transformation: $P = logit(p)$ to avoid boundary effect of density estimation. The mixture model density becomes $g(P|\lambda) = \lambda\frac{e^P}{(1+e^P)^2} + (1-\lambda)g_1(P)$;
3. Estimate $g(P|\lambda)$ by adaptive kernel smoothing method; [Silverman, 1986]
4. Given λ, compute $\hat{g}_1(P)$;
5. Transform $\hat{g}_1(P)$ back to $\hat{f}_1(p)$.

With any of these five methods, we will finally come up with the fitted mixture density:

$$\hat{f}(p|\hat{\lambda}) = \hat{\lambda}\hat{f}_0(p) + (1-\hat{\lambda})\hat{f}_1(p) \tag{2.11}$$

2.3 GENOME-WIDE POWER(EDR) PREDICTION

The ultimate goal of our approach is to predict EDR under targeted sample size(N') with FDR controlled at false discovery rate $\alpha\%$. Given a fitted p-value mixture model, we were able to perform a re-sampling procedure in estimating EDR. The procedure will be repeat for many times for the power prediction. Originally, we proposed both frequentist and bayesian approach. For the convenience of statistics transformation, we determine to use bayesian approach described below. Since previously we had applied both parametric and semi-parametric ways in the estimation procedure, the re-sampling method could also be parametric sampling and non-parametric sampling(e.g., Metropolis Hasting). In both approach, we proposed method to sample DE status based on posterior probability for purpose of power prediction.

2.3.1 Posterior sampling approach based on parametric model

We already computed the Z statistics $(Z_1, ..., Z_g)$ for a given pilot data. Given fitted mixture density (2.11), we can compute posterior probability of whether a certain gene comes from DE gene:

$$P(I_g = 1|\hat{\lambda}, \hat{r}, \hat{s}, p_g) = \frac{(1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s})}{(1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s}) + \hat{\lambda}} \tag{2.12}$$

where I_g is a binary variable indicating the DE gene status. When $I_g = 1$, g^{th} gene comes from DE component. Otherwise, it comes from non-DE component. $\hat{r}$, $\hat{s}$ and $\widehat{lambda}$ were estimated from mixture model fitting. We then implement the following procedure:

1. In the b^{th} simulation, $I^{(b)} = \{I_1^{(b)}, ..., I_2^{(b)}, ..., I_G^{(b)}\}$ are randomly generated from $P(I_g = 1|\hat{\lambda}, \hat{r}, \hat{s}, p_g)$. $(1 \leq g \leq G)$
2. Only p-values from alternative distribution will be transformed by:

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \sqrt{\frac{N'}{N}} + (1 - I_g^{(b)}) \times Z_g \tag{2.13}$$

   where N': predicted sample size. Remember that in formula (2.8), $Var(\hat{\beta}_1)$ could be written as a function of sample size N. Therefore, if we assume effect size of DE gene remains the same as sample size increases, we can apply transformation (2.13). It is also based on the assumption that, as sample size increases, p-value of non-DE gene will remain non-significant, yet DE gene will be more significant.
3. Compute p-value based on 2-sided test: $p_g^{(b)}(I_g^{(b)} = 1) = 2 \times (1 - \Phi(|Z_g^{(b)}|))$
4. Control empirical FDR at $\alpha\%$:
   a. Order p-values so that: $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(j)} \leq ... \leq p_{(G)}$
   b. For $p_{(j)}$, compute $FDR(p_{(j)}) = \frac{\sum_{g=1}^{G}(1-I_g^{(b)}) \times \mathbb{1}_{\{p_g^{(b)} \leq p_{(j)}\}}}{\sum_{g=1}^{G} \mathbb{1}_{\{p_g^{(b)} \leq p_{(j)}\}}}$
   c. $p^{(b)} = \underset{p_g^{(b)}}{\text{argmax}}(FDR(p_g^{(b)}) \leq \alpha)$, where $p^{(b)}$ is the p-value cut-off for b^{th} simulated sample.
5. $\widehat{EDR}^{(b)} = \frac{\hat{R_1}^{(b)}}{\hat{G_1}^{(b)}}$, at the mean time, we can compute: $\widehat{TP}^{(b)} = \frac{\hat{R_1}^{(b)}}{R^{(b)}}$, $\widehat{TN}^{(b)} = \frac{\hat{A_0}^{(b)}}{A^{(b)}}$, where $\hat{R}_1^{(b)} = \sum_{g=1}^{G} I_g^{(b)} \cdot \mathbb{1}_{\{p_g^{(b)} < p^{(b)}\}}$, $\hat{R}_0^{(b)} = \sum_{g=1}^{G}(1 - I_g^{(b)}) \cdot \mathbb{1}_{\{p_g^{(b)} < p^{(b)}\}}$, $\hat{A}_1^{(b)} = \sum_{g=1}^{G} I_g^{(b)} \cdot \mathbb{1}_{\{p_g^{(b)} \geq p^{(b)}\}}$ and $\hat{A}_0^{(b)} = \sum_{g=1}^{G}(1 - I_g^{(b)}) \cdot \mathbb{1}_{\{p_g^{(b)} \geq p^{(b)}\}}$, $\hat{G}_1^{(b)} = \hat{A}_1^{(b)} + \hat{R}_1^{(b)}$, $R^{(b)} = \sum_{g=1}^{G} \mathbb{1}_{\{p_g^{(b)} < p^{(b)}\}}$, $A^{(b)} = \sum_{g=1}^{G} \mathbb{1}_{\{p_g^{(b)} \geq p^{(b)}\}}$ according to Table 2.

We repeated step 1∼5 for B=100 times and computed estimated EDR and evaluate TN rate by averaging across the B repeats:

$$\hat{E}_{\widehat{TP}} = \sum_{b=1}^{B} \widehat{TP}^{(b)}/B, \hat{E}_{\widehat{TN}} = \sum_{b=1}^{B} \widehat{TN}^{(b)}/B, \hat{E}_{\widehat{EDR}} = \sum_{b=1}^{B} \widehat{EDR}^{(b)}/B \tag{2.14}$$

2.3.2 Metropolis Hasting approach based on semi-parametric model

It has been argued that the p-value distribution of DE component sometimes does not strictly follow beta-distribution. In that case, the application of beta distribution will be harmful to the inference of EDR. In the previous section, we proposed ways to model p-value distribution by semi-parametric methods to address this problem. We simulated p-value from alternative distribution by Metropolis Hasting:

2.3.2.1 Bootstrap p-value from estimated density

1. Assuming y $\in \hat{f}_1$, proposal distribution is $N(\bar{y}, 5 \times Var(y))$
2. Initialize: $Y_0 = \bar{y}$
3. In each iteration t, we simulate a number(Y^*) from $N(\bar{y}, 5 \times Var(y))$, and compute acceptance probability:
$$r = \frac{\hat{f}_1(Y^*)}{\hat{f}_1(Y_{t-1})} \tag{2.15}$$
We also generate a random number u from uniform distribution Uniform(0,1). If u $\leq$ r, assign $Y_t = Y^*$, otherwise $Y_t = Y_{t-1}$. And t^{th} number of generated sample is Y_{t-1}.
4. The desirable acceptance rate is around 20 $\sim$ 50%.

In our simulations, we generated markov chain with 20,000 steps and remove the the first 2000 number generated. (burn in period) We implemented both (2.3.2.1) and (2.3.2.2) in our simulation studies. Due to their similar performance, here we'll only show the result of (2.3.2.2).

2.3.2.2 Posterior sampling approach based on semi-parametric model Similar with parametric approach, we generated posterior probability for each p-value (2.12) based on semi-parametric model (substitute $\hat{f}_1(p_g|\hat{r}, \hat{s})$ with $\hat{f}_1(p_g)$), and simulated DE gene status for each bootstrap sample. After the simulation of p-values or DE status, we followed the step (2) $\sim$ (5) as in section 2.3.1.

2.4 SIMULATION STUDIES

To mimic the real world situation, parameters of our simulation settings was based on a real data downloaded from SRA with relatively high read depth. (http://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP023266) We estimated the model parameters from real data and simulated data with some variability of the estimated parameters.

2.4.1 SRA data description and preprocessing

This SRA data was based on a rat study, where the RNA-seq data of noninfectious HIV-1 transgenic (HIV-1Tg) rat was compared with F344 control rats. [Li et al., 2013c] The primary goal of this study was to identify differentially expressed genes and enriched pathways affected by the gag-pol-deleted HIV-1 genome. The authors sequenced RNA transcripts in three brian regions (prefrontal cortex(PFC), hippocampus(HIP), and striatum(STR)) of HIV-1Tg and F344 rats by RNA deep sequencing. A total of 72 RNA samples were sequenced (12 animals per group × 2 strains × 3 brain regions). The differential expression signal in this data set was relatively weak, with fold change cut-off specified $\geq 20\%$. The final number of declared DE genes for the three brain regions were 197, 154, and 171 out of a total of 14,750 genes from the original paper.

Following deep-sequencing of 50-bp paired-end reads of RNA-Seq, we used Bowtie /Tophat /Cufflinks suites(version 2.0.10) to align these reads onto gene regions based on the Rn4 rat reference genome. Then 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].

2.4.2 Differential analysis and Compare Wald test with exact/likelihood ratio test in real data analysis

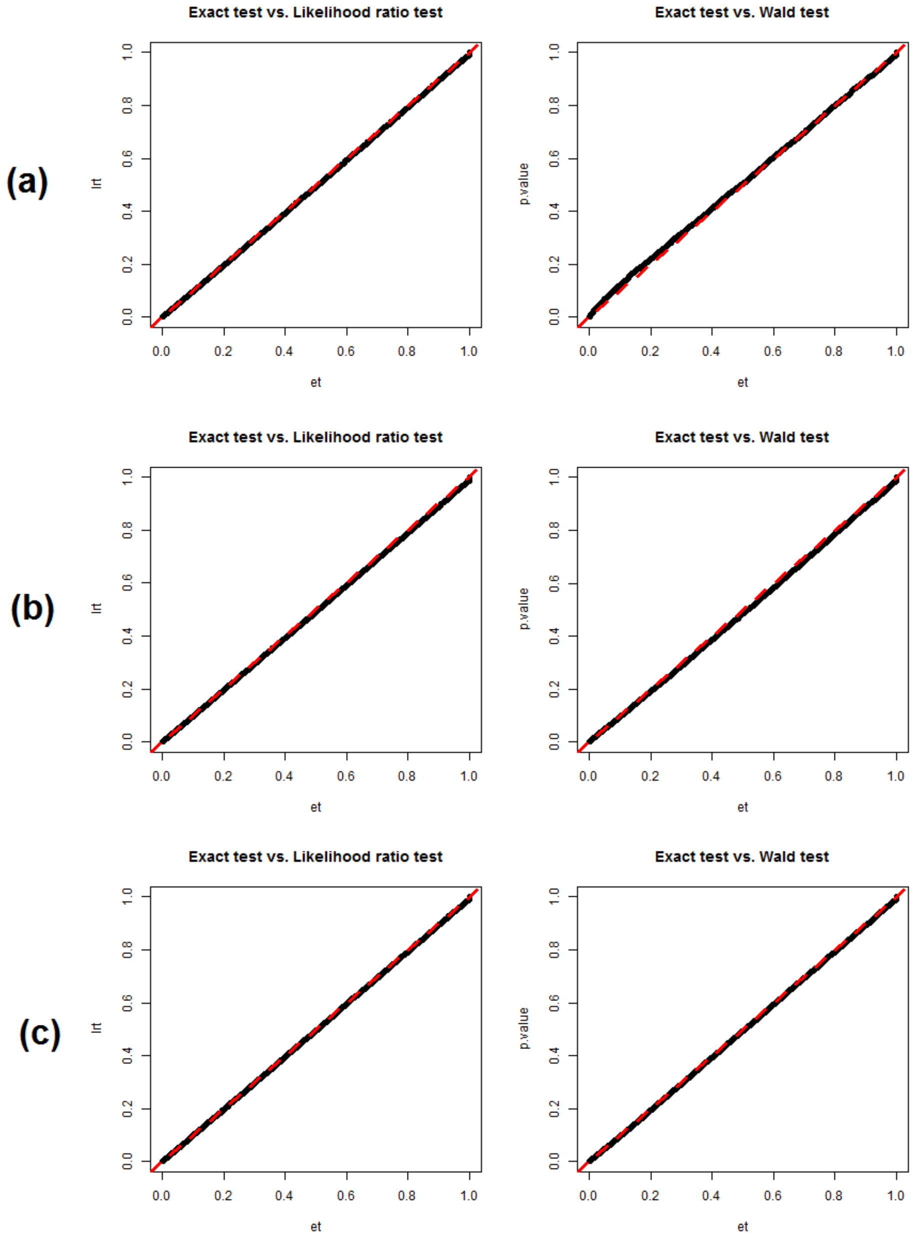
To confirm the validity of Wald test when pilot sample size is small, we first compared the p-value distribution of Wald test and exact/likelihood ratio test performed in "edgeR". (R function "exactTest" and "glmLRT" respectively.) Dispersion parameter δ was estimated by "estimateCommonDisp" in "edgeR". The results are shown in Figure 3.

Figure 3 indicated an almost perfect concordance of p-value distribution between (1) Exact test vs. likelihood ratio test; and (2) Wald test vs. exact test for all three real datasets. We fitted the p-value from Wald test into three parameter BUM mixture model for all three real datasets. Figure 4 showed the fitted p-value density. (red: mixture density; black: uniform component) p-value distributions of PFC and STR data were not ideal to be fitted into mixture parametric model, while HIP data worked well under this assumption.

2.4.3 Simulation settings

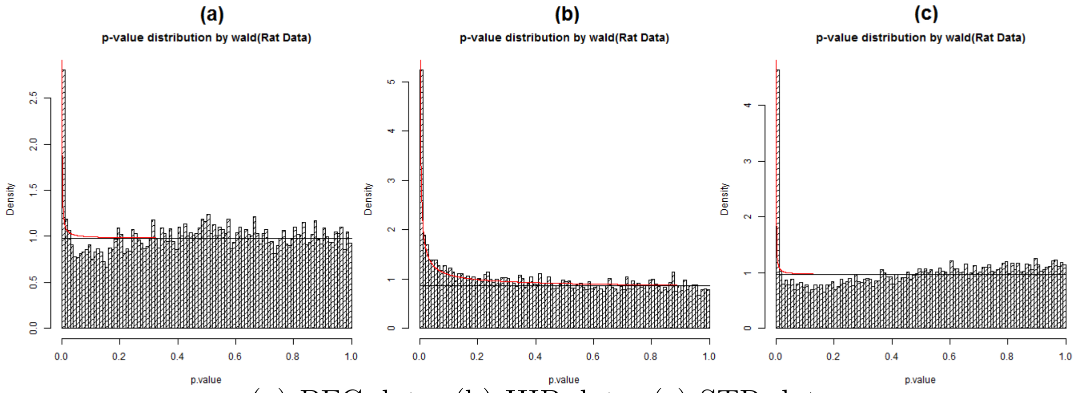
To mimic data structure in real case, we simulated data based on model estimated from real data. We started with HIP data (N=12, R=μG, G =14750, μ ≈600) since that its p-value distribution had good fitting into p-value mixture model and therefore we computed the mean counts per gene(μ_g) and generated data sets based on the its empirical distribution. We also calculated log-fold-change(lfc_g), which is simply, $\hat{e}_{g1}/\hat{e}_{g0}$, and lfc_g was fitted into a truncated normal distribution with four different cut-offs. Common dispersion parameter δ was estimated by "edgeR". See Table 4 for the notations for parameters in simulation settings.

To evaluate predicted EDR under different simulation settings, we generated data with slightly different dispersion and fold change parameters. Based on real data, we estimated dispersion to be $\hat{\delta} = 50$. We selected $\delta = 40, 45, 50, 55, 60$ in our simulation. In real data analysis, we estimated mixture model parameters of p-values. $\hat{\lambda}$ by Storey's method was roughly 0.90, meaning 10% genes come from DE component. The distribution of lfc from



(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).

Figure 3: Compare distribution of p-values by using Wald test and exact/likelihood ratio test.



(a) PFC data; (b) HIP data; (c) STR data.

Figure 4: p-value distribution by Mixture model

Table 4: Notations of parameters in simulation settings

Parameter	Explanation
N	Number of biological replicates in each group in pilot study
N'	Number of biological replicates in each group in future study
G	Total number of genes in pilot study
R	Total number of reads for each sample in pilot study (R=$\mu\times$G)
R'	Total number of reads for each sample in future study
e_{gij}	Observed reads for subject j in group i
μ	Average reads per genes per sample
μ_g	Average reads for gene g per sample
μ_{g1}	Average reads for gene g in group A
μ_{g0}	Average reads for gene g in group B
δ	Common dispersion parameter
m%	Percentage of DE gene
lfc_g	log-fold-change for gene g

DE genes followed N(0,0.04) truncated at certain cut-offs. If we selected the top 10% genes with largest lfc then the corresponding cut-off was around 0.2 (fc=1.15). We also added in several alternative cut-offs, including 0.26, 0.32 and 0.38. (corresponds to 1.2; 1.25 and 1.30 at fold change scale) In the original paper of HIP data, the reported significance of DE gene was defined as p≤0.005 (FDR≤0.2) with a fold change (FC) $\geq$ 20%, which coincided with our selection. The average number of reads for each gene was set to be 650. ($\mu = 650$) This corresponds to the data generated by 1 lane and with alignment rate 50%. The total number of genes(G) is set to be 10,000. The total number of reads per sample is R=6.5M.

The simulation studies is summarized in Figure 5. The details of simulation steps to generate data with (N, R) is as follows:

1. Randomly sample μ_g from empirical distribution estimated from real data.

2. Generate lfc from specified truncated normal distribution

3. Assign DE gene label: generate random number r_g from Uniform(0,1), if r_g ≤0.10 then gth gene is DE, otherwise, it is non-DE gene.

4. Generate expression value for each sample: if gth gene is DE, $e_{g1j} \sim NB(\mu_g \times 2^{lfc_g/2}, \delta)$, $e_{g0j} \sim NB(\mu_g \times 2^{-lfc_g/2}, \delta)$; otherwise $e_{gi.} \sim NB(\mu_g, \delta)$

We simulated 50 datasets under each simulation setting with pilot data sample size(N) $N = 2, 4, 8, 16$. For each pilot data with sample size N in the left flow chart of Figure 5, EDR was predicted for targeted sample size(N') $N' = 5, 10, 20, 30, 40, 50, 100$, which could be denoted by $TP_N(N')$, $TN_N(N')$, $EDR_N(N')$. Posterior sampling procedure was repeated for 20 times. On the right side of Figure 5, $EDR(N')$ was computed from simulated data under the target sample size (N'). The performance of power calculation was evaluated by comparing $EDR_N(N')$ with EDR(N').

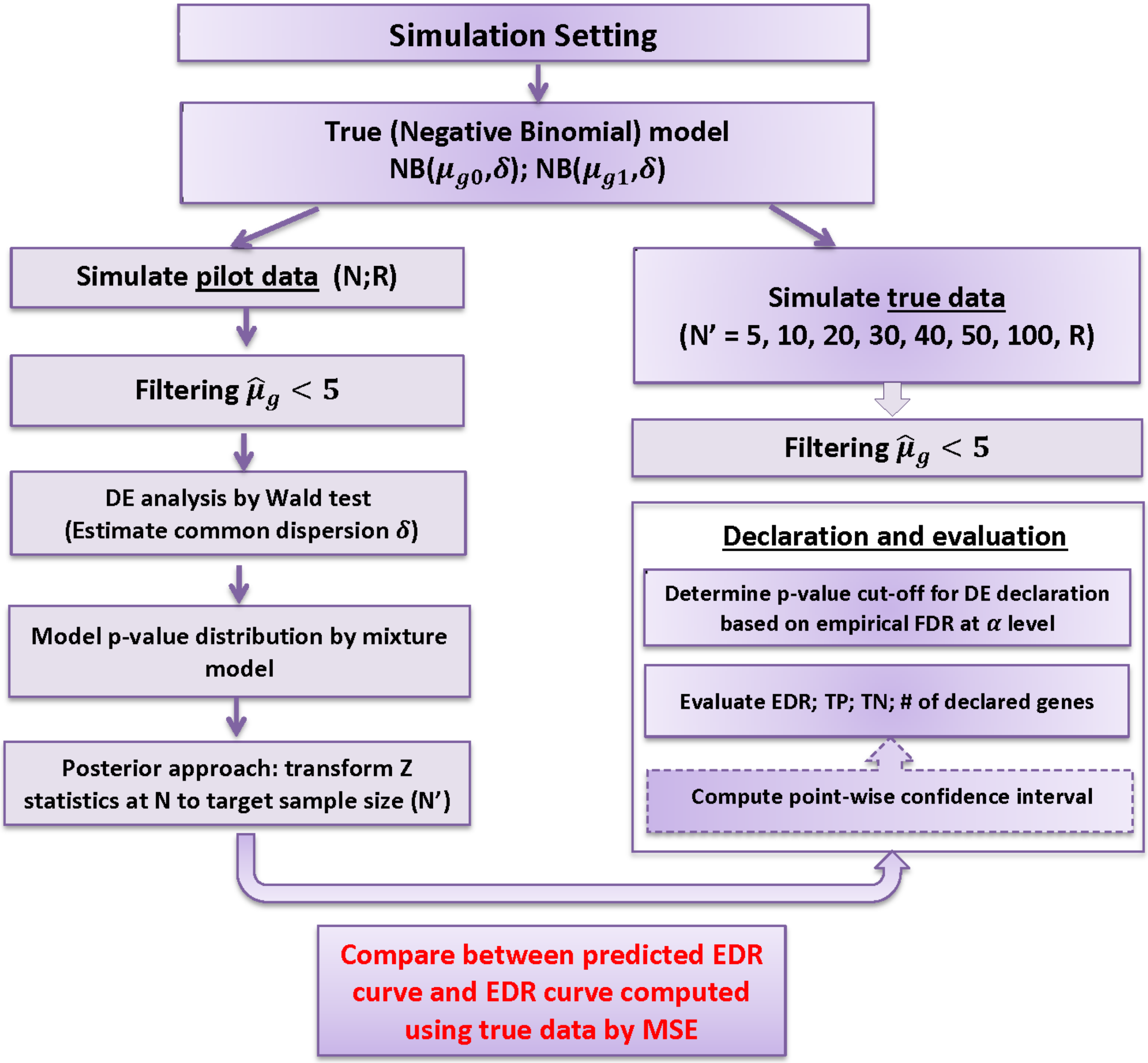

Simulation Setting
True (Negative Binomial) model
NB(μ_{g0},δ); NB(μ_{g1},δ)
Simulate pilot data (N;R)
Filtering $\widehat{\mu}_g < 5$
DE analysis by Wald test
(Estimate common dispersion δ)
Model p-value distribution by mixture model
Posterior approach: transform Z statistics at N to target sample size (N')
Simulate true data
(N' = 5, 10, 20, 30, 40, 50, 100, R)
Filtering $\widehat{\mu}_g < 5$
Declaration and evaluation
Determine p-value cut-off for DE declaration based on empirical FDR at α level
Evaluate EDR; TP; TN; # of declared genes
Compute point-wise confidence interval
Compare between predicted EDR curve and EDR curve computed using true data by MSE

Figure 5: Flowchart of SeqDEsign when predicting EDR at targeted sample size N'

2.4.4 Compare ROC curve of Wald test and exact test

Wald test applies to approximations (plugged-in standard deviation and chi-squared approximation) and sometimes raised concerns of accuracy compared to exact test. To demonstrate the validity of Wald test, we first compared it with exact test under negative binomial distribution using “exactTest” function in “edgeR” package for the two simulation settings described above. Since we know true labels of DE gene under simulation settings, we can compare the Receiver operating characteristic (ROC) curve and area under curve (AUC) of Wald test and exact test.

We compared the two tests in 12 of the total 20 simulation setting. (Dispersion is choose to be 40, 50, 60; fc is choose to be ≥ 1.15, ≥ 1.20, ≥ 1.25, ≥ 1.30). In each setting, we generated 50 datasets and performed the two tests. Pilot data sample size N was fixed as N=4. For both methods, common dispersion parameter was estimated by “estimateCommonDisp” due to the small sample size of pilot data. Then, we performed the two tests separately for simulated data under each setting and generated ROC curve by comparing the declared genes with the true labels. Figure 6 show the ROC curve(with boxplot of 50 datasets). The ROC curves of exact and Wald test almost overlapped with each other, indicating good concordance.

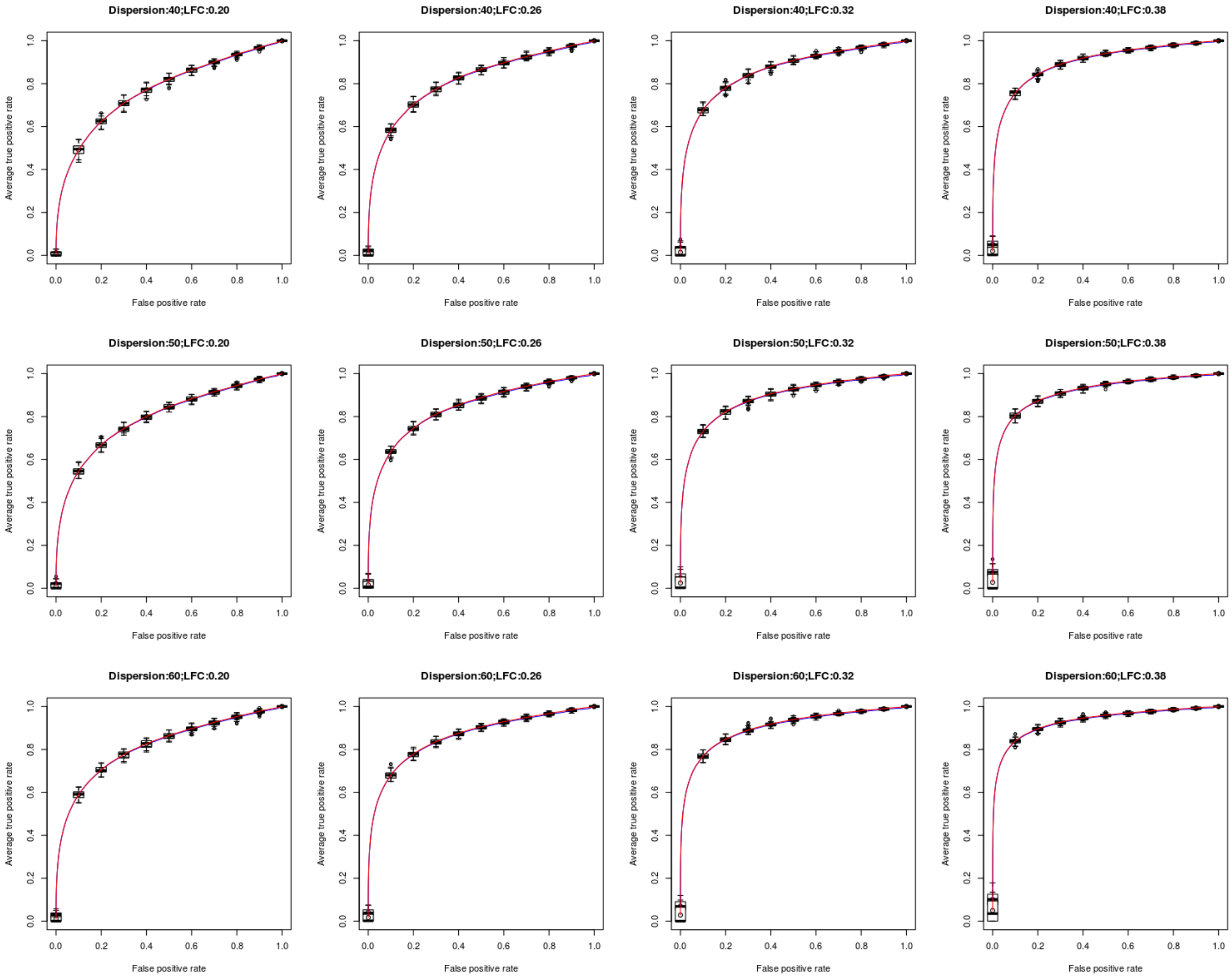


Figure 6: ROC curve comparing exact and wald test under 12 simulation settings.

Table 5: 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	0.7693(0.0099)	0.7711(0.0098)	0.7944(0.0088)	0.7967(0.0088)	0.8160(0.0098)	0.8186(0.0098)
≥ 1.20	0.8145(0.0086)	0.8172(0.0086)	0.8401(0.0075)	0.8430(0.0075)	0.8594(0.0071)	0.8625(0.0071)
≥ 1.25	0.8606(0.0074)	0.8631(0.0073)	0.8852(0.0074)	0.8881(0.0072)	0.9002(0.0053)	0.9032(0.0052)
≥ 1.30	0.8983(0.0057)	0.9008(0.0057)	0.9155(0.0061)	0.9182(0.0060)	0.9293(0.0050)	0.9322(0.0048)

Table 5 showed the mean and standard deviation of AUC for both cases. In all cases, Wald test had slightly higher mean AUC compared with exact test in edgeR. But there was no statistically significant difference between the performance of Wald test and exact test.

2.4.5 Performance of BUM model estimation methods under different simulation settings

Previously, we introduced several alternative approaches in modeling the mixture model of p-value distribution. In any of these methods, we performed sampling scheme for DE gene labels based on empirical Bayes method. In each simulation setting, we computed the True EDR curve with its point-wise confidence interval by normal approximation(bounded between 0 and 1). We compared it with predicted EDR curve and corresponding confidence interval. We performed simulation studies to a total of 20 settings. (4 log-fold-change settings(lfc∼N(0,0.04), truncated at +/-0.20; lfc∼N(0,0.04), truncated at +/-0.26; lfc∼N(0,0.04), truncated at +/-0.32, lfc∼N(0,0.04), truncated at +/-0.38) and 5 dispersion parameter settings(δ=40, 45, 50, 55, 60).)

To evaluate accuracy of power prediction with different pilot data sample size(N) in simulations, we generated mean squared error of the 7 data points(N'=5, 10, 20, 30, 40, 50, 100) by comparing between predicted EDR and true EDR to evaluate the five approaches.

Figure 7 ∼ Figure 11 showed the predicted EDR when N=2 for all five different approaches. We observed that Storey and Tibshirani's method BUM model and CDD BUM model outperformed the other methods in most simulation settings. We further summarized the performance of four approaches by mean square error(MSE) in Figure 15 and Table 6 ∼ Table 8, where Storey and Tibshirani's method BUM model and CDD BUM model had the best performance in most of the simulation settings. The MSE is computed by:

$$MSE(N) = \sum_{N'} (\overline{\widehat{EDR}_N}(N') - \overline{EDR}(N'))^2 \tag{2.16}$$

where $\overline{\widehat{EDR}_N}(N')$ is the average of predicted EDR under (N',R) given (N,R), and $\overline{EDR}(N')$ is the average of true EDR.

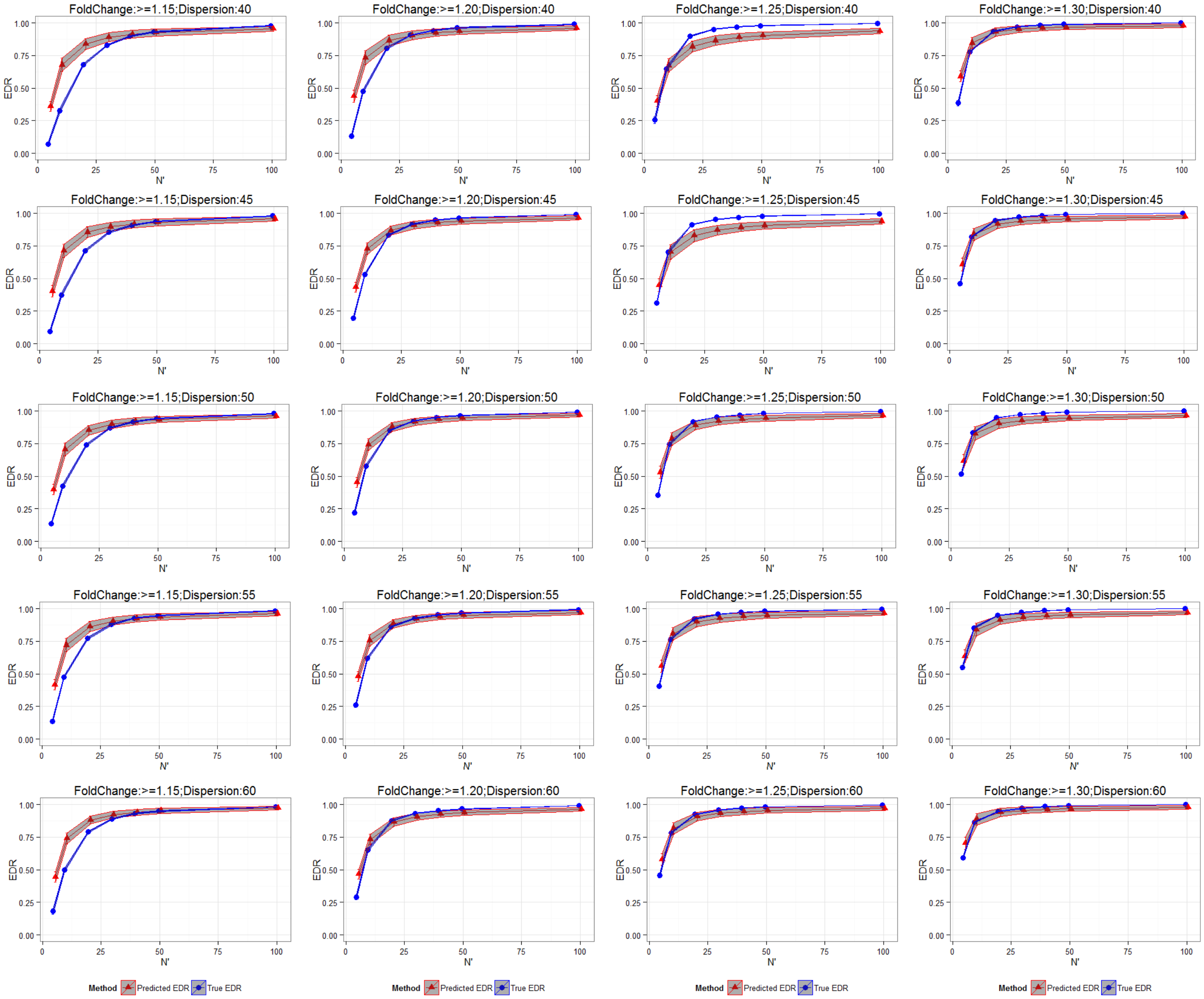


Figure 7: Predicted and True EDR curve for 3-parameter BUM model (N=2)

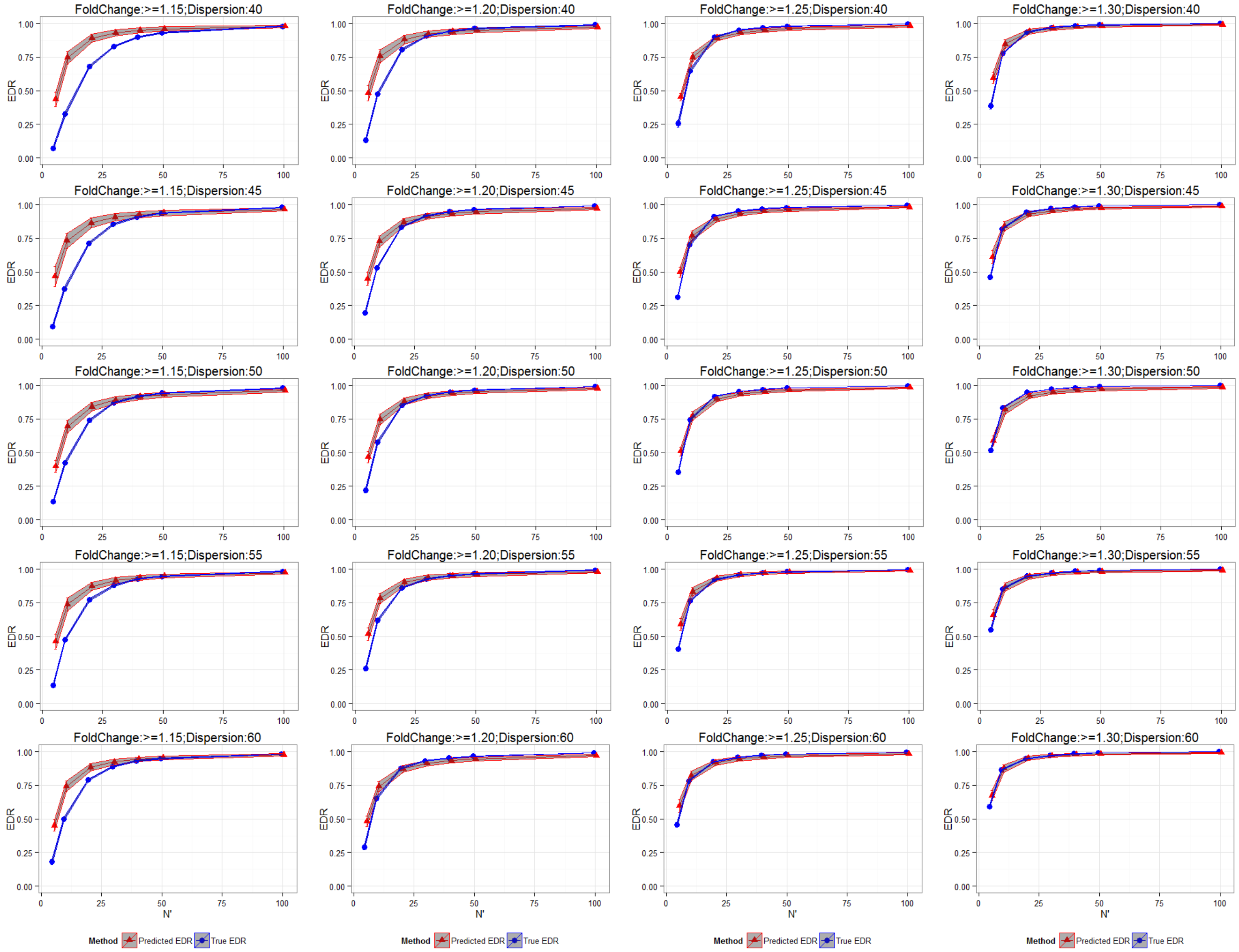


Figure 8: Predicted and True EDR curve for Storey's method BUM model (N=2)

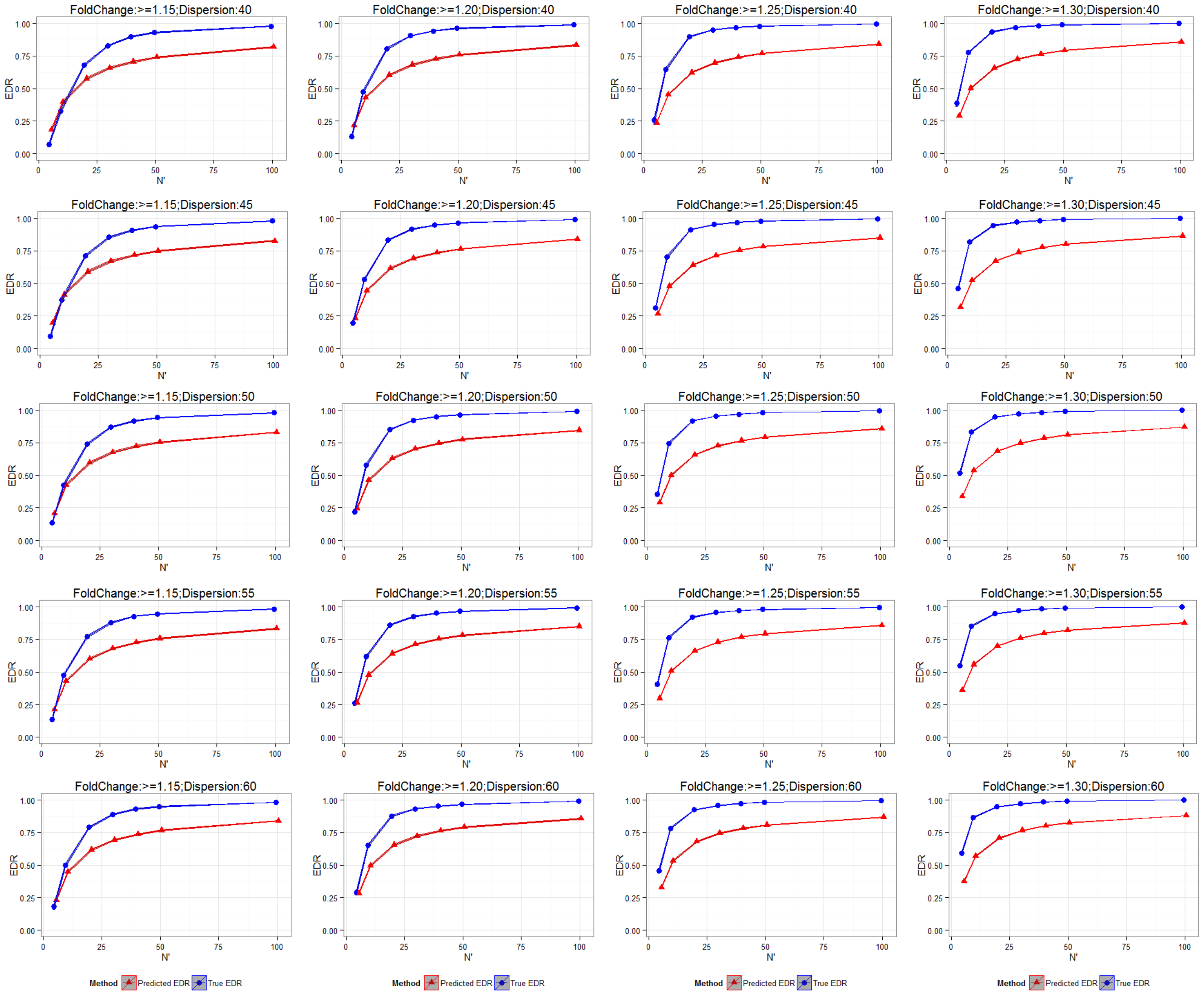


Figure 9: Predicted and True EDR curve for 2-parameter BUM model (N=2)

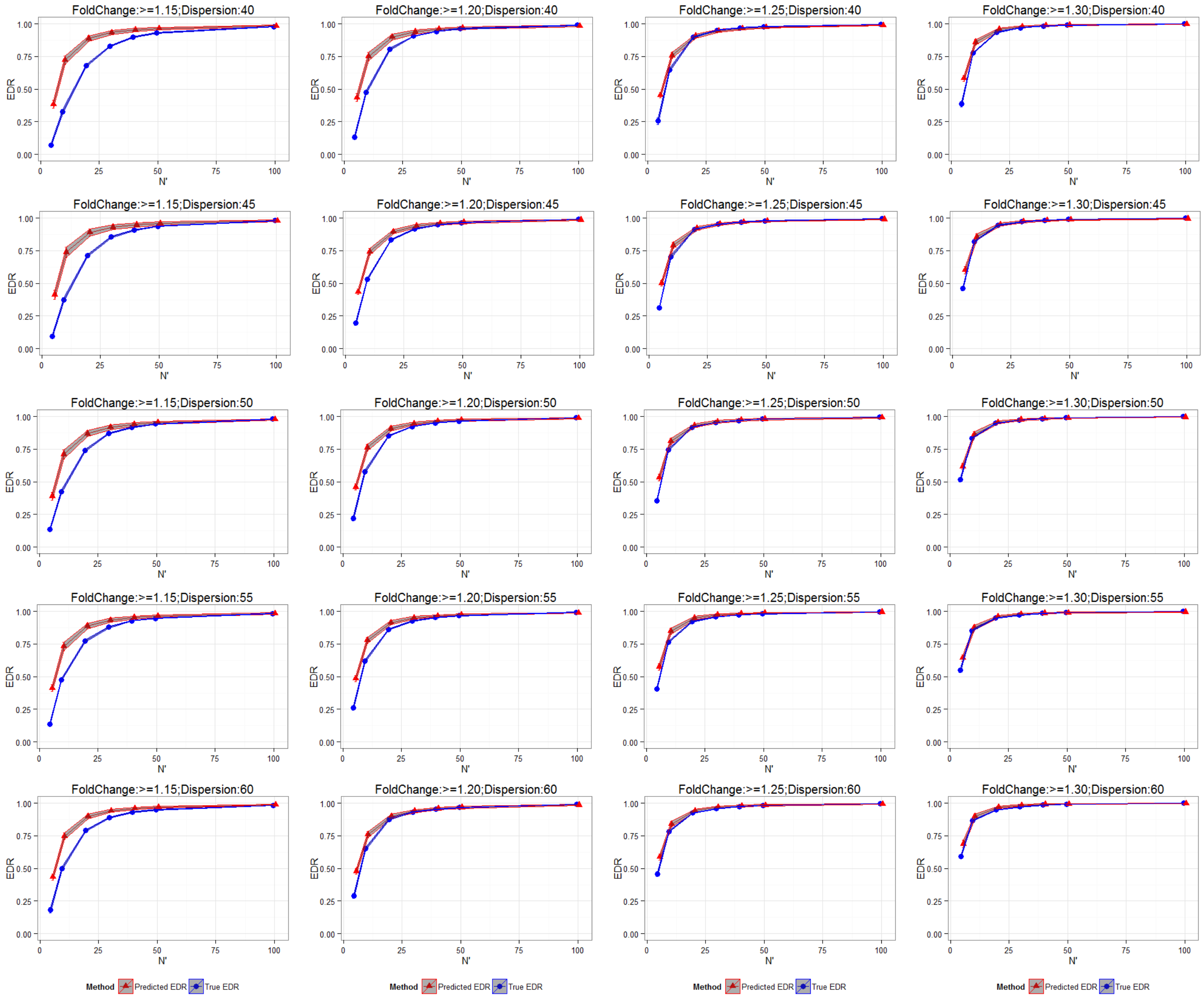


Figure 10: Predicted and True EDR curve for CDD BUM model (N=2)

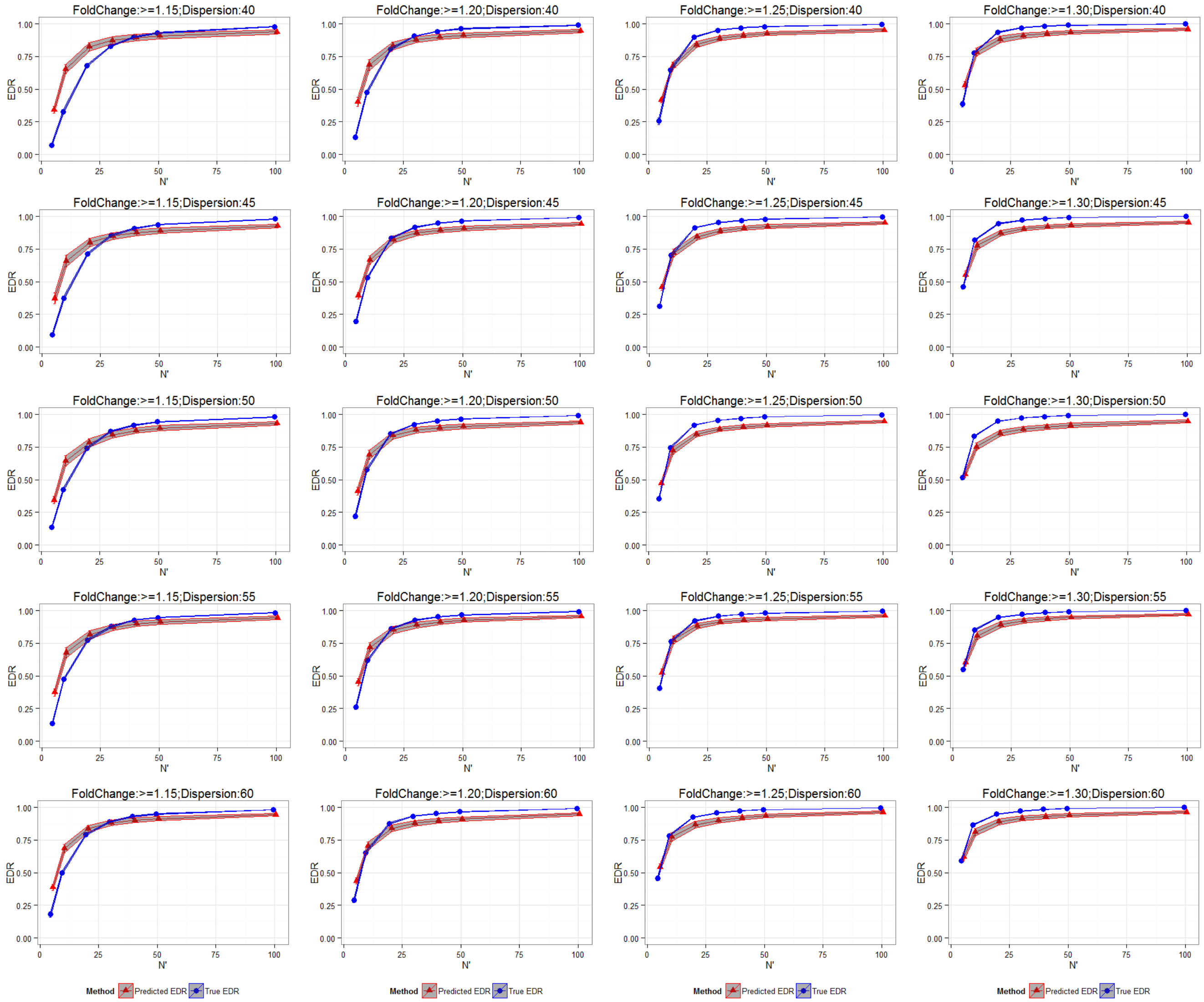


Figure 11: Predicted and True EDR curve for semi-parametric BUM model (N=2)

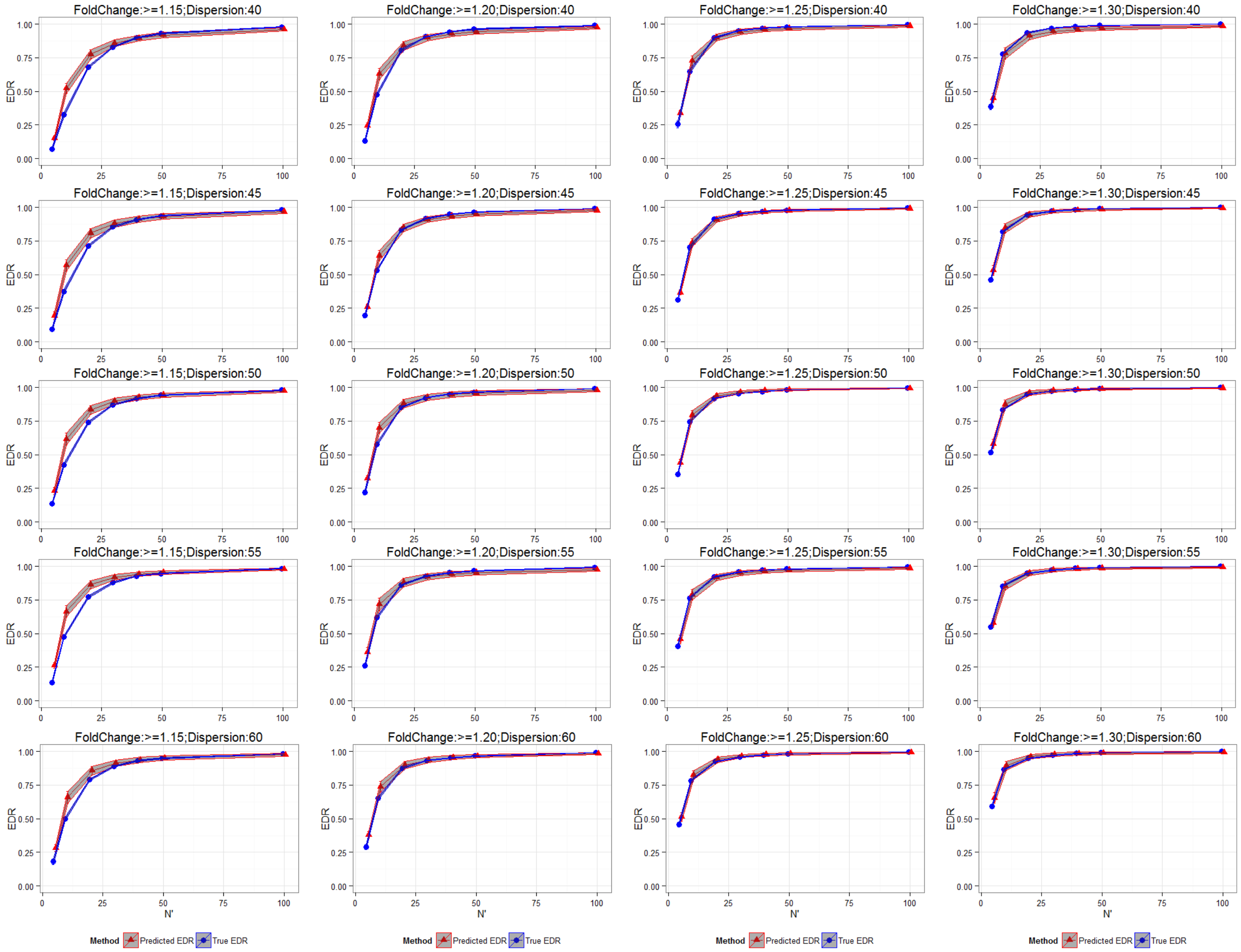


Figure 12: Predicted and True EDR curve for Storey and Tibshirani's method BUM model (N=4)

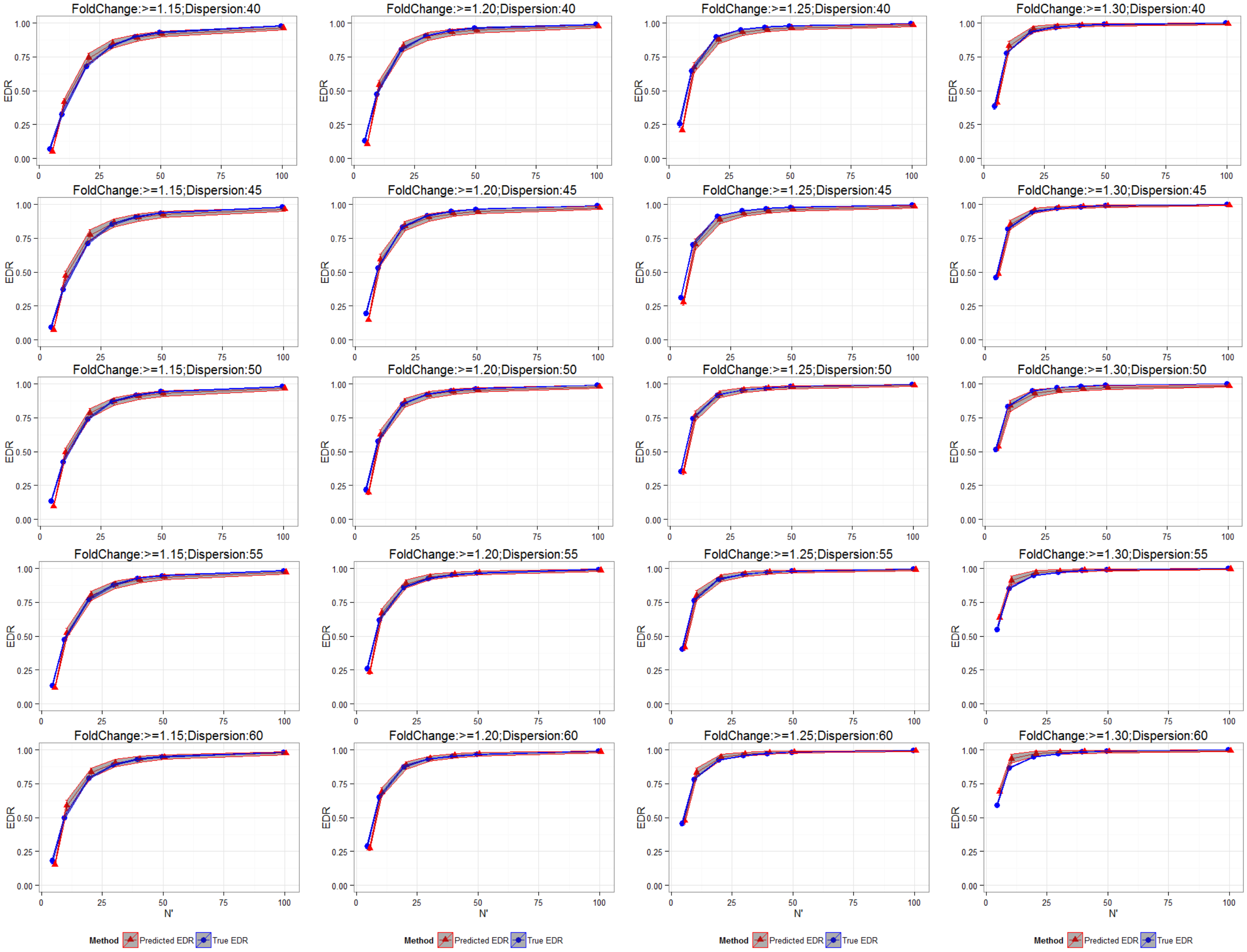


Figure 13: Predicted and True EDR curve for Storey and Tibshirani's method BUM model (N=8)

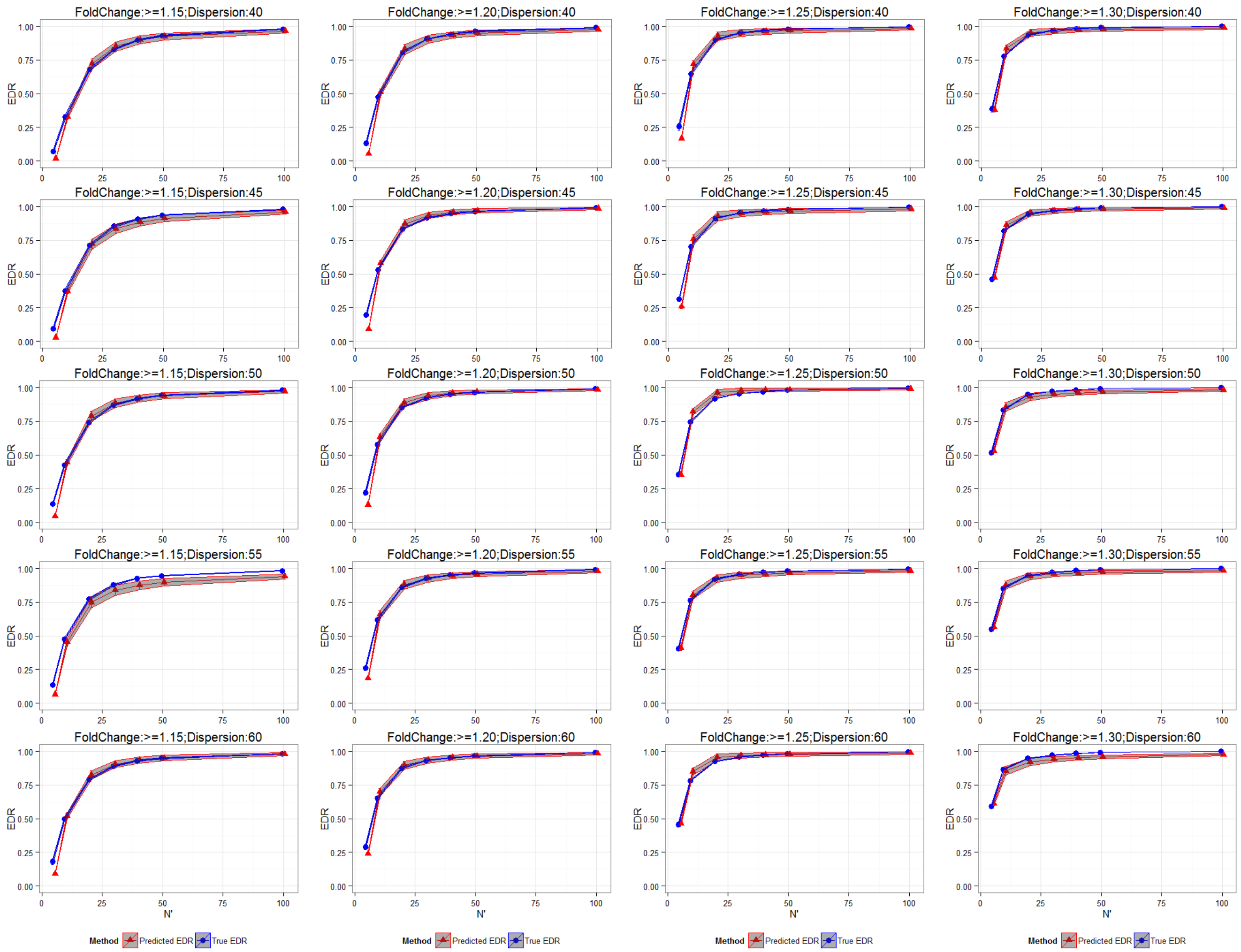
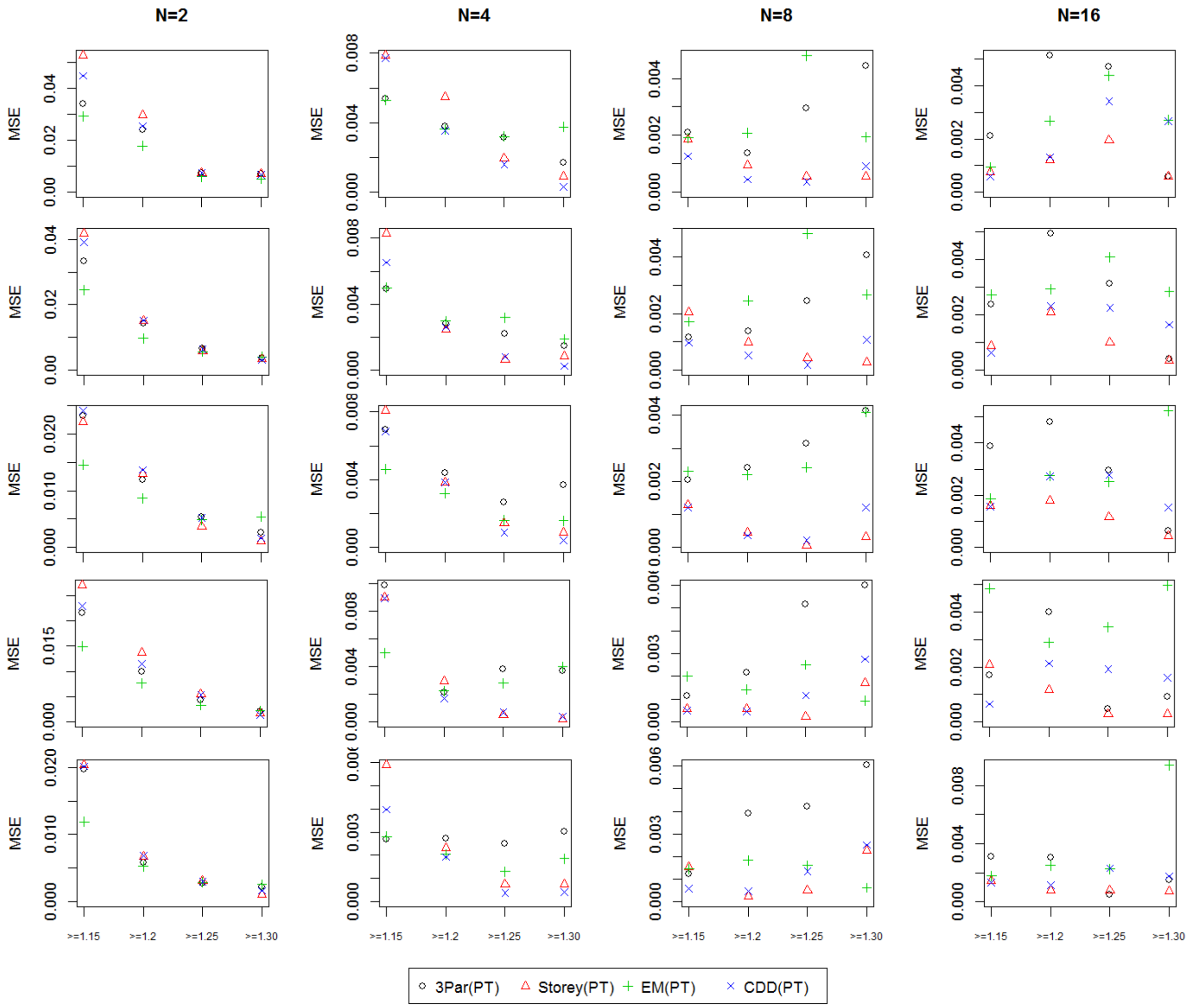


Figure 14: Predicted and True EDR curve for Storey and Tibshirani's method BUM model (N=16)



Five rows indicate five different dispersion parameter setting: δ=40, 45, 50, 55, 60.

Figure 15: Compare four different power prediction approaches by mean square error (MSE)

Table 6: Summary table for MSE for simulation settings $\delta = 40$ and all lfc settings

Method	N=2	N=4	N=8	N=16
3-parameter BUM model	0.0180	0.0035	0.0027	0.0031
Storey's method BUM model	0.0241	0.0040	0.0009	0.0011
Semi-parametric model	0.0145	0.0040	0.0023	0.0027
2-parameter BUM model	0.0354	0.0426	0.0425	0.0304
CDD BUM model	0.0210	0.0033	0.0007	0.0020

Due to the good performance and computation simplicity, we sticked to Storey and Tibshirani's method BUM model as our approach for mixture model fitting for later methods. Figure 12, Figure 13 and Figure 14 showed the predicted EDR curve for 3 parameter BUM model when N increases to 4, 8, 16. The accuracy of EDR prediction was improved as pilot sample size increased.

2.5 MODEL DIAGNOSTICS

Since our EDR prediction methods depend on the fitting of parametric mixture model, it is important to evaluate the fitting of mixture model in two aspects: (1) Fitting of beta component; (2) Estimation of λ by different methods. We focused on Storey and Tibshirani's method BUM model and CDD BUM model due to their good performance.

2.5.1 Fitting of beta component

We compared the distribution of fitted beta component with the empirical distribution of true DE genes. Specifically, we extracted the true DE genes according to underlying true

Table 7: Summary table for MSE for simulation settings $\delta = 50$ and all lfc settings

Method	N=2	N=4	N=8	N=16
3-parameter BUM model	0.0110	0.0044	0.0029	0.0031
Storey's method BUM model	0.0099	0.0035	0.0005	0.0012
Semi-parametric model	0.0084	0.0027	0.0027	0.0031
2-parameter BUM model	0.0347	0.0410	0.0367	0.0240
CDD BUM model	0.0111	0.0030	0.0024	0.0021

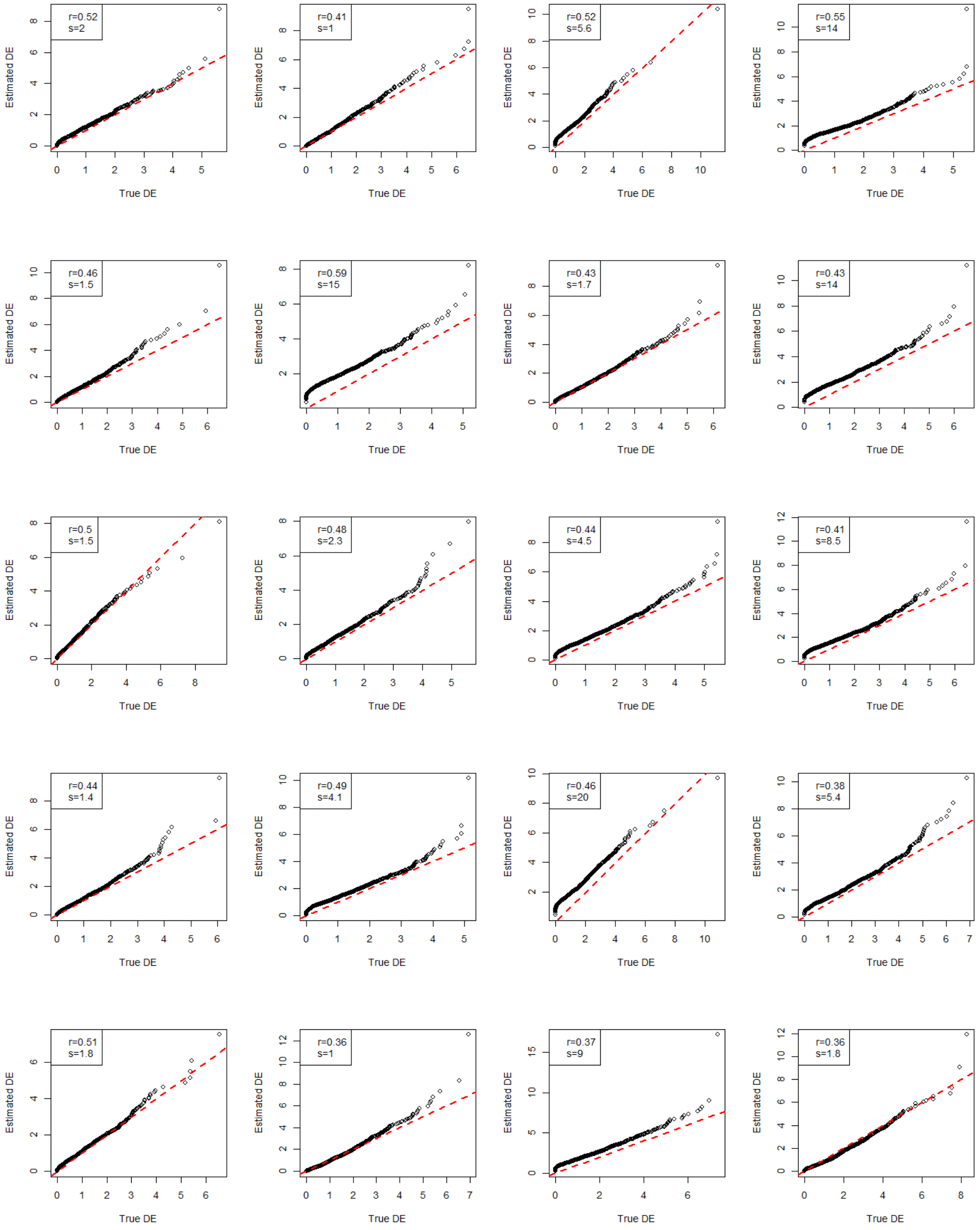
label and compared with the distribution of fitted beta distribution $f_1(p|\hat{r}, \hat{s})$ by Quantile-Quantile(QQ) plot of -log10 p-values. In Figure 16, the DE gene p-value distribution was fitted well by estimated beta distribution using Storey and Tibshirani's method BUM model according to Quantile-Quantile plot of -log10 p-value. For CDD BUM model, the fitting of beta component was good as well. (not shown here) Figure 17 showed the empirical p-value distribution for each simulation setting when N=2. As dispersion and fold change cut-off increased, p-value distribution of DE genes became sharper, while non-DE gene still followed uniform distribution.

2.5.2 Estimation of λ

Here we compared the estimation of λ with our underlying truth (λ ≈0.9) under different simulation settings for N=2.

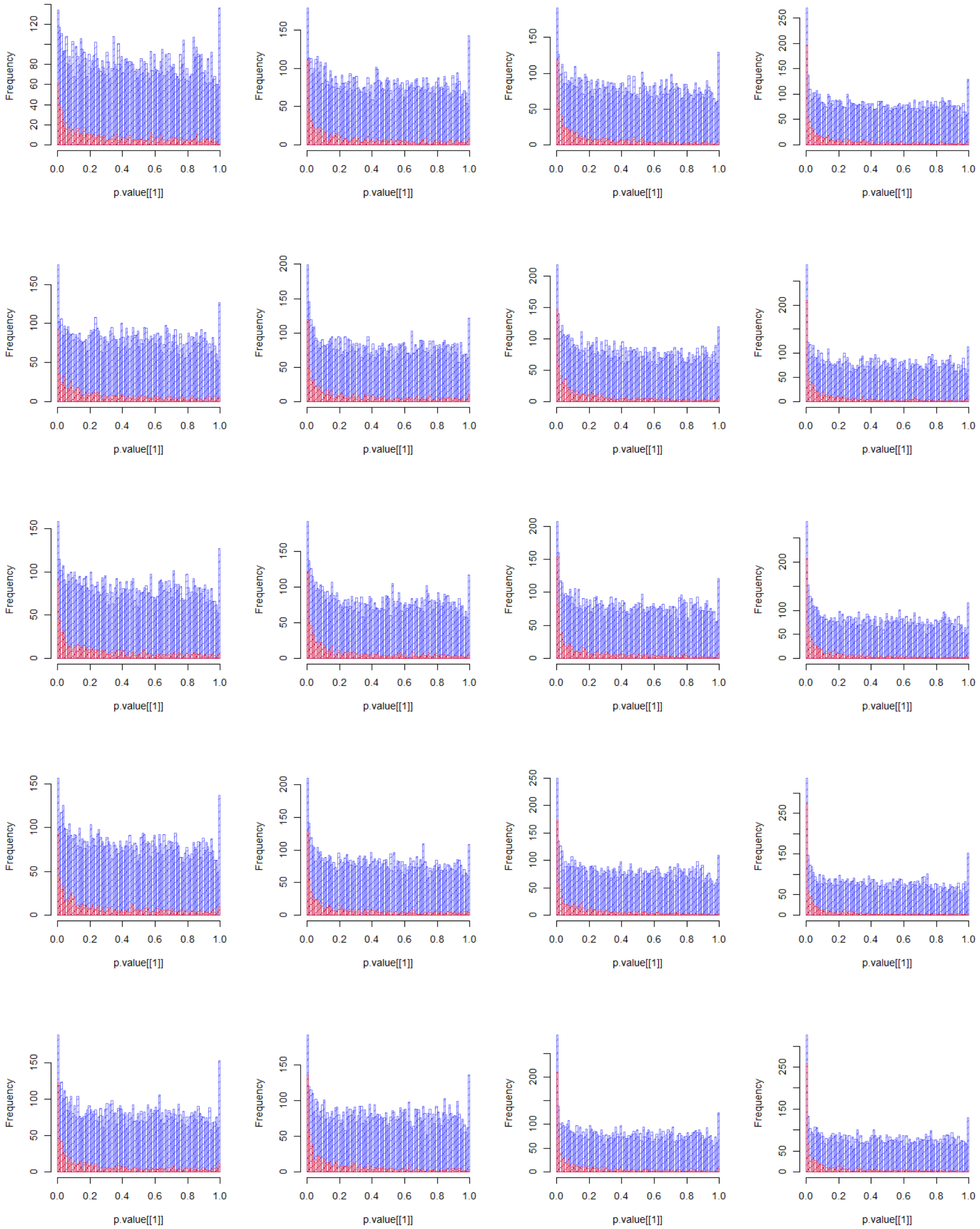
Table 8: Summary table for MSE for simulation settings $\delta = 60$ and all lfc settings

Method	N=2	N=4	N=8	N=16
3-parameter BUM model	0.0076	0.0027	0.0038	0.0022
Storey's method BUM model	0.0078	0.0024	0.0011	0.0009
Semi-parametric model	0.0056	0.0020	0.0014	0.0040
2-parameter BUM model	0.0337	0.0382	0.0310	0.0194
CDD BUM model	0.0079	0.0017	0.0012	0.0016



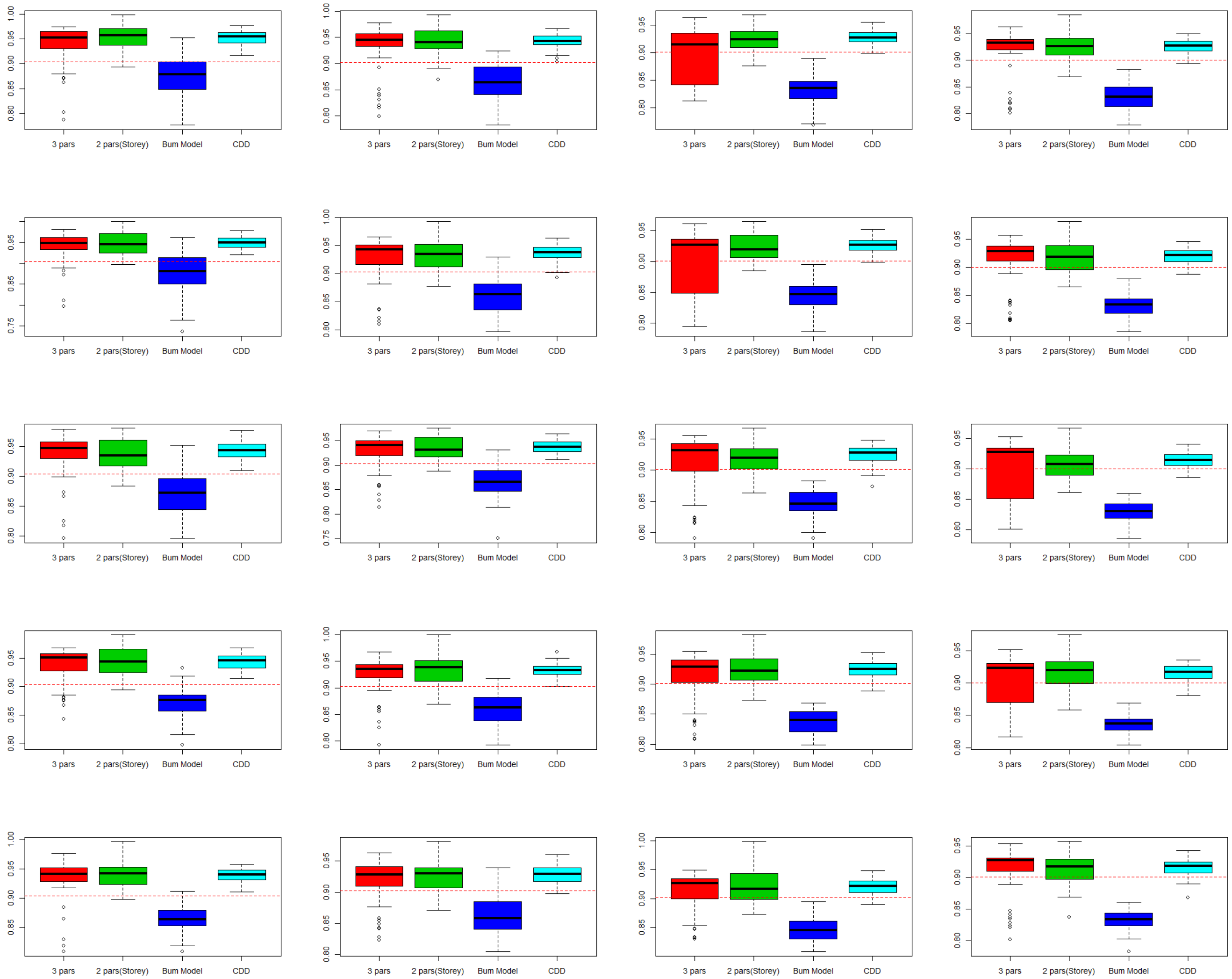
Five rows indicate different dispersion parameter (1) δ =40;(2) δ =45;(3) δ =50;(4) δ =55;(5) δ =60. Four columns indicate different fold change cut-off (1) fc $\geq$ 1.15; (2) fc $\geq$ 1.20; (3) fc $\geq$ 1.25; (4) fc $\geq$ 1.30.

Figure 16: Quantile-Quantile plot for 20 simulation settings



Five rows indicate different dispersion parameter (1) δ =40;(2) δ =45;(3) δ =50;(4) δ =55;(5) δ =60. Four columns indicate different fold change cut-off (1) fc $\geq$ 1.15; (2) fc $\geq$ 1.20; (3) fc $\geq$ 1.25; (4) fc $\geq$ 1.30.

Figure 17: p-value distribution in different simulation settings(N=2)



Five rows indicate different dispersion parameter (1) δ =40;(2) δ =45;(3) δ =50;(4) δ =55;(5) δ =60. Four columns indicate different fold change cut-off (1) fc $\geq$ 1.15; (2) fc $\geq$ 1.20; (3) fc $\geq$ 1.25; (4) fc $\geq$ 1.30.

Figure 18: Lambda estimate of five methods in different simulation settings(N=2)

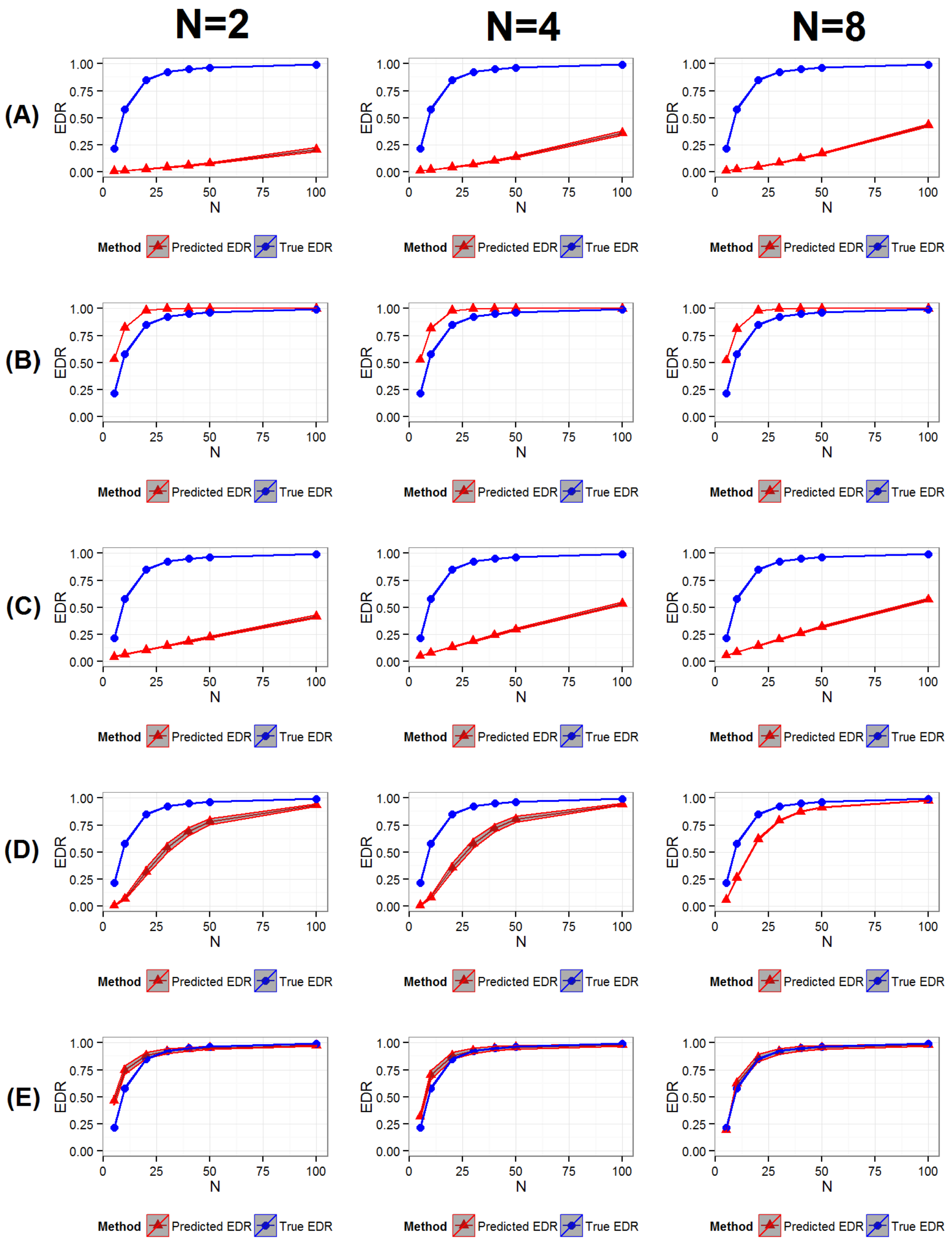
For all four methods, λ tended to be over-estimated(Figure 18). As pilot sample size increased, the estimation was more towards underlying truth(not shown here). To note, CDD method actually gave the most accurate estimates among all methods, especially when the data signal was stronger.

2.6 COMPARISON WITH OTHER EXISTING METHODS IN SIMULATION SETTINGS

In chapter 1, we introduced several existing methods for the power calculation of RNA-Seq data. Here, we compared our proposed method with four other methods: (1) Poisson model; (2) RNASeqPower; (3) NB exact test; and (4) Scotty in the simulation setting when $\delta = 50$ and $lfc \geq 0.26$ (fc ≥ 1.2) with N=2, 4, 8 and 16. Since the methods depends on different assumptions, we need to compare them in a more reasonable way. In general, we estimated input parameters from pilot data or provided the underlying truth if favorable to the existing method to facilitate a fair comparison.

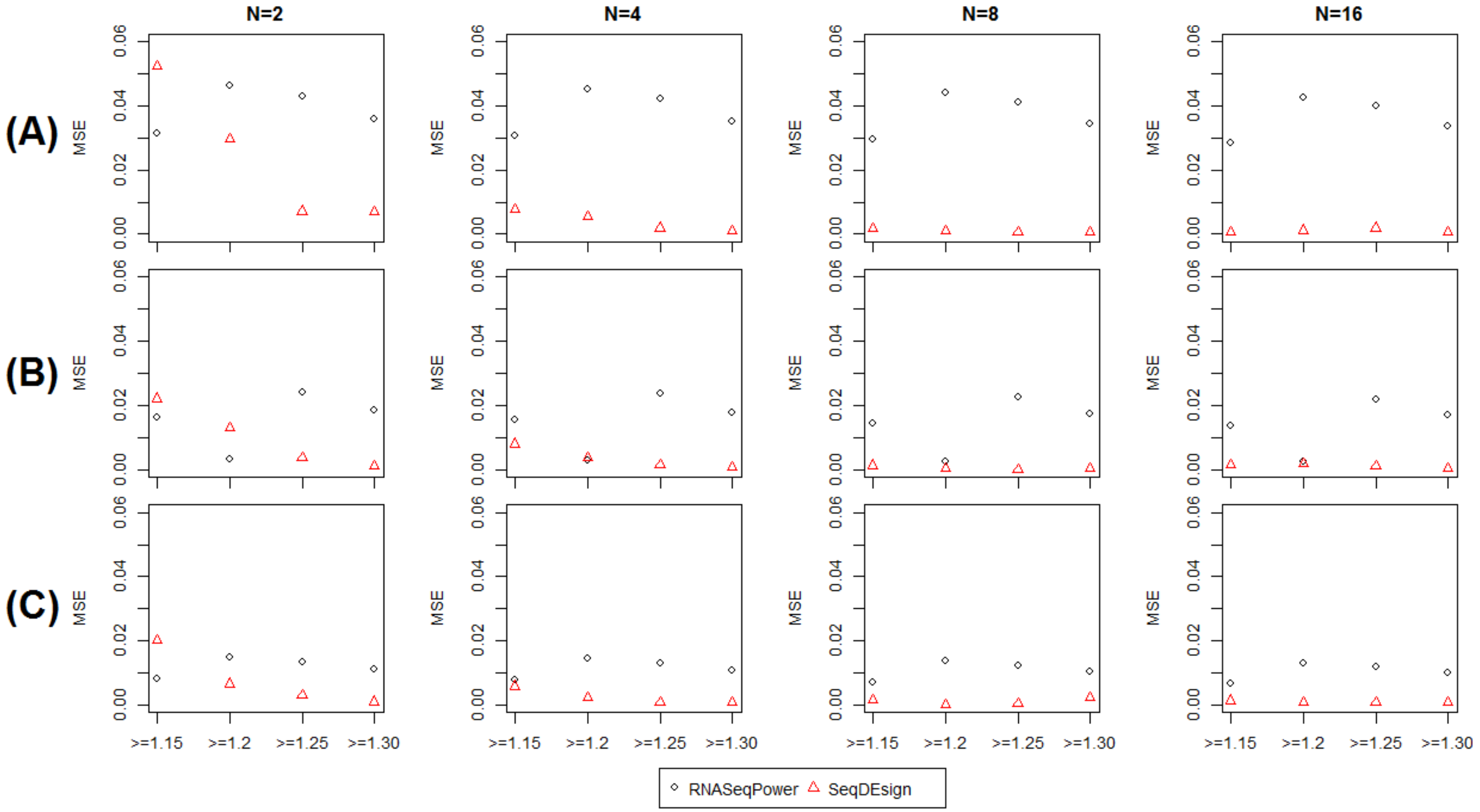
Similar to what we did for SeqDEsign, in all the other tests, we first filtered out genes with small mean counts across samples. Several cut-offs for mean counts(1,2,5,8,10) were tested and compared. The basic rule is that we want to remove a small number of genes with shallow coverage which influence the fitting of mixture model. Through the comparison, we determined to use 5 reads/gene as the cut-off for all simulation studies. We implemented power calculation by Poisson model based on R code provided by Li and others[Li et al., 2013a]. In detail, we gave it the true proportion of prognostic genes, which is around 10% of total genes, and assumed 80% of them are true rejections, which was suggested in the example of the original paper. The FDR was set as 0.05. We gave the true minimum DE gene fold change 1.2 as the input parameter. We then applied RNASeqPower method by R function "rnapower" in package "RNASeqPower". Depth was estimated by averaging the read count align across to each gene and samples. Biological coefficient of variation(BCV) was estimated as $BCV = \sqrt{1/\delta_{0.50}}$, where $\delta_{0.50}$ was the 50% quantile across tag-wise dispersion (δ_g) estimated by "estimateTagwiseDisp" in R package "edgeR". We set effect size 1.20 and the false positive rate 0.05. For NB exact test method, we implemented their R function "est_power" in package "RnaSeqSampleSize". The parameter setting was similar as it was in Poisson test. We only need to additionally specified the estimate of maximum tag-wise dispersion parameter, which was estimated by "edgeR". Lastly, we implemented "Scotty" in MATLAB code downloaded from https://github.com/mbusby/Scotty. To compute power, we need to provide p-value cut-off to use as the metric of power, which is simply the declared DE gene cut-off. To make it comparable with all methods which have FDR control, we estimated the the exact p-value cut-off corresponds to FDR at 5% for each pilot data and input this for "Scotty". The other parameters were specified the same as other methods.

According the results in Figure 19, our method performed overall the best, especially for power prediction when N' was smaller (N'≤40). When pilot sample size was two, the prediction was a little conservative, but bias was removed if we increased pilot sample size to N=4 or 8. As pilot sample size increased from N=2 to 4 and 8, the predicted EDR gradually converged to true EDR curve. "Scotty" also had similar property but the convergence rate was smaller and the prediction was not as accurate as SeqDEsign. Except for SeqDEsign, RNASeqPower also had good accuracy for EDR prediction, but it tended to over-estimate EDR in general. Especially for smaller N', the bias was larger. For example, when N=8, to attain EDR of 80%, RNASeqPower required about N'=10 samples, and SeqDEsign required N'=18 samples. Methods based on Wald test of Poisson model (A) and exact test based on negative binomial model (C) didn't provide satisfying EDR prediction and the prediction accuracy didn't improve as pilot sample size increased.



(A) Poisson model; (B) RNASeqPower; (C) NB exact test; (D) Scotty; (E) SeqDEsign.

Figure 19: Method comparison for setting ($\delta = 50$ and fc $\geq$1.20).



The three rows indicate (A)$\delta = 40$;(B)$\delta = 50$;(C)$\delta = 60$, x axis of each figure is fc cut-off.

Figure 20: Comparisons between RNASeqPower and SeqDEsign for 12 simulation settings

We further compared SeqDEsign with RNASeqPower in more simulation settings(Figure 20). In Figure 20, we showed the MSE of RNASeqPoewr(black circle) and SeqDEsign(red triangle) in simulation settings with different N, fc cut-off(x axis) and δ (three rows). SeqDEsign outperformed RNASeqPower in almost all simulation settings.

2.7 REAL DATA APPLICATION

We applied our method to the real data HIP by assuming different pilot data sample size. We randomly sub-sampled N=2, 3, 4, 5, ..., 11, 12. (when N $\leq$ 10, we sub-sampled D=100 data for each case) Then we predicted EDR at N'=N, N+1, ..., 12, 20, 30, 40, 50, 100.

We observed that when pilot sample size is very small(N=2), p-value distribution didn't strictly follow BUM model(Figure 21). For example, many p-value distribution had heavier tail in the right hand size, In this case, if we still fit the mixture model, the corresponding EDR will be over-estimated. However, this kind of p-value distribution indicated that there were very few DE genes in the data. We defined the mean of all p-values as $\bar{p}$. To recognize the p-value distribution where mixture model estimation might fail, we compared $\bar{p}$ with 0.5, which is mean of uniform(0,1). If $\bar{p} \geq 0.5$, this indicates p-value distribution is not skew to the right and doesn't satisfy our model assumption. The associated EDR will be set to 0. Follow the same procedure of EDR computation, we compared the EDR predicted curve at each N with EDR curve generated under N=12(maximum pilot sample size). The results was shown in Figure 22. When N $\geq$ 6, the EDR prediction $\widehat{EDR}_N(N')$ seemed to converge well to $\widehat{EDR}_{12}(N')$. Although we did not know the underlying truth of EDR(N'), in this dataset, the result reasonably suggested that N$\geq$ 6 was needed and SeqDEsign was performing well.

Furthermore, we also generated results for the other methods in Figure 23. (D=10) Poisson model and NB exact test did not not look reasonable since their predicted EDR was only 15$\sim$25% even with sample size N'=100. For RNASeqPower and Scotty, the EDR curves

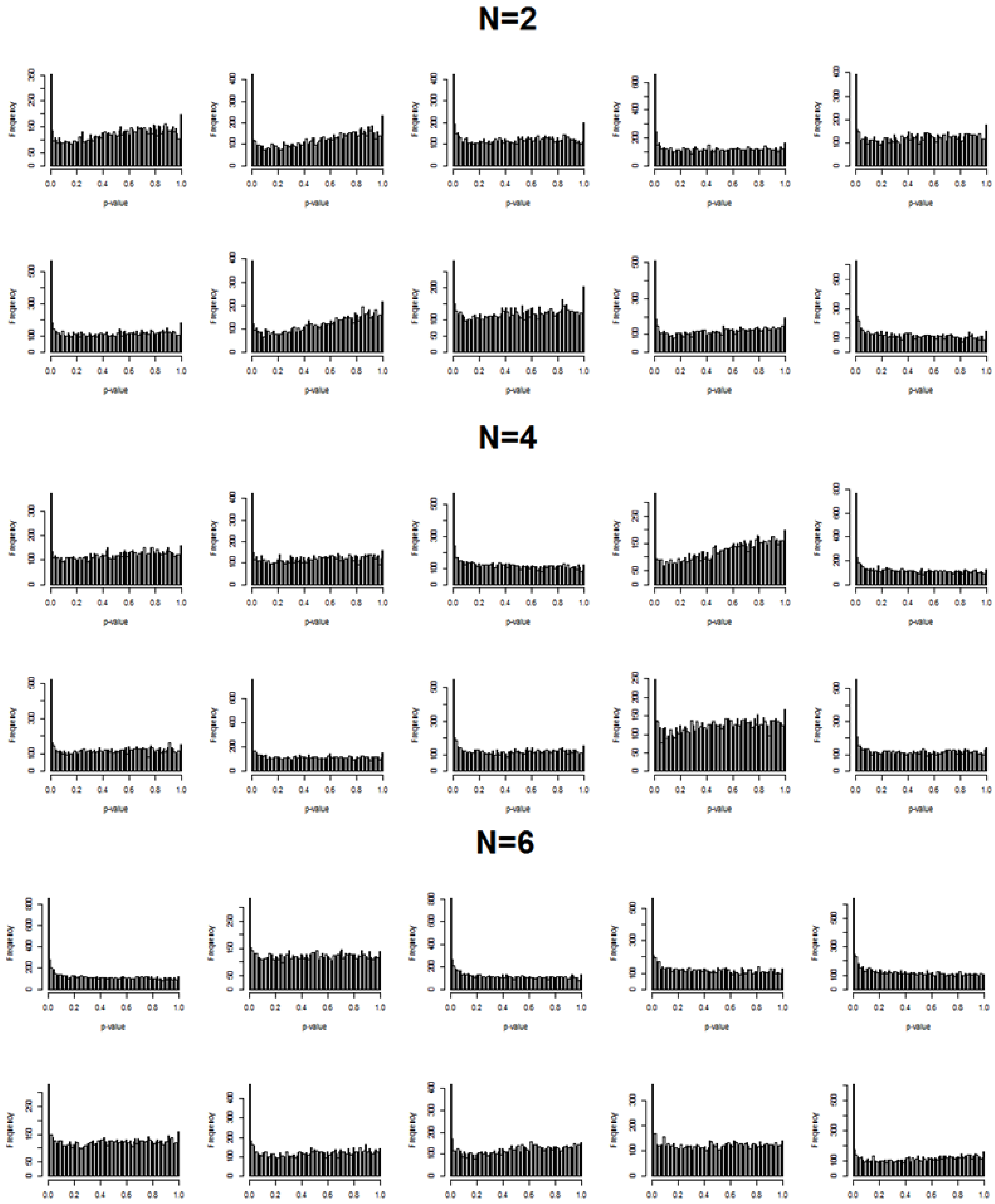


Figure 21: p-value distribution in 10 subset of real data(N=2,4,6)

were more reasonable but the result was suspicious since they claimed similar prediction performance for small pilot sample size(N=2) and large pilot sample size(N=12). The result of our SeqDEsign was more reasonable in that small pilot sample size(N=2,4,6) provided variable and inaccurate power prediction and the performance converged once the pilot sample size was large enough (N>6).

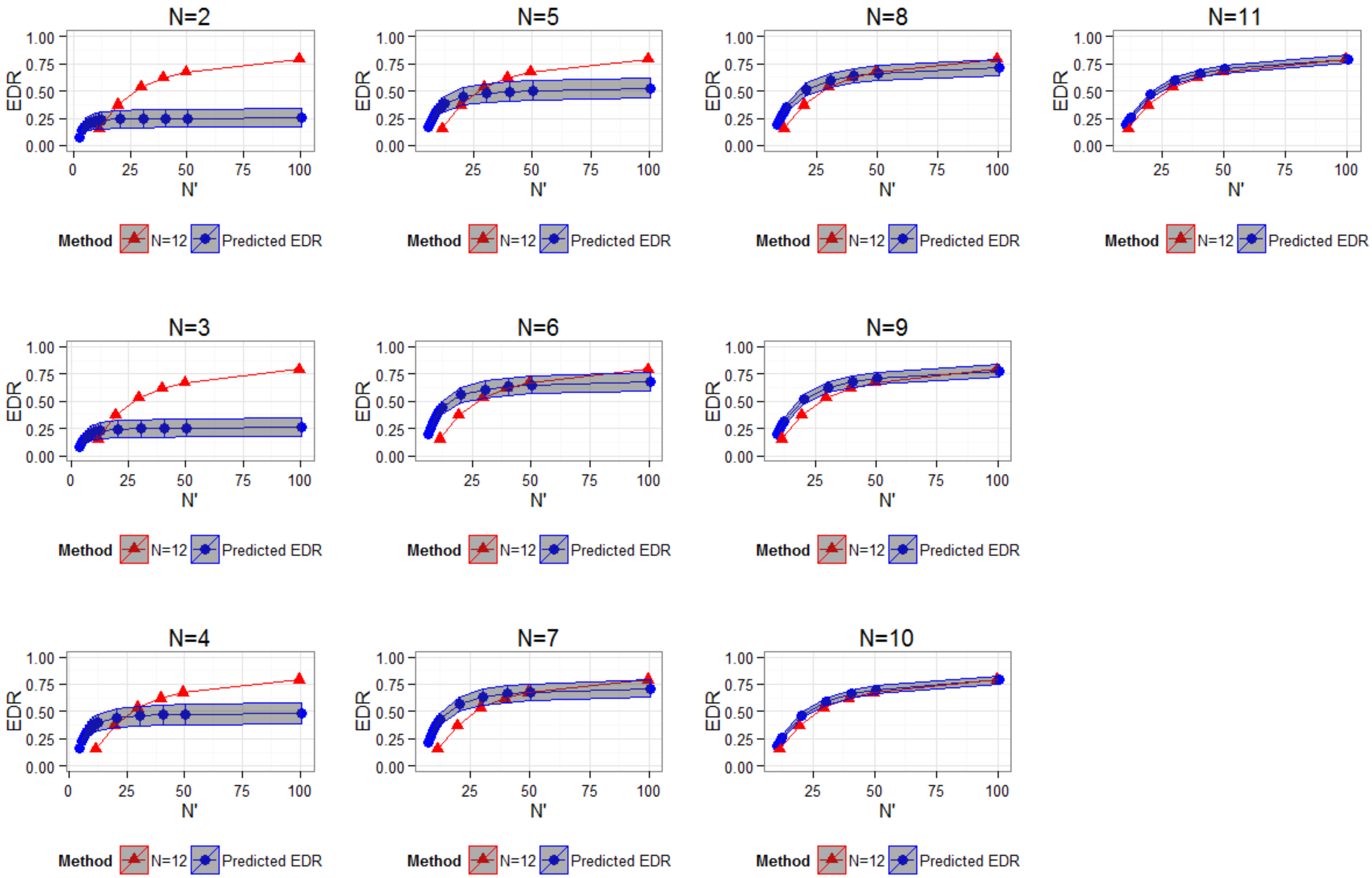
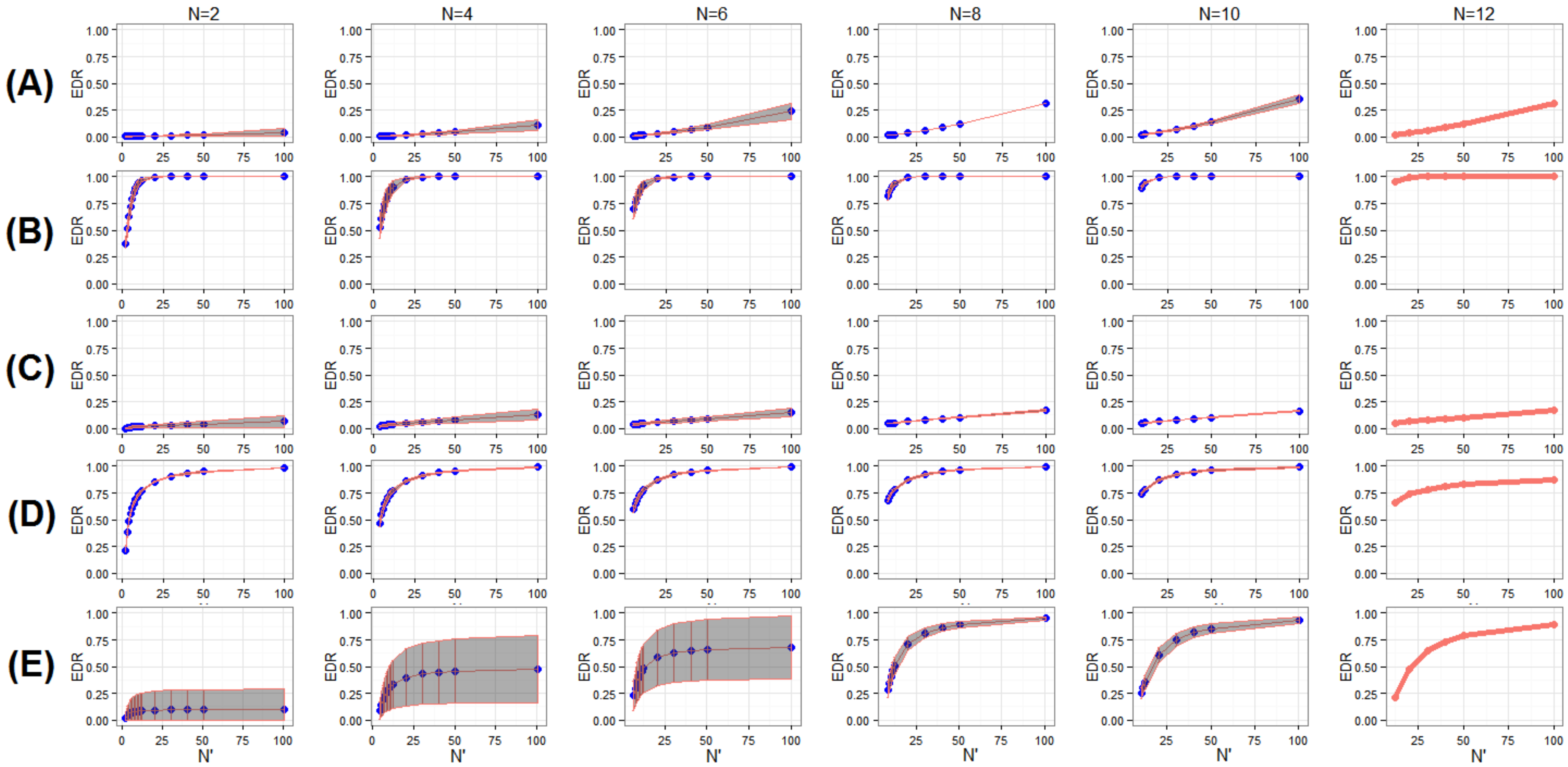


Figure 22: Real Data Application:SeqDEsign



(A) Poisson model; (B) RNASeqPower; (C) NB exact test; (D) Scotty; (E) SeqDEsign.

Figure 23: Real Data Application: Compare with four existing methods

2.8 UNBALANCED EXPERIMENTAL DESIGN

Unbalanced design is very common in genomic and clinical studies. For example, in cancer studies, due to the homogeneity of normal samples, researchers usually recruit less subjects in normal groups. Previously we focused on balanced design, where both groups had pilot sample size $N_0 = N_1 = N$ and the predicted sample size was N'. Now, we will investigate how to modify existing approaches to predict EDR at targeted unbalanced studies with sample size (N_0' and N_1'). The prediction of unbalanced experimental design involves the change of sample size and allocation ratio simultaneously. Accordingly, there are two alternative approaches to realize the power prediction.

2.8.1 Down-sampling(DS) method

From a pilot study with balanced design, we will first take sub-samples of pilot data that have the targeted allocation ratio θ'. Then, we follow the same procedure as in Chapter 2 for EDR prediction at targeted sample size.

2.8.2 Model-based(MB) method

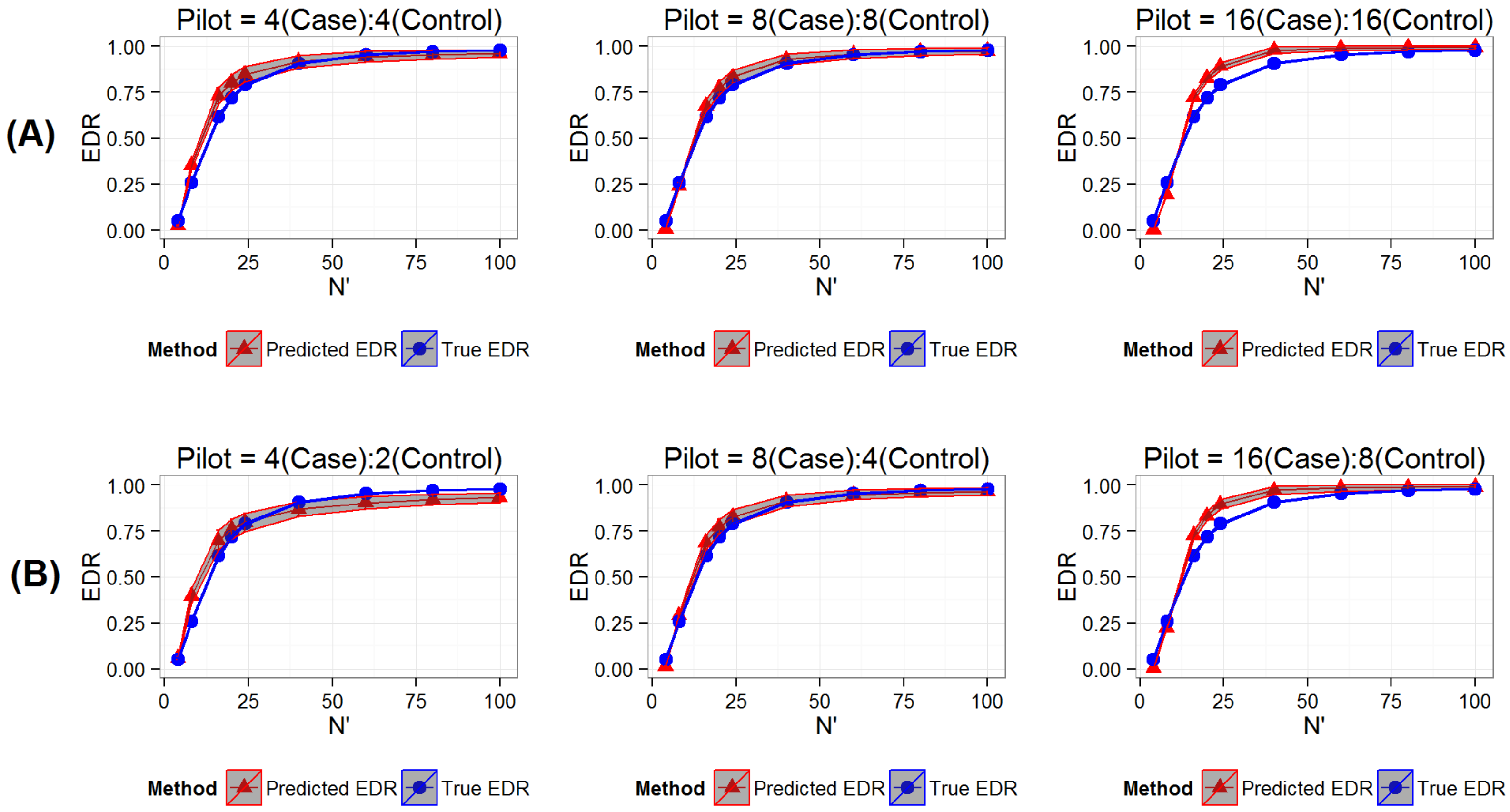
In chapter 2, we had introduced the resampling procedure based on posterior probability. In the b^{th} resampling, model-based method was solely based on the following two-way Z statistic transformation:

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \frac{\sqrt{N' \times (\frac{1+\theta e^{\hat{\beta}_{g1}}}{\theta R e^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{1+\theta}{\theta\hat{\delta}})}}{\sqrt{N \times (\frac{1+\theta' e^{\hat{\beta}_{g1}}}{\theta' R e^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{1+\theta'}{\theta'\hat{\delta}})}} + (1 - I_g^{(b)}) \times Z_g \tag{2.17}$$

where θ is the allocation ratio of pilot data and θ' is the allocation ratio for targeted design.

2.8.3 Simulation study

We conducted a simulation study to test for the EDR prediction of unbalanced design. In this setting, δ=50 and $fc \geq 1.20$. We started with pilot data in balanced design (N=4,8,16) and wanted to predict EDR in unbalanced design ($N_1' = 2 \cdot N_0'$ and $N_0' = 8, 10, 12, 20, 30, 40, 50$). Under this simulation setting, we generated 50 datasets. Resamping procedure was repeat for B=50 times. We compared the performance of Down-sampling and Model-based method for those simulation settings.



(A) Model-based method; (B) Down-sampling method.

Figure 24: Predict EDR of unbalanced experimental design

Table 9: Summary table for MSE of (A) Model-based and (B) Down-sampling method

Method	N=2	N=4	N=8
Model-based	0.00365	0.00060	0.00028
Down-sampling	0.00419	0.00204	0.00075
SeqDEsign	0.0182	0.0174	0.0168

Figure 24 showed the performance of methods for the two approaches. (A) is model-based method and (B) is Down-sampling approach. The x axis (N) is the sample size for predicted case group. Since allocation ratio is constant(2), control group sample size is N/2. In terms of MSE as indicated in Table 9, the two methods had similar good performances in predicting EDR of unbalanced design. For computational convenience, consistency to our previous approach and slightly better performance, we will adopt the model-based approach for unbalanced design.

We further compared model-based approach with RNASeqPower which can also predict power of unbalanced design under the simulation model. Figure 25 indicated that our method had better performance.

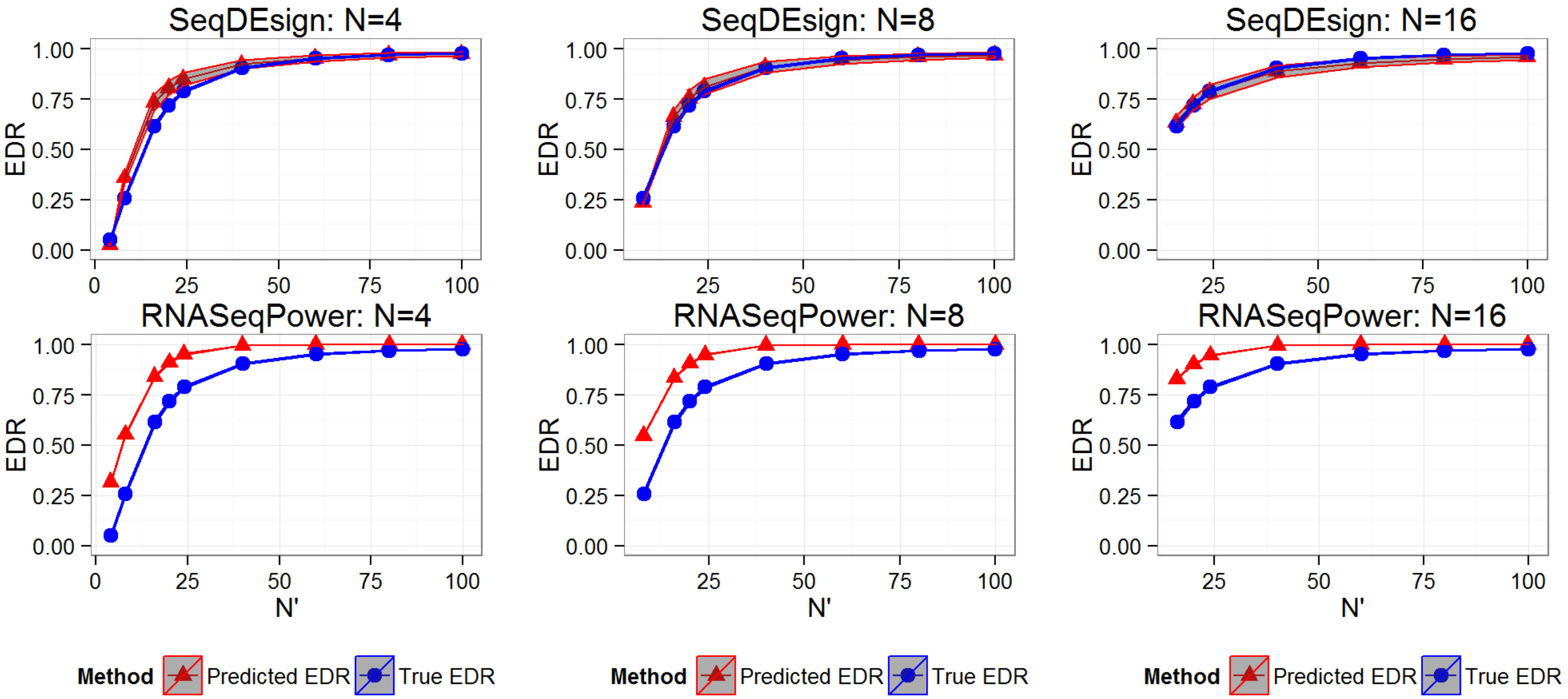


Figure 25: Comparison between SeqDEsign and RNASeqPower under unbalanced design setting

2.9 SUMMARY AND DISCUSSION

In this chapter, we have proposed and compared several alternative approaches to predict EDR at targeted sample size(N') based on a pilot data (N, R). By evaluating each method in terms of MSE, we found that Storey and Tibshirani's method BUM model and CDD BUM model gave us overall best EDR prediction. We further compared Storey and Tibshirani's method BUM model with four other existing methods. Our proposed methods gave asymptotically best EDR prediction as pilot sample size increased. Even when pilot sample size was small (e.g., N=2), the prediction was only a little bit conservative which was acceptable. In the comparative study with four other methods, our methods performed overall best in the simulation settings.

For parametric model fitting, originally we also proposed parametric bootstrap approach as in PowerAtlas[Gadbury et al., 2004]. But since posterior approach had similar and better performance and is more convenient for two way EDR prediction, we decided to adopt posterior approach.

Our methods showed superior characteristics over existing methods in the investigation of EDR prediction of targeted sample size. To design a practical algorithm of power calculation for RNA-Seq, we will also take consideration of the read depth in the EDR computation in the next chapter.

3.0 SAMPLE SIZE, READ DEPTH AND GENOME-WIDE POWER PREDICTION

So far, we have investigated the genome-wide power prediction in the direction of increasing sample size. For Next Generation Sequencing data like RNA-Seq, higher read depth generates more reads, which increases the statistical power to detect DE genes [Liu et al., 2013, Tarazona et al., 2011]. Therefore, it is important that power calculation method should consider the impact of read depth, yet only two existing methods (RNASeqPower and Scotty) included this impact factor. Given the superior performance of SeqDEsign, we now further extend the current framework to predict EDR in various read depth selections. In this chapter, we will first discuss two alternative approaches proposed and compare their performance. By fitting two-way inverse power law model, we will construct 3-dimensional predicted power surface. A consequent cost benefit analysis of optimal experiment design based on this 3-dimensional power surface will be discussed in chapter 4.

3.1 NESTED DOWN-SAMPLING(NDS) AND MODEL-BASED(MB) APPROACH FOR GENOME-WIDE POWER PREDICTION FOR A FUTURE READ DEPTH

Here we proposed and compared two approaches to extend power calculation to two-dimensional EDR prediction including both sample size and sequencing depth (i.e. $\widehat{EDR}_{(N,R)}(N', R')$). The first approach is nested down-sampling(NDS) procedure and the other approach is model-based(MB) method. In both methods, we used Storey and Tibshirani's method in mixture model fitting.

3.1.1 Nested down-sampling(NDS) method

Suppose we start with pilot data with sample size N and read depth $R = 650G$. We can write the data matrix as $S = \{\overrightarrow{s_{01}}, ..., \overrightarrow{s_{0N}}, ..., \overrightarrow{s_{11}}, ..., \overrightarrow{s_{1N}}\}$. Each sample is consist of G elements, each presenting the read counts aligned to a certain gene. For subject i in group j, $\overrightarrow{s_{ij}} = \{e_{1ij}, ..., e_{gij}, ..., e_{Gij}\}$. We can expand all reads into a larger pool by rewriting $\overrightarrow{s_{ij}} = \{\underbrace{1, ..., 1}_{e_{1ij}}, \underbrace{2, ..., 2}_{e_{2ij}}, ..., \underbrace{G, ..., G}_{e_{Gij}}\}$, indicating which gene each read is coming from. We would then sample p% of reads without replacement from $\overrightarrow{s_{ij}}$ and collapse the reads into gene level again so the resulting sample is s_{ij}^p. We could further perform downward sampling from s_{ij}^p and repeat the procedure to generate sample with smaller read depth. Following this fashion, we can represent the nested samples as $\overrightarrow{s_{ij}}^{650}$, $\overrightarrow{s_{ij}}^{500}$,$\overrightarrow{s_{ij}}^{400}$,$\overrightarrow{s_{ij}}^{300}$,$\overrightarrow{s_{ij}}^{200}$,$\overrightarrow{s_{ij}}^{100}$. The superscript indicates the average coverage for each gene. Then we can follow the same steps as proposed in chapter 2.3 and Figure 5. If we represent the pilot data as S: N × R, we first down-sample it to S':N × R' and then estimate $\widehat{EDR}_{(N,R')}(N', R')$ and compared with $\widehat{EDR}(N', R')$.

3.1.2 Model-based(MB) method

While the nested downward sampling approach is more straightforward, it requires more computation time since we need to repeat the subsampling procedure for multiple times, followed by repeated sampling of DE gene status. Besides, nested downward sampling procedure could only predict EDR at R' $\leq$ R. We have to applied surface fitting procedure to get predicted EDR at R' $\geq$ R. In comparison, an easier way is to modified our previous posterior procedure in the statistics transformation step. The underlying assumption is that estimate of β_{g0} and β_{g1} will not be changed with sample size and read depth.

Suppose we have a pilot data with sample size N and read depth R. If we perform the hypothesis testing for each gene g with Z statistics $Z_g^{(b)}$ in the b^{th} simulation, the transformation step is achieved by applying the following transformation:

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \frac{\sqrt{N'} \times (\frac{1+\theta e^{\hat{\beta}_{g1}}}{\theta R e^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{(1+\theta)}{\theta\hat{\delta}})}{\sqrt{N} \times (\frac{1+\theta e^{\hat{\beta}_{g1}}}{\theta R' e^{\hat{\beta}_{g0}+\hat{\beta}_{g1}}} + \frac{(1+\theta)}{\theta\hat{\delta}})} + (1 - I_g^{(b)}) \times Z_g \tag{3.1}$$

After computation of the transformed Z statistics, we can follow the proposed steps to compute predicted EDR(Pow(N',R')).

3.2 TWO-WAY INVERSE POWER LAW SURFACE FITTING

Inverse power law model is frequently fitted to model learning curve created by a small training dataset in the machine learning field. Curve fitting is carried out by nonlinear least square optimization[Figueroa et al., 2012]. Here we propose two-way inverse power law curve fitting to model the power surface by a function of sample size and read depth. The power function could be written as:

$$EDR = Pow(N', R') = a - b \times N'^{-c} - d \times R'^{-e} \tag{3.2}$$

where a,b,c,d,e are all positive numbers. As N and R increase to infinity, EDR will be approximate to 1. Therefore, we constrain a to be 1 exactly. Specifically, we estimate the remaining four parameters by R function "optim" in "stats" R package using BFGS quasi-Newton method. For NDS method, by fitting two-way inverse power law, we can extrapolate EDR of experimental design with higher read depth(R'> R) than the pilot data.

3.3 SIMULATION STUDIES

3.3.1 Simulation settings

The flowchart of SeqDEsign for two way EDR prediction was shown in Figure 26. As described previously, here we want to compare two approaches to predict EDR(N',R'): (1) Nested downward sampling; (2) Transformation of statistics in bootstrapping step.

In simulation studies, we followed the similar settings as in chapter 2. The pilot data was generated with sample size N=2, 4 and 8 with dispersion parameter δ=40, 50, 60. In each combination of N and R, we had four different log-fold-change cur-off for DE genes. (0.2, 0.26, 0.32 and 0.38) The read depth of pilot data was fixed at R=G×650. (G=10^4) To generate the true EDR curve, we simulated data under each predicted setting (N',R'). N' were selected to be 5, 10, 20, 30, 40, 50, 100, and R' were selected to be 1M, 2M, 3M, 4M, 5M, 6M, 6.5M(pilot), 7M, 8M, 9M, 10M, 11M and 12M. In each pilot data setting, D=10 data were generated. For true data, we also generated D=10 data follow the same distribution. Posterior procedure was repeated for B=20 times.

3.3.2 Comparisons between Model-based and Nested down-sampling method

We first compared the predicted EDR and true EDR by mean square error without surface fitting. For Model based method, MSE under pilot study (N,R) was computed as

$$MSE(N,R) = \sum_{N',R'} (\overline{\widehat{EDR}}_{(N,R)}(N',R') - \overline{EDR}(N',R'))^2 \tag{3.3}$$

For Nested down-sampling method, MSE under pilot study (N,R) was computed as

$$MSE(N,R) = \sum_{N',R'} (\overline{\widehat{EDR}}_{(N,R')}(N',R') - \overline{EDR}(N',R'))^2 \tag{3.4}$$

,where $\overline{\widehat{EDR}}_{(N,R)}(N',R')$ is the average of predicted EDR under setting (N',R'), and $\overline{EDR}(N',R')$ is the average of true EDR.

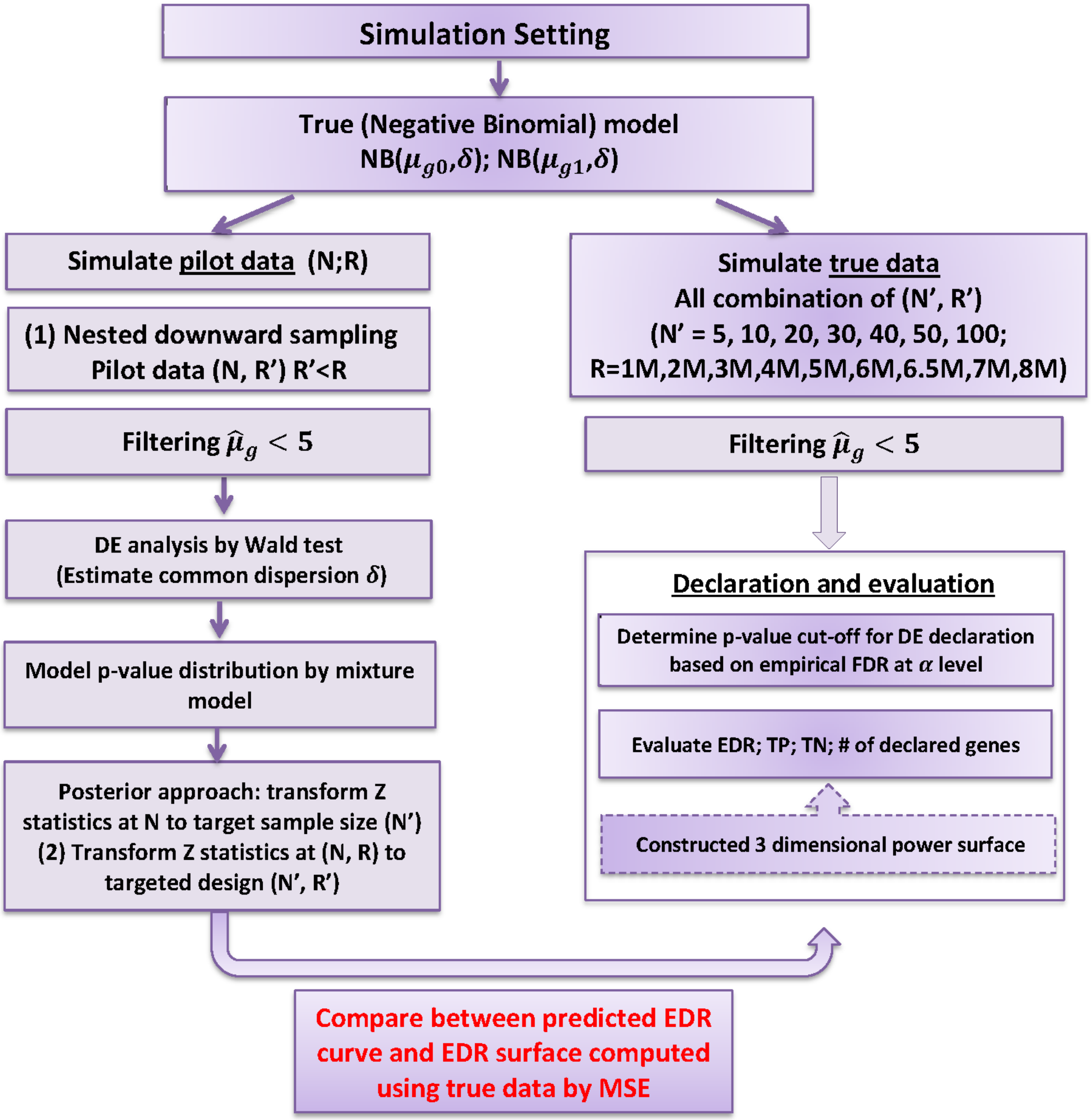
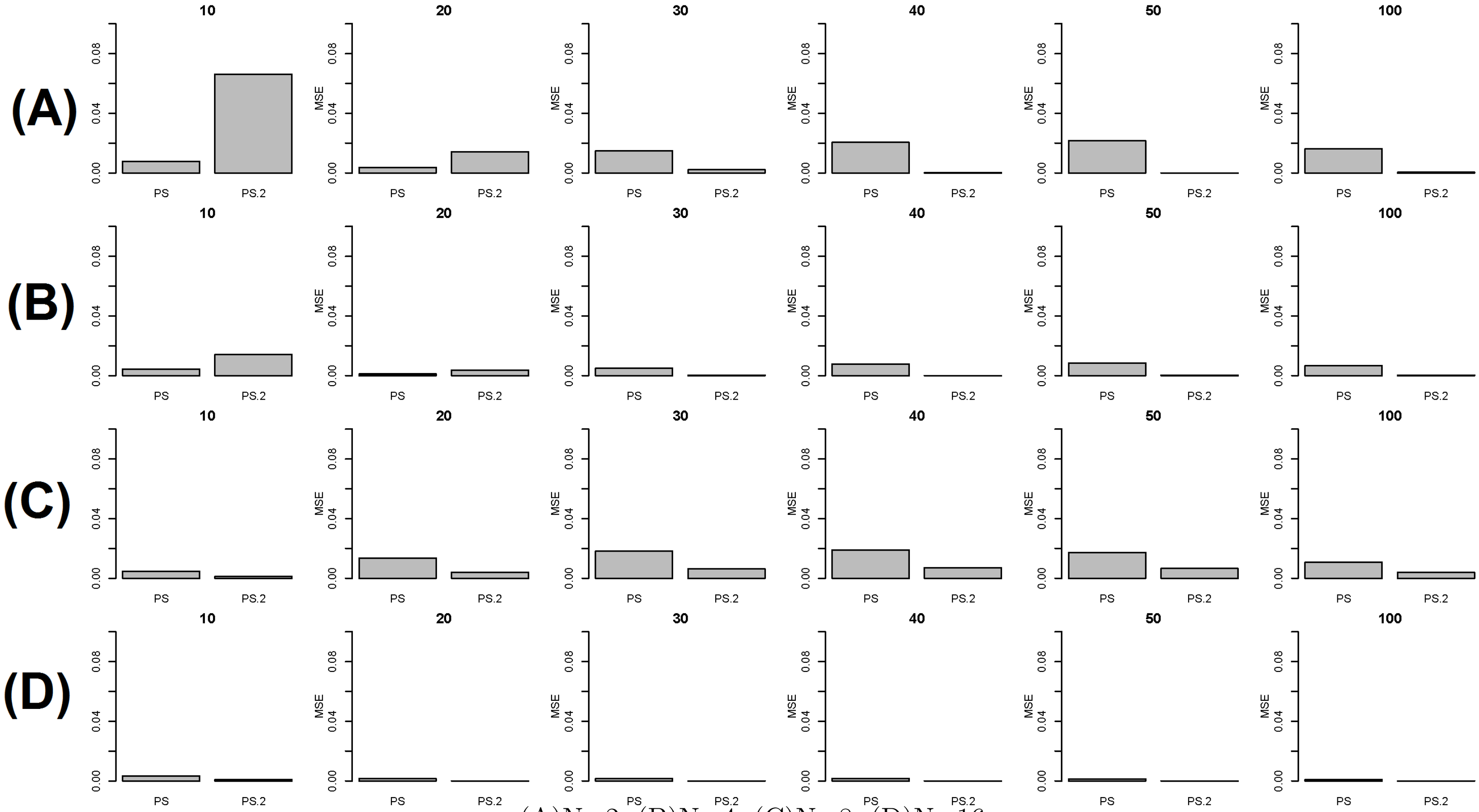
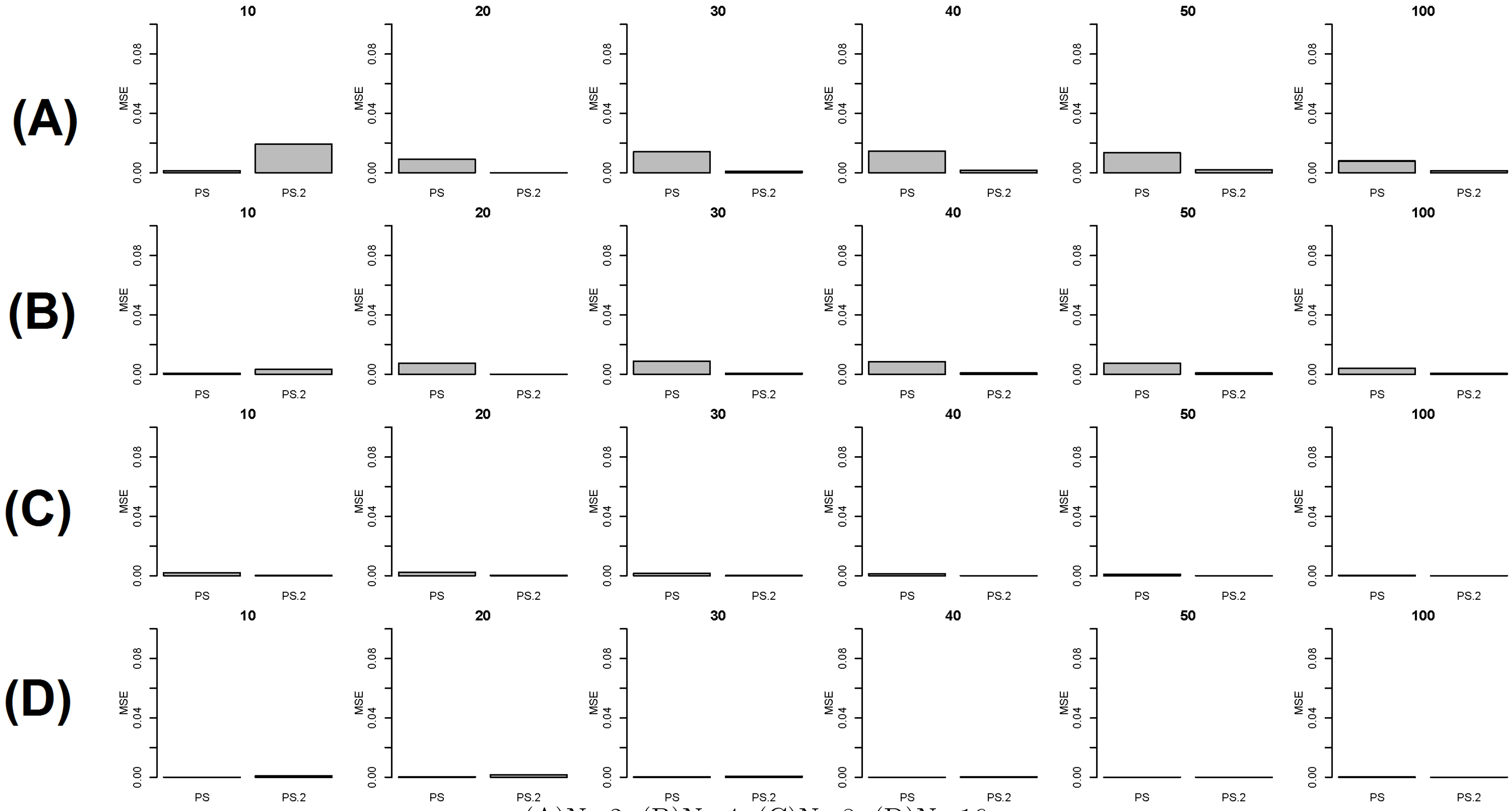

Simulation Setting
True (Negative Binomial) model
NB(μ_{g0},δ); NB(μ_{g1},δ)
Simulate pilot data (N;R)
(1) Nested downward sampling
Pilot data (N, R') R'<R
Filtering $\hat{\mu}_g < 5$
DE analysis by Wald test
(Estimate common dispersion δ)
Model p-value distribution by mixture model
Posterior approach: transform Z statistics at N to target sample size (N')
(2) Transform Z statistics at (N, R) to targeted design (N', R')
Simulate true data
All combination of (N', R')
(N' = 5, 10, 20, 30, 40, 50, 100;
R=1M,2M,3M,4M,5M,6M,6.5M,7M,8M)
Filtering $\hat{\mu}_g < 5$
Declaration and evaluation
Determine p-value cut-off for DE declaration based on empirical FDR at α level
Evaluate EDR; TP; TN; # of declared genes
Constructed 3 dimensional power surface
Compare between predicted EDR curve and EDR surface computed using true data by MSE

Figure 26: Flowchart of SeqDEsign when predicting EDR at targeted sample size N' and read depth R'



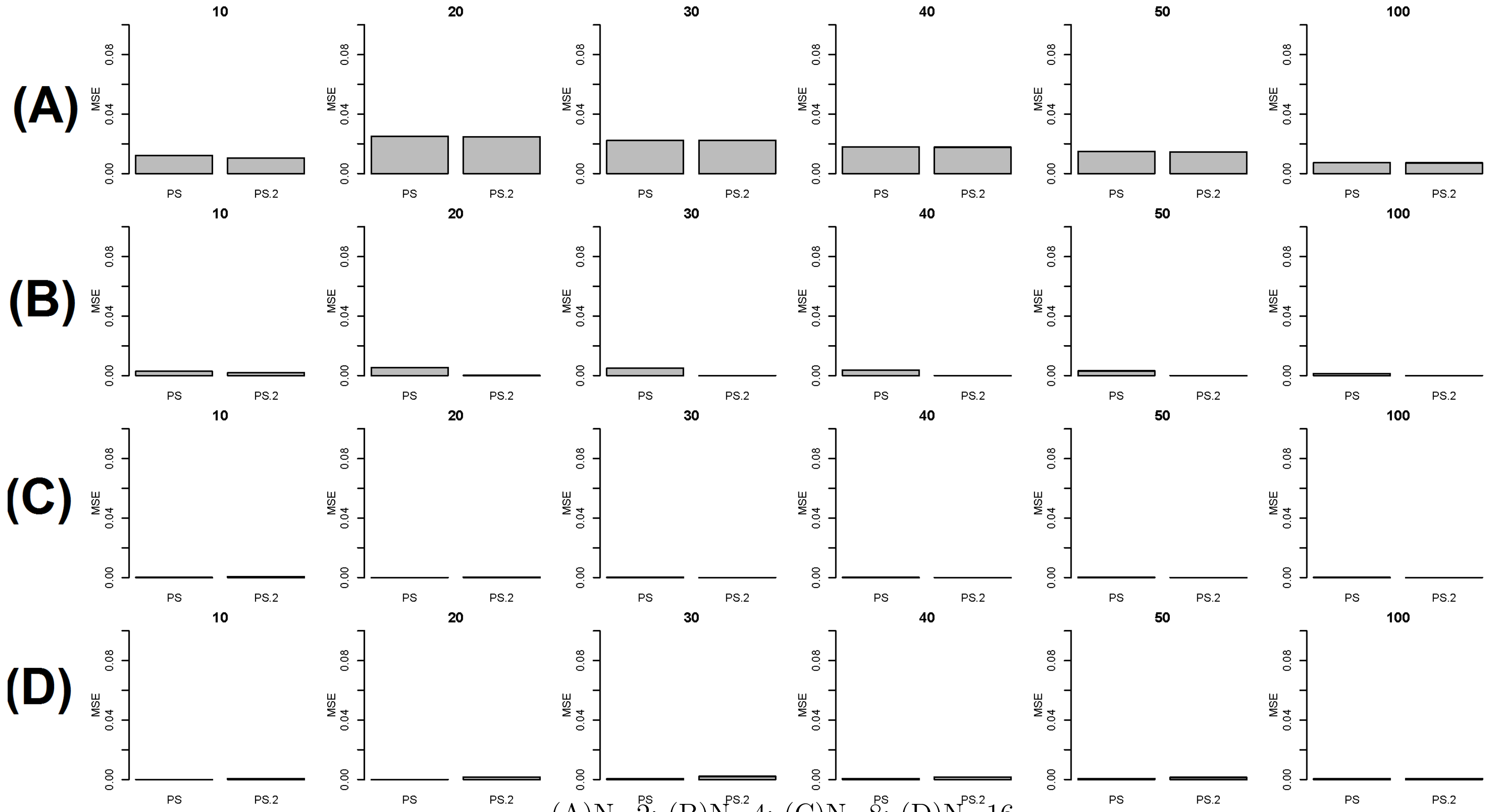
(A)N=2; (B)N=4; (C)N=8; (D)N=16.

Figure 27: Compare between MB and NDS(fc≥1.15)



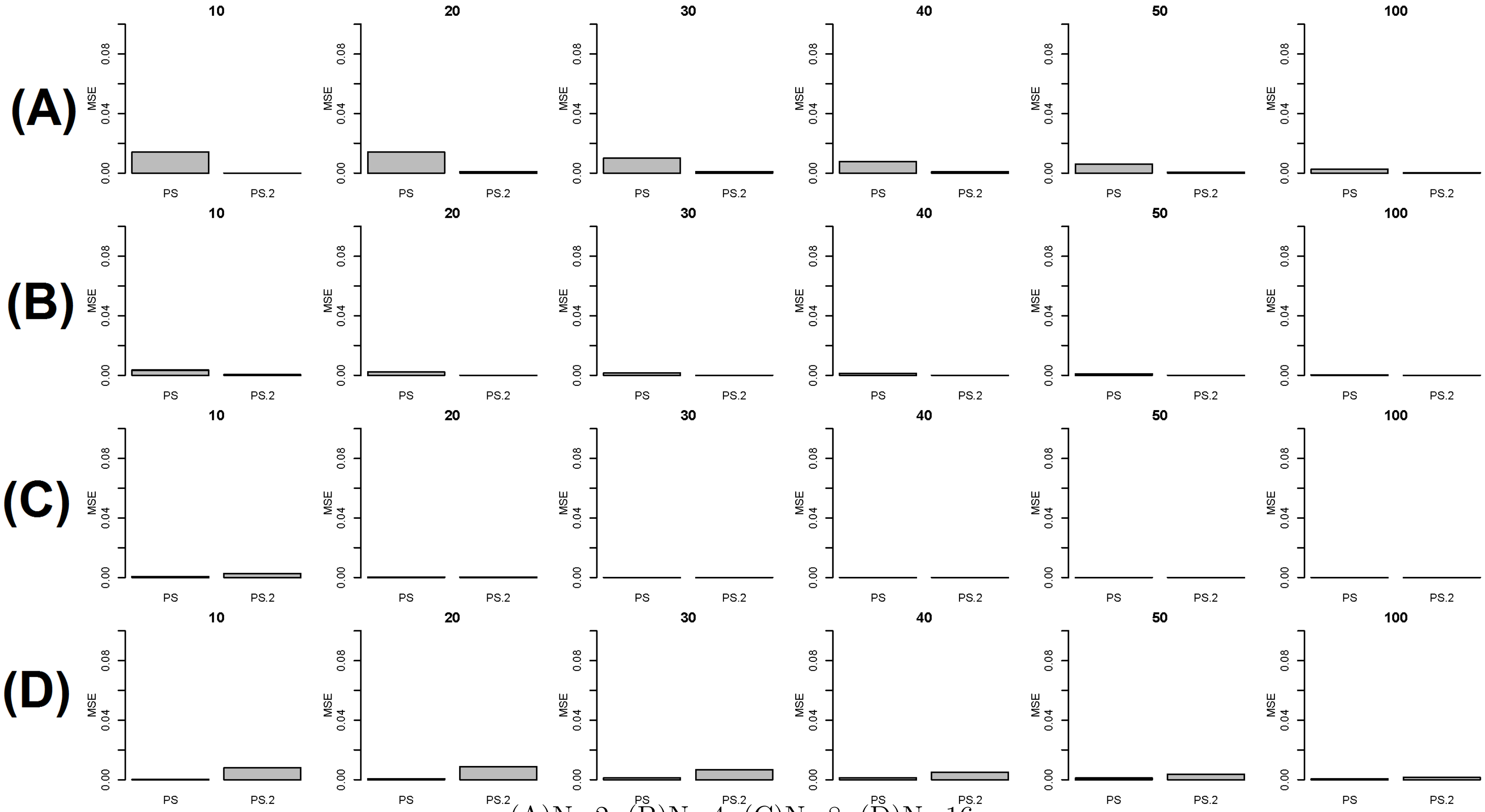
(A)N=2; (B)N=4; (C)N=8; (D)N=16.

Figure 28: Compare between MB and NDS(fc≥1.20)



(A)N=2; (B)N=4; (C)N=8; (D)N=16.

Figure 29: Compare between MB and NDS(fc≥1.25)



(A)N=2; (B)N=4; (C)N=8; (D)N=16.

Figure 30: MB and NDS(fc≥1.30)

Figure 27 $\sim$ Figure 30 showed the MSE of two methods comparing with predicted true EDR under different simulation settings(4 different fc cut-off) separated by N'(MSE(N,R,N')). PS is nested down-ward sampling method and PS.2 is model-based method. Overall, model-based method has smaller MSE in more scenarios comparing with nested down-sampling method and have several appealing advantages: (1) it gives better EDR prediction; (2) it is computationally more efficient; (3) it could predict EDR at a wider range of read depth.

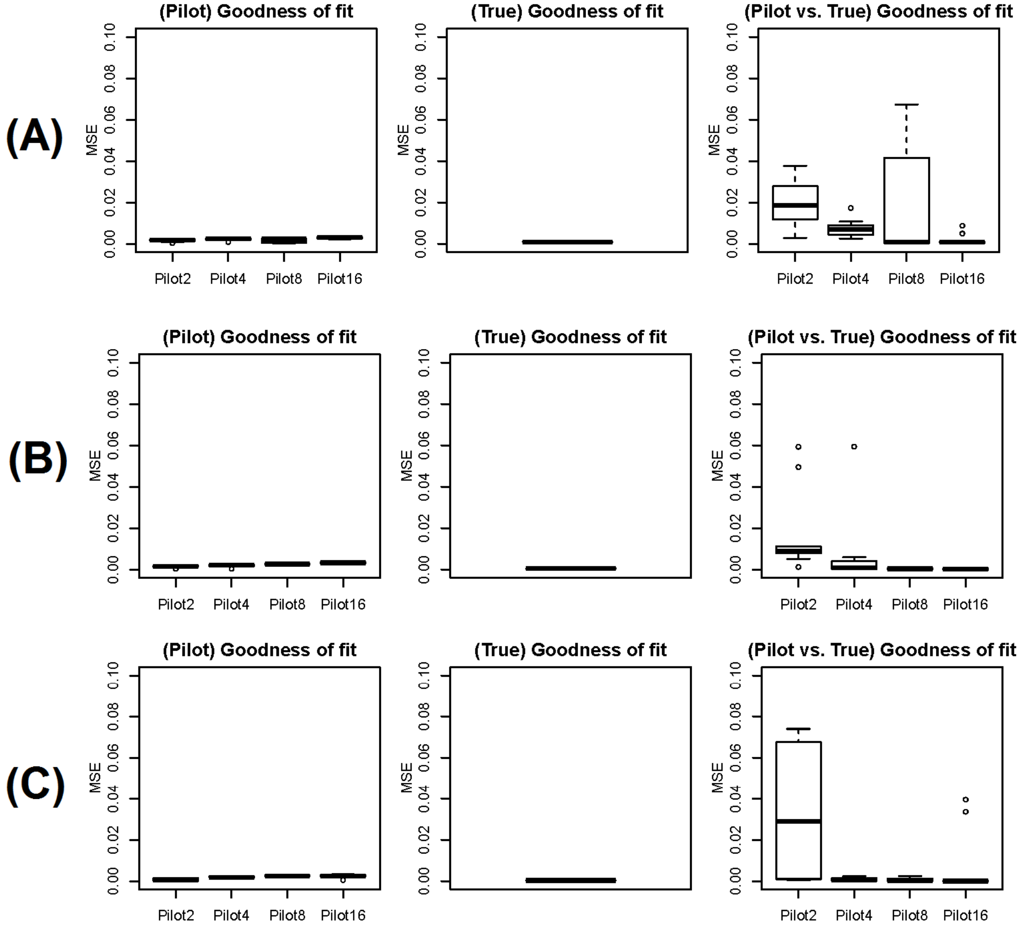
3.3.3 Inverse power law fitting

Our ultimate goal is to predict best optimal experimental design(N*,R*) for RNA-Seq. Therefore, it's necessary for us to predict EDR surface, which is a function of N and R. Here, given the estimated EDR from model-based method, we fitted two-way inverse power law previously proposed. Under each simulation setting, we have D=10 pilot data simulated from same underlying model. Then ten predicted EDR surface could be constructed.

Figure 31 showed the Goodness-of-fit of inverse power law fitting for predicted EDR surface, true EDR surface and third column is the MSE of predicted EDR surface comparing with true EDR surface. The EDR surface fitting for predicted EDR and true EDR was pretty good. There was however a bump in MSE between predicted EDR surface and true EDR surface when pilot sample size N=8 (A). After diagnosis of the problematic data, we found that there're some problem about the mixture model fitting when using Storey and Tibshirani's method under certain situation settings. Due to the previous equally good performance of CDD method, we also tried this approach. It estimated λ by non-parametric method and the resulting EDR prediction was shown in Figure 32. The result was much better without any outliers.

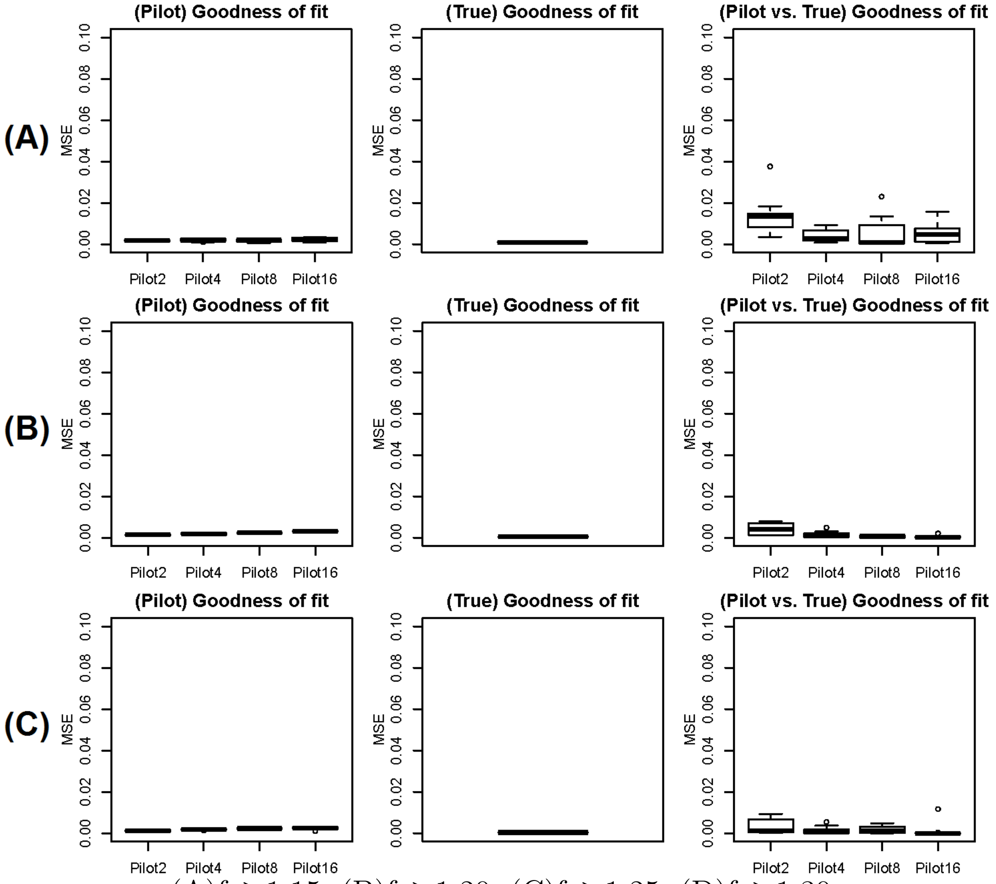
3.3.4 Fitted EDR surface

Here we demonstrate the EDR surface prediction in several examples. Figure 33(N=2), Figure 34(N=4) and Figure 35(N=16) are 3 dimensional plots with both true EDR sur-



(A)fc≥1.15; (B)fc≥1.20; (C)fc≥1.25; (D)fc≥1.30.

Figure 31: Goodness-of-fit of inverse power law fitting model(Storey and Tishirani's method)



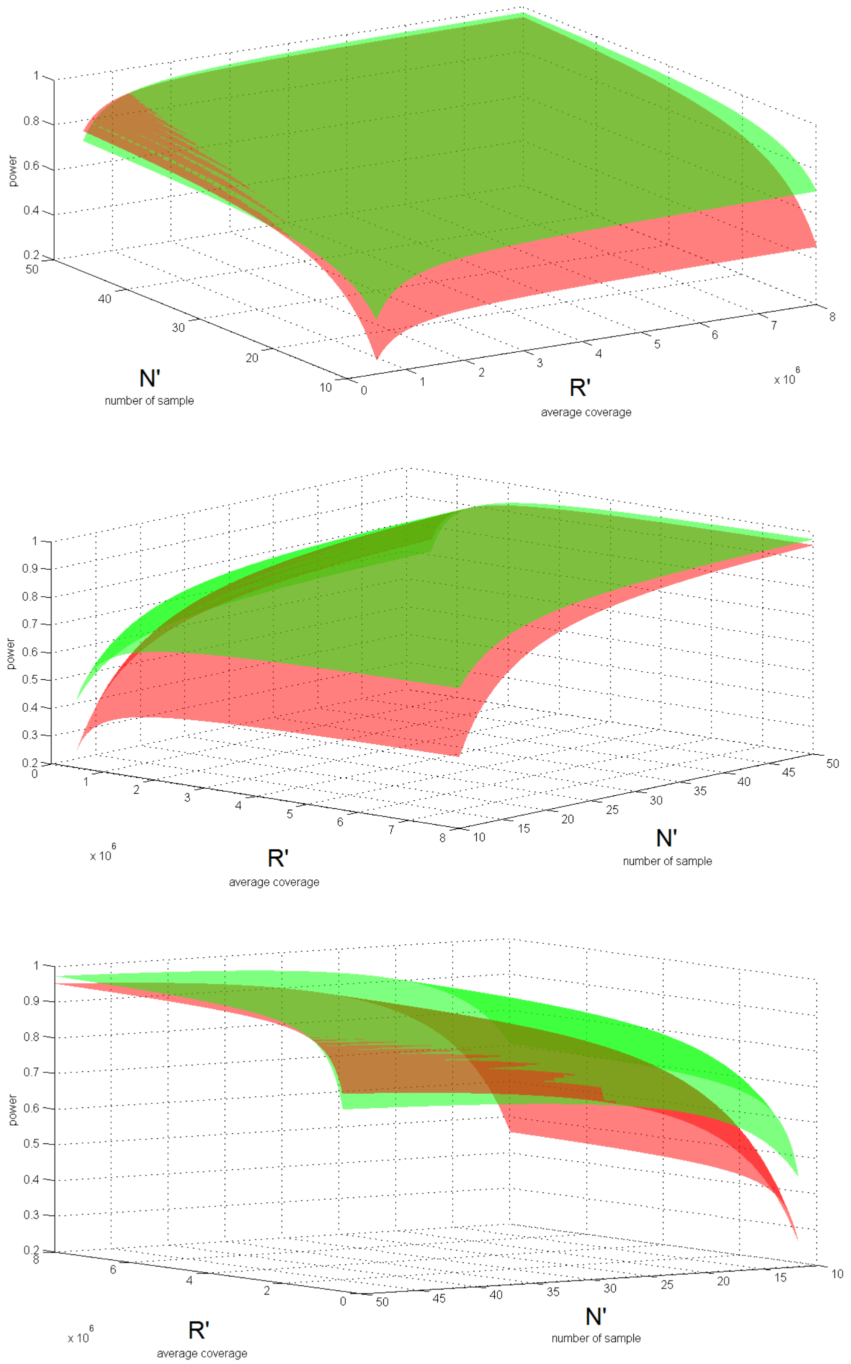
(A)fc≥1.15; (B)fc≥1.20; (C)fc≥1.25; (D)fc≥1.30.

Figure 32: Goodness-of-fit of inverse power law fitting model(CDD)

face(red) and predicted EDR surface(green). The increment of pilot study sample size helps the EDR prediction according to our results. The estimated true EDR surface is $Pow(N', R') = 1 - 21.07590 \times N^{1.595355} - 10^6 \times R^{1.173074}$, while predicted EDR surface for N=2 is $\widehat{Pow}(N', R') = 1 - 16.3058 \times N^{1.75834} - 10^6 \times R^{1.151303}$, for N=4, $\widehat{Pow}(N', R') = 1 - 24.97404 \times N^{-1.778650} - 10^6 * R^{-1.137705}$, and for N=16, $\widehat{Pow}(N', R') = 1 - 41.23273 \times N^{-1.862118} - 10^6 * R^{-1.136118}$

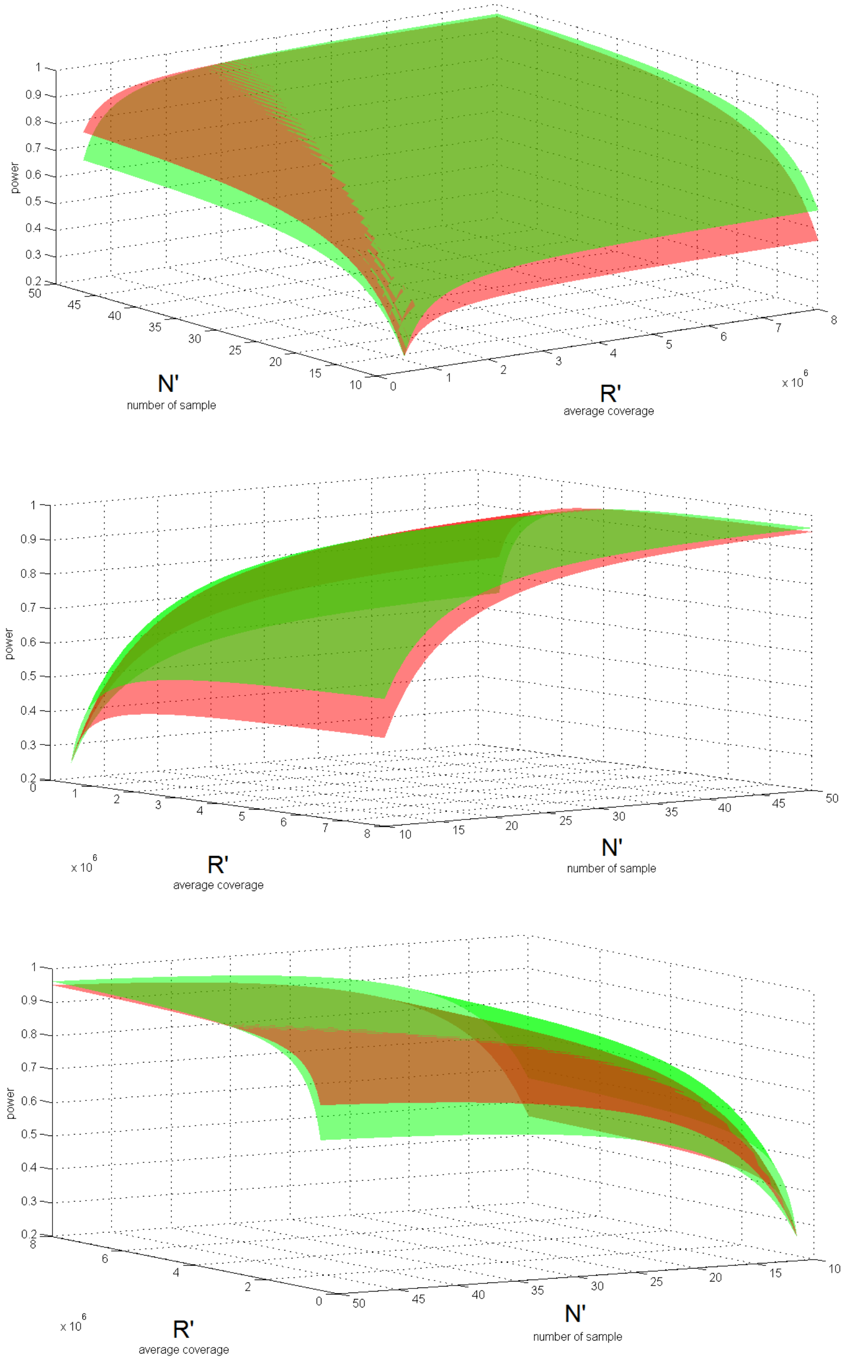
3.4 SUMMARY

In this chapter, we extended EDR prediction to the direction of read depth and constructed 3-dimensional power surface by inverse power law fitting. Comparing with underlying true EDR surface, our method had fairly good EDR prediction. Model-based method had advantages of better prediction accuracy, wider EDR prediction coverage and higher computational efficiency compared with nested down-sampling approach. We will adopt the model-based approach in the later sections and the software package.



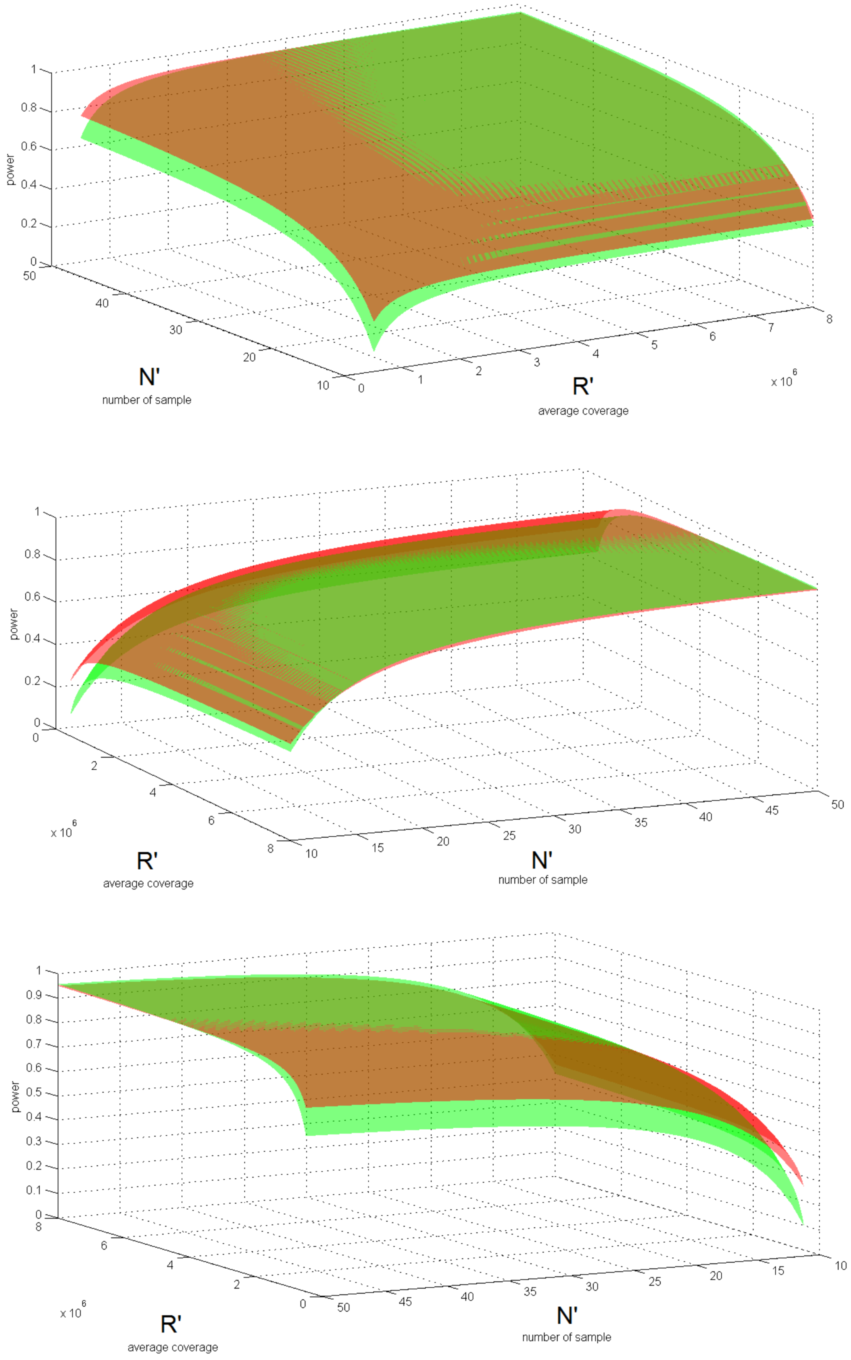
red surface-True EDR; green-Predicted EDR.

Figure 33: Compare true EDR and predicted EDR surface(δ=50,fc$\geq$1.20,pilot N=2)



red surface-True EDR; green-Predicted EDR.

Figure 34: Compare true EDR and predicted EDR surface(δ=50,fc$\geq$1.20,pilot N=4)



red surface-True EDR; green-Predicted EDR.

Figure 35: Compare true EDR and predicted EDR surface(δ=50,fc$\geq$1.20,pilot N=16)

4.0 COST BENEFIT ANALYSIS FOR RNA-SEQ EXPERIMENT

In RNA-Seq experiment, we usually have limited budget, which brings us to develop methods for power calculation in designing a powerful and affordable experiment in detecting DE genes. Therefore, it's important to design the cost function and generate the optimal design given cost constraints. In this chapter, we further defined a reasonable cost function. Based on fitted three dimensional EDR surface and cost constraints, we conducted cost-benefit analysis to compute the optimal design. By comparing our predicted optimal design with the underlying true optimal design in simulation studies, we demonstrated the superiority of our method. We also discussed ways to predict desirable experimental design to meet certain criteria (a EDR lower bound, sample size upper bound, etc.) when there is no clear constraint of cost function. Specifically, we want to answer the following practical questions about desirable experiment design (N*,R*):

Q1: With a fixed maximum total cost C, what is the optimal design?

Q2: To reach a certain EDR level (EDR'), what are all possible experimental design?

Q3: With a maximum sample size N_{max} and a targeted EDR (EDR') to reach, what are all possible experimental design?

Q4: Given a dataset with (N,R) from a RNA-Seq experiment, is it worthwhile to increase the sample size or sequencing depth?

We designed cost functions and provided visualization tools to assist users to select the optimal experiment design according to their specific needs.

4.1 DESIGN OF COST FUNCTION

A reasonable cost function C could be defined as follows:

$$C = B(N', R') = 2 \times N' \times (A + B \times R'/10^6) \tag{4.1}$$

Here N' is the targeted sample size for each group and R' is targeted total reads. To give better interpretation, here we divide R' by 10^6. So A is the sample collection cost per sample(which includes cost to recruit a patient, collect and preprocess the sample and parify mRNA etc.), and B is the sequencing cost for per sample per million aligned reads. We use A=\$570; B=\$400 throughout this section to demonstrate the method.

4.2 OPTIMAL EXPERIMENT DESIGN WITH COST CONSTRAIN

With the cost function defined previously and the fitted three dimensional EDR surface by two-way inverse power law, we can compute optimal RNA-Seq experiment design under cost constraints. Below, we showed one simulation example of the computation of optimal design(N^*,R^*).

Figure 36 showed the fitted three dimensional power surface. Both power and cost are increasing functions of N and R. (EDR=Pow(N,R) and C=B(N,R)) There is a requirement that the total cost of RNA-Seq experiment should not exceed C_{max}. When cost function intersects with power surface, the experimental design corresponds to the optimal power is the optimal design.

$$(N^*(C), R^*(C)) = \underset{N', R', C < 2 \times N' \times (A + B \times R'/10^6)}{\text{argmax}} (\hat{a} - \hat{b} \times N'^{(-\hat{c})} - \hat{d} \times R'^{(-\hat{e})}) \tag{4.2}$$

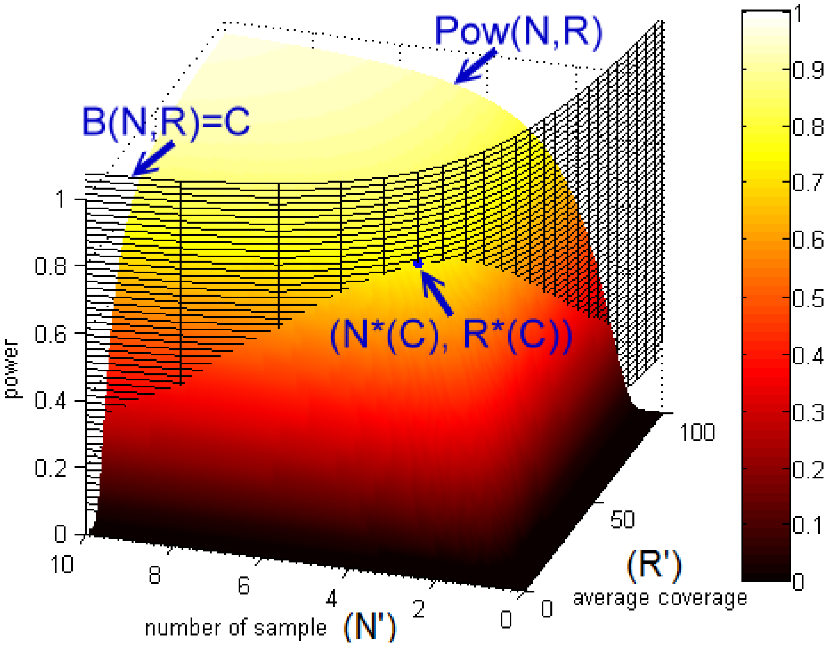
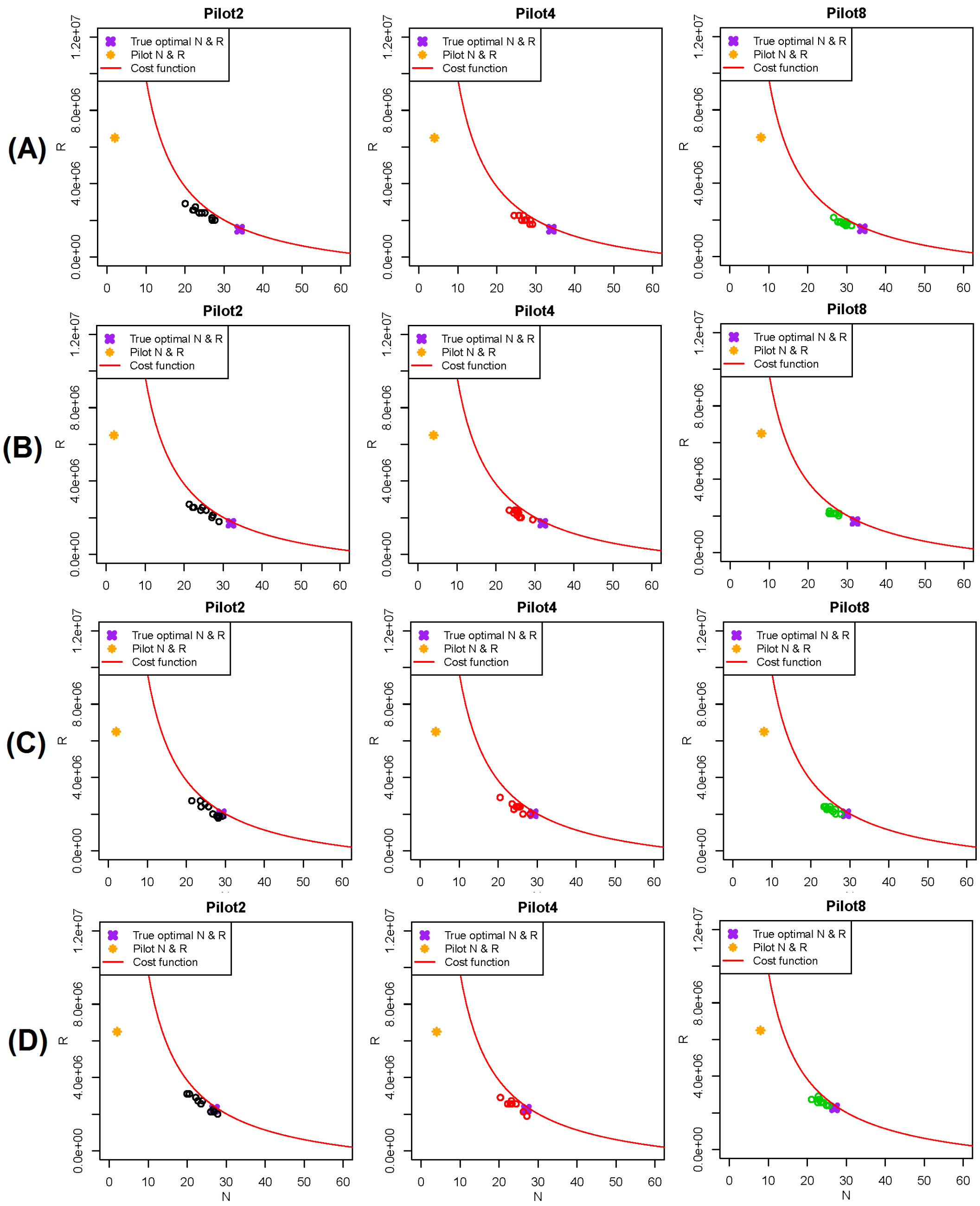


Figure 36: Three dimensional power surface

4.3 SIMULATION

We have conducted cost benefit analysis based on simulation settings in chapter 3, when $\delta = 50$, fc≥1.15;≥1.20;≥1.25;≥1.30, and N=2,4,8.

To address Q1 previously discussed, here we defined the cost function as in (4.1). The total cost C was set to be 80,000. Figure 37 showed the result of optimal design under each setting. In each sub-plot, the red curve is the cost function, orange star is the pilot study design and purple star is the true optimal experimental design computed from simulated true data. There are another 10 dots, which indicated our predicted pilot study design from 10 pilot datasets simulated from the sample model. In all settings, the optimal design predictions were pretty accurate. With increase of pilot data sample size, the accuracy of optimal design prediction also improves.



(A)fc≥1.15; (B)fc≥1.20; (C)fc≥1.25; (D)fc≥1.30.

Figure 37: Optimal Design given pilot study and cost function

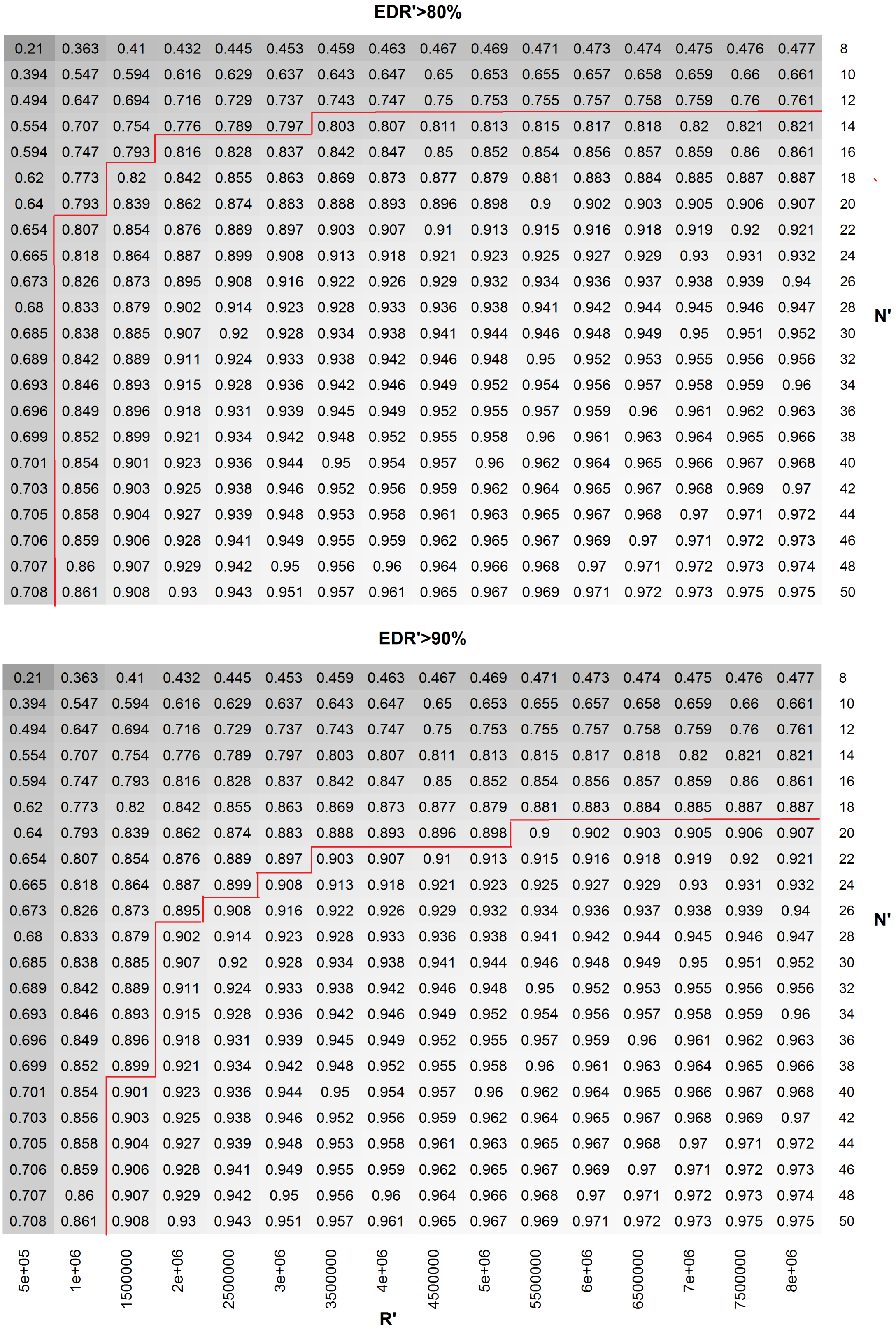
To illustrate the solution to Q2$\sim$Q3, now we focus on setting (δ=50,fc$\geq$1.20,N=4), where the EDR surface is: $\widehat{Pow}(N', R') = 1 - 32.588 \times N^{1.998} - 10^6 \times R^{-1.150}$. We now ask the following specific questions:

(a) What experimental design(N*,R*) are favored to reach EDR' (for example, 80%, 90%)? In Figure 38, we showed the predicted power under each experimental design. The red lines indicates boundary of desirable design with power greater than EDR'.

(b) What experimental design(N*,R*) are favored to reach EDR' (for example, 80%, 90%) with maximum sample size N_{max}? Assume N_{max} is 30, Figure 39 showed the resulting desirable designs:

Now assuming that we already have a RNA-Seq dataset with (N'=8,R'=6.5×10^6). The question is that if we have already achieve our desirable power EDR' (for example, 80%), and if we need to sequence extra samples to reach that?

To answer the question, we first computed the predicted EDR at N'=8, which is EDR'=34%. With fixed R'=6.5M, we then computed the minimum samples required to reach EDR=80%, which is 14. (Figure 40)



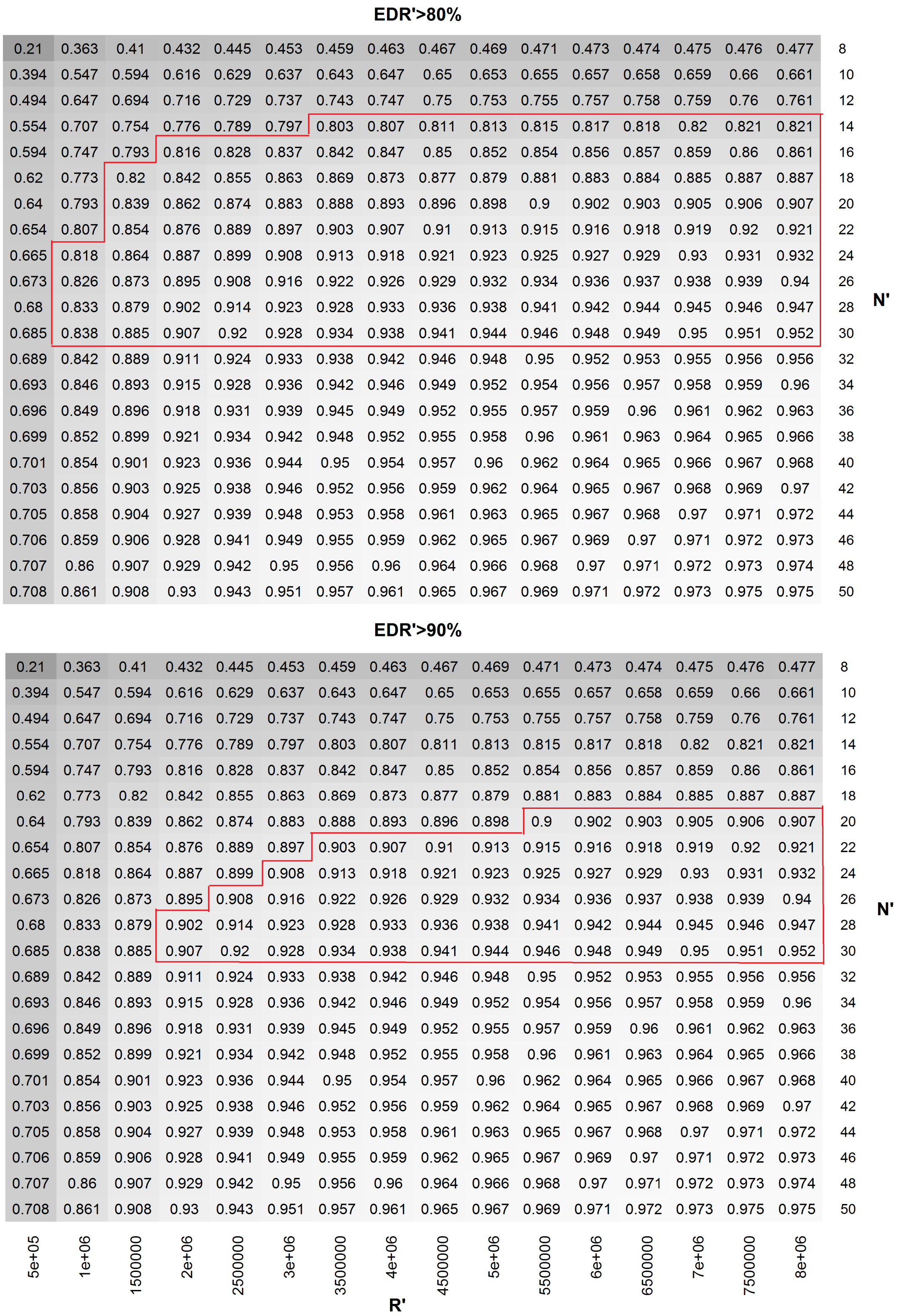
EDR'>80%

5e+05	1e+06	1500000	2e+06	2500000	3e+06	3500000	4e+06	4500000	5e+06	5500000	6e+06	6500000	7e+06	7500000	8e+06	N'
0.21	0.363	0.41	0.432	0.445	0.453	0.459	0.463	0.467	0.469	0.471	0.473	0.474	0.475	0.476	0.477	8
0.394	0.547	0.594	0.616	0.629	0.637	0.643	0.647	0.65	0.653	0.655	0.657	0.658	0.659	0.66	0.661	10
0.494	0.647	0.694	0.716	0.729	0.737	0.743	0.747	0.75	0.753	0.755	0.757	0.758	0.759	0.76	0.761	12
0.554	0.707	0.754	0.776	0.789	0.797	0.803	0.807	0.811	0.813	0.815	0.817	0.818	0.82	0.821	0.821	14
0.594	0.747	0.793	0.816	0.828	0.837	0.842	0.847	0.85	0.852	0.854	0.856	0.857	0.859	0.86	0.861	16
0.62	0.773	0.82	0.842	0.855	0.863	0.869	0.873	0.877	0.879	0.881	0.883	0.884	0.885	0.887	0.887	18
0.64	0.793	0.839	0.862	0.874	0.883	0.888	0.893	0.896	0.898	0.9	0.902	0.903	0.905	0.906	0.907	20
0.654	0.807	0.854	0.876	0.889	0.897	0.903	0.907	0.91	0.913	0.915	0.916	0.918	0.919	0.92	0.921	22
0.665	0.818	0.864	0.887	0.899	0.908	0.913	0.918	0.921	0.923	0.925	0.927	0.929	0.93	0.931	0.932	24
0.673	0.826	0.873	0.895	0.908	0.916	0.922	0.926	0.929	0.932	0.934	0.936	0.937	0.938	0.939	0.94	26
0.68	0.833	0.879	0.902	0.914	0.923	0.928	0.933	0.936	0.938	0.941	0.942	0.944	0.945	0.946	0.947	28
0.685	0.838	0.885	0.907	0.92	0.928	0.934	0.938	0.941	0.944	0.946	0.948	0.949	0.95	0.951	0.952	30
0.689	0.842	0.889	0.911	0.924	0.933	0.938	0.942	0.946	0.948	0.95	0.952	0.953	0.955	0.956	0.956	32
0.693	0.846	0.893	0.915	0.928	0.936	0.942	0.946	0.949	0.952	0.954	0.956	0.957	0.958	0.959	0.96	34
0.696	0.849	0.896	0.918	0.931	0.939	0.945	0.949	0.952	0.955	0.957	0.959	0.96	0.961	0.962	0.963	36
0.699	0.852	0.899	0.921	0.934	0.942	0.948	0.952	0.955	0.958	0.96	0.961	0.963	0.964	0.965	0.966	38
0.701	0.854	0.901	0.923	0.936	0.944	0.95	0.954	0.957	0.96	0.962	0.964	0.965	0.966	0.967	0.968	40
0.703	0.856	0.903	0.925	0.938	0.946	0.952	0.956	0.959	0.962	0.964	0.965	0.967	0.968	0.969	0.97	42
0.705	0.858	0.904	0.927	0.939	0.948	0.953	0.958	0.961	0.963	0.965	0.967	0.968	0.97	0.971	0.972	44
0.706	0.859	0.906	0.928	0.941	0.949	0.955	0.959	0.962	0.965	0.967	0.969	0.97	0.971	0.972	0.973	46
0.707	0.86	0.907	0.929	0.942	0.95	0.956	0.96	0.964	0.966	0.968	0.97	0.971	0.972	0.973	0.974	48
0.708	0.861	0.908	0.93	0.943	0.951	0.957	0.961	0.965	0.967	0.969	0.971	0.972	0.973	0.975	0.975	50

EDR'>90%

5e+05	1e+06	1500000	2e+06	2500000	3e+06	3500000	4e+06	4500000	5e+06	5500000	6e+06	6500000	7e+06	7500000	8e+06	N'
0.21	0.363	0.41	0.432	0.445	0.453	0.459	0.463	0.467	0.469	0.471	0.473	0.474	0.475	0.476	0.477	8
0.394	0.547	0.594	0.616	0.629	0.637	0.643	0.647	0.65	0.653	0.655	0.657	0.658	0.659	0.66	0.661	10
0.494	0.647	0.694	0.716	0.729	0.737	0.743	0.747	0.75	0.753	0.755	0.757	0.758	0.759	0.76	0.761	12
0.554	0.707	0.754	0.776	0.789	0.797	0.803	0.807	0.811	0.813	0.815	0.817	0.818	0.82	0.821	0.821	14
0.594	0.747	0.793	0.816	0.828	0.837	0.842	0.847	0.85	0.852	0.854	0.856	0.857	0.859	0.86	0.861	16
0.62	0.773	0.82	0.842	0.855	0.863	0.869	0.873	0.877	0.879	0.881	0.883	0.884	0.885	0.887	0.887	18
0.64	0.793	0.839	0.862	0.874	0.883	0.888	0.893	0.896	0.898	0.9	0.902	0.903	0.905	0.906	0.907	20
0.654	0.807	0.854	0.876	0.889	0.897	0.903	0.907	0.91	0.913	0.915	0.916	0.918	0.919	0.92	0.921	22
0.665	0.818	0.864	0.887	0.899	0.908	0.913	0.918	0.921	0.923	0.925	0.927	0.929	0.93	0.931	0.932	24
0.673	0.826	0.873	0.895	0.908	0.916	0.922	0.926	0.929	0.932	0.934	0.936	0.937	0.938	0.939	0.94	26
0.68	0.833	0.879	0.902	0.914	0.923	0.928	0.933	0.936	0.938	0.941	0.942	0.944	0.945	0.946	0.947	28
0.685	0.838	0.885	0.907	0.92	0.928	0.934	0.938	0.941	0.944	0.946	0.948	0.949	0.95	0.951	0.952	30
0.689	0.842	0.889	0.911	0.924	0.933	0.938	0.942	0.946	0.948	0.95	0.952	0.953	0.955	0.956	0.956	32
0.693	0.846	0.893	0.915	0.928	0.936	0.942	0.946	0.949	0.952	0.954	0.956	0.957	0.958	0.959	0.96	34
0.696	0.849	0.896	0.918	0.931	0.939	0.945	0.949	0.952	0.955	0.957	0.959	0.96	0.961	0.962	0.963	36
0.699	0.852	0.899	0.921	0.934	0.942	0.948	0.952	0.955	0.958	0.96	0.961	0.963	0.964	0.965	0.966	38
0.701	0.854	0.901	0.923	0.936	0.944	0.95	0.954	0.957	0.96	0.962	0.964	0.965	0.966	0.967	0.968	40
0.703	0.856	0.903	0.925	0.938	0.946	0.952	0.956	0.959	0.962	0.964	0.965	0.967	0.968	0.969	0.97	42
0.705	0.858	0.904	0.927	0.939	0.948	0.953	0.958	0.961	0.963	0.965	0.967	0.968	0.97	0.971	0.972	44
0.706	0.859	0.906	0.928	0.941	0.949	0.955	0.959	0.962	0.965	0.967	0.969	0.97	0.971	0.972	0.973	46
0.707	0.86	0.907	0.929	0.942	0.95	0.956	0.96	0.964	0.966	0.968	0.97	0.971	0.972	0.973	0.974	48
0.708	0.861	0.908	0.93	0.943	0.951	0.957	0.961	0.965	0.967	0.969	0.971	0.972	0.973	0.975	0.975	50

Figure 38: Desirable experiment design with constraint on minimum EDR



EDR'>80%

0.21	0.363	0.41	0.432	0.445	0.453	0.459	0.463	0.467	0.469	0.471	0.473	0.474	0.475	0.476	0.477	8
0.394	0.547	0.594	0.616	0.629	0.637	0.643	0.647	0.65	0.653	0.655	0.657	0.658	0.659	0.66	0.661	10
0.494	0.647	0.694	0.716	0.729	0.737	0.743	0.747	0.75	0.753	0.755	0.757	0.758	0.759	0.76	0.761	12
0.554	0.707	0.754	0.776	0.789	0.797	0.803	0.807	0.811	0.813	0.815	0.817	0.818	0.82	0.821	0.821	14
0.594	0.747	0.793	0.816	0.828	0.837	0.842	0.847	0.85	0.852	0.854	0.856	0.857	0.859	0.86	0.861	16
0.62	0.773	0.82	0.842	0.855	0.863	0.869	0.873	0.877	0.879	0.881	0.883	0.884	0.885	0.887	0.887	18
0.64	0.793	0.839	0.862	0.874	0.883	0.888	0.893	0.896	0.898	0.9	0.902	0.903	0.905	0.906	0.907	20
0.654	0.807	0.854	0.876	0.889	0.897	0.903	0.907	0.91	0.913	0.915	0.916	0.918	0.919	0.92	0.921	22
0.665	0.818	0.864	0.887	0.899	0.908	0.913	0.918	0.921	0.923	0.925	0.927	0.929	0.93	0.931	0.932	24
0.673	0.826	0.873	0.895	0.908	0.916	0.922	0.926	0.929	0.932	0.934	0.936	0.937	0.938	0.939	0.94	26
0.68	0.833	0.879	0.902	0.914	0.923	0.928	0.933	0.936	0.938	0.941	0.942	0.944	0.945	0.946	0.947	28
0.685	0.838	0.885	0.907	0.92	0.928	0.934	0.938	0.941	0.944	0.946	0.948	0.949	0.95	0.951	0.952	30
0.689	0.842	0.889	0.911	0.924	0.933	0.938	0.942	0.946	0.948	0.95	0.952	0.953	0.955	0.956	0.956	32
0.693	0.846	0.893	0.915	0.928	0.936	0.942	0.946	0.949	0.952	0.954	0.956	0.957	0.958	0.959	0.96	34
0.696	0.849	0.896	0.918	0.931	0.939	0.945	0.949	0.952	0.955	0.957	0.959	0.96	0.961	0.962	0.963	36
0.699	0.852	0.899	0.921	0.934	0.942	0.948	0.952	0.955	0.958	0.96	0.961	0.963	0.964	0.965	0.966	38
0.701	0.854	0.901	0.923	0.936	0.944	0.95	0.954	0.957	0.96	0.962	0.964	0.965	0.966	0.967	0.968	40
0.703	0.856	0.903	0.925	0.938	0.946	0.952	0.956	0.959	0.962	0.964	0.965	0.967	0.968	0.969	0.97	42
0.705	0.858	0.904	0.927	0.939	0.948	0.953	0.958	0.961	0.963	0.965	0.967	0.968	0.97	0.971	0.972	44
0.706	0.859	0.906	0.928	0.941	0.949	0.955	0.959	0.962	0.965	0.967	0.969	0.97	0.971	0.972	0.973	46
0.707	0.86	0.907	0.929	0.942	0.95	0.956	0.96	0.964	0.966	0.968	0.97	0.971	0.972	0.973	0.974	48
0.708	0.861	0.908	0.93	0.943	0.951	0.957	0.961	0.965	0.967	0.969	0.971	0.972	0.973	0.975	0.975	50

N'

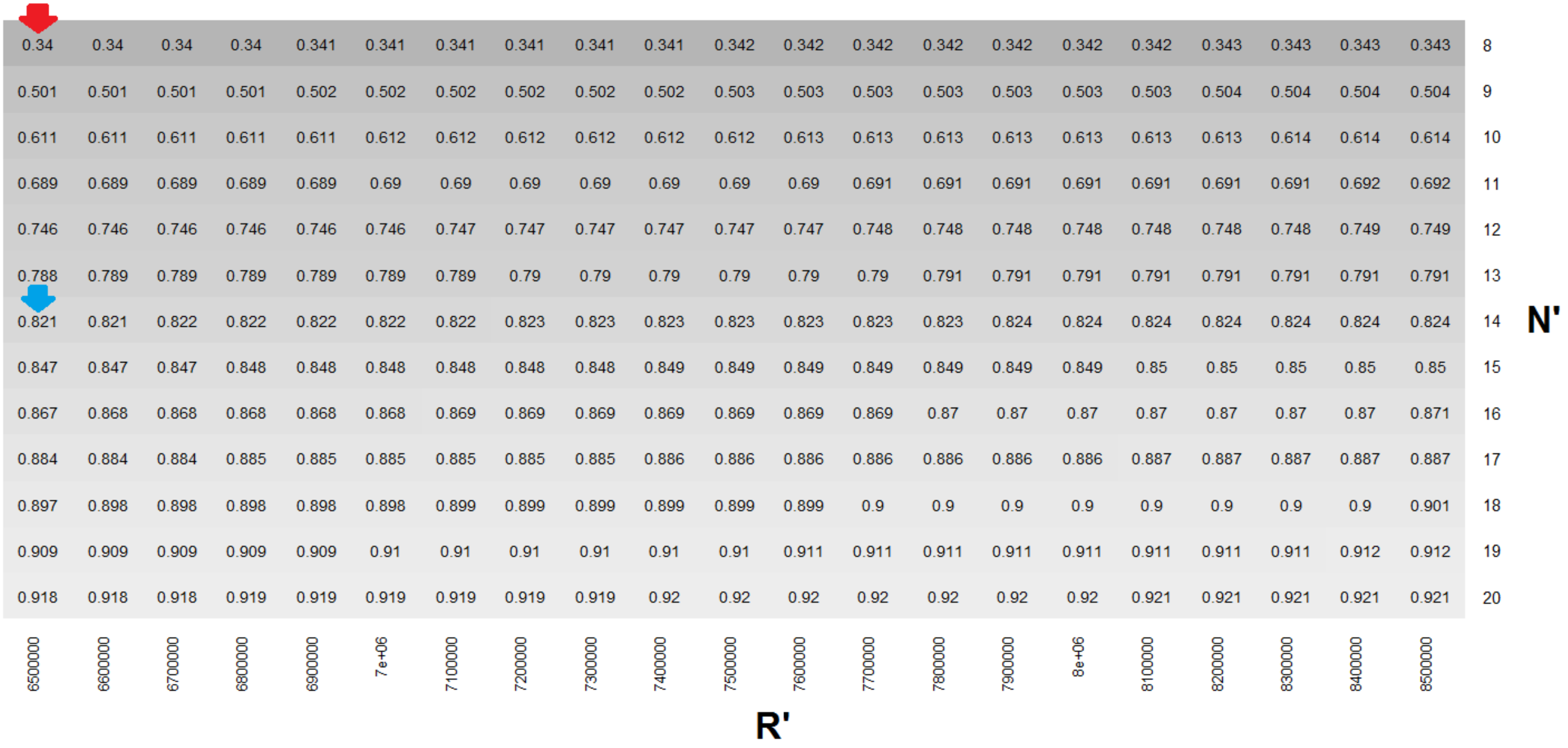
EDR'>90%

0.21	0.363	0.41	0.432	0.445	0.453	0.459	0.463	0.467	0.469	0.471	0.473	0.474	0.475	0.476	0.477	8
0.394	0.547	0.594	0.616	0.629	0.637	0.643	0.647	0.65	0.653	0.655	0.657	0.658	0.659	0.66	0.661	10
0.494	0.647	0.694	0.716	0.729	0.737	0.743	0.747	0.75	0.753	0.755	0.757	0.758	0.759	0.76	0.761	12
0.554	0.707	0.754	0.776	0.789	0.797	0.803	0.807	0.811	0.813	0.815	0.817	0.818	0.82	0.821	0.821	14
0.594	0.747	0.793	0.816	0.828	0.837	0.842	0.847	0.85	0.852	0.854	0.856	0.857	0.859	0.86	0.861	16
0.62	0.773	0.82	0.842	0.855	0.863	0.869	0.873	0.877	0.879	0.881	0.883	0.884	0.885	0.887	0.887	18
0.64	0.793	0.839	0.862	0.874	0.883	0.888	0.893	0.896	0.898	0.9	0.902	0.903	0.905	0.906	0.907	20
0.654	0.807	0.854	0.876	0.889	0.897	0.903	0.907	0.91	0.913	0.915	0.916	0.918	0.919	0.92	0.921	22
0.665	0.818	0.864	0.887	0.899	0.908	0.913	0.918	0.921	0.923	0.925	0.927	0.929	0.93	0.931	0.932	24
0.673	0.826	0.873	0.895	0.908	0.916	0.922	0.926	0.929	0.932	0.934	0.936	0.937	0.938	0.939	0.94	26
0.68	0.833	0.879	0.902	0.914	0.923	0.928	0.933	0.936	0.938	0.941	0.942	0.944	0.945	0.946	0.947	28
0.685	0.838	0.885	0.907	0.92	0.928	0.934	0.938	0.941	0.944	0.946	0.948	0.949	0.95	0.951	0.952	30
0.689	0.842	0.889	0.911	0.924	0.933	0.938	0.942	0.946	0.948	0.95	0.952	0.953	0.955	0.956	0.956	32
0.693	0.846	0.893	0.915	0.928	0.936	0.942	0.946	0.949	0.952	0.954	0.956	0.957	0.958	0.959	0.96	34
0.696	0.849	0.896	0.918	0.931	0.939	0.945	0.949	0.952	0.955	0.957	0.959	0.96	0.961	0.962	0.963	36
0.699	0.852	0.899	0.921	0.934	0.942	0.948	0.952	0.955	0.958	0.96	0.961	0.963	0.964	0.965	0.966	38
0.701	0.854	0.901	0.923	0.936	0.944	0.95	0.954	0.957	0.96	0.962	0.964	0.965	0.966	0.967	0.968	40
0.703	0.856	0.903	0.925	0.938	0.946	0.952	0.956	0.959	0.962	0.964	0.965	0.967	0.968	0.969	0.97	42
0.705	0.858	0.904	0.927	0.939	0.948	0.953	0.958	0.961	0.963	0.965	0.967	0.968	0.97	0.971	0.972	44
0.706	0.859	0.906	0.928	0.941	0.949	0.955	0.959	0.962	0.965	0.967	0.969	0.97	0.971	0.972	0.973	46
0.707	0.86	0.907	0.929	0.942	0.95	0.956	0.96	0.964	0.966	0.968	0.97	0.971	0.972	0.973	0.974	48
0.708	0.861	0.908	0.93	0.943	0.951	0.957	0.961	0.965	0.967	0.969	0.971	0.972	0.973	0.975	0.975	50
5e+05	1e+06	1500000	2e+06	2500000	3e+06	3500000	4e+06	4500000	5e+06	5500000	6e+06	6500000	7e+06	7500000	8e+06	

N'

R'

Figure 39: Desirable experiment design with constraint on minimum EDR and maximum N



N'	6500000	6600000	6700000	6800000	6900000	7e+06	7100000	7200000	7300000	7400000	7500000	7600000	7700000	7800000	7900000	8e+06	8100000	8200000	8300000	8400000	8500000
8	0.34	0.34	0.34	0.34	0.341	0.341	0.341	0.341	0.341	0.341	0.342	0.342	0.342	0.342	0.342	0.342	0.342	0.343	0.343	0.343	0.343
9	0.501	0.501	0.501	0.501	0.502	0.502	0.502	0.502	0.502	0.502	0.503	0.503	0.503	0.503	0.503	0.503	0.503	0.504	0.504	0.504	0.504
10	0.611	0.611	0.611	0.611	0.611	0.612	0.612	0.612	0.612	0.612	0.612	0.613	0.613	0.613	0.613	0.613	0.613	0.613	0.614	0.614	0.614
11	0.689	0.689	0.689	0.689	0.689	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.691	0.691	0.691	0.691	0.691	0.691	0.691	0.692	0.692
12	0.746	0.746	0.746	0.746	0.746	0.746	0.747	0.747	0.747	0.747	0.747	0.747	0.748	0.748	0.748	0.748	0.748	0.748	0.748	0.749	0.749
13	0.788	0.789	0.789	0.789	0.789	0.789	0.789	0.79	0.79	0.79	0.79	0.79	0.79	0.791	0.791	0.791	0.791	0.791	0.791	0.791	0.791
14	0.821	0.821	0.822	0.822	0.822	0.822	0.822	0.823	0.823	0.823	0.823	0.823	0.823	0.823	0.824	0.824	0.824	0.824	0.824	0.824	0.824
15	0.847	0.847	0.847	0.848	0.848	0.848	0.848	0.848	0.848	0.849	0.849	0.849	0.849	0.849	0.849	0.849	0.85	0.85	0.85	0.85	0.85
16	0.867	0.868	0.868	0.868	0.868	0.868	0.869	0.869	0.869	0.869	0.869	0.869	0.869	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.871
17	0.884	0.884	0.884	0.885	0.885	0.885	0.885	0.885	0.885	0.886	0.886	0.886	0.886	0.886	0.886	0.886	0.887	0.887	0.887	0.887	0.887
18	0.897	0.898	0.898	0.898	0.898	0.898	0.899	0.899	0.899	0.899	0.899	0.899	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.901
19	0.909	0.909	0.909	0.909	0.909	0.91	0.91	0.91	0.91	0.91	0.91	0.911	0.911	0.911	0.911	0.911	0.911	0.911	0.911	0.912	0.912
20	0.918	0.918	0.918	0.919	0.919	0.919	0.919	0.919	0.919	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.921	0.921	0.921	0.921	0.921

Figure 40: Desirable experiment design for a given RNA-Seq experiment

5.0 DISCUSSION AND FUTURE WORK

5.1 DISCUSSION

In the existing and our methods, the definition of power actually varies. In SeqDEsign, we focused on genome-wide power, which is similar to the concept of sensitivity (recall rate or true positive rate(TPR)) in ROC analysis. We conducted more than 10,000 hypothesis tests simultaneously and estimate genome-wide power as the proportion of true positive among the true DE genes. It's different from the concept of power as in RNASeqPower, where only one test is conducted conceptually for the computation of power. To our knowledge, genome-wide power definition is more appropriate for the case of genomic data since the heterogeneity across genes could be maintained. Consequently, it's somewhat unfair to compare these different definitions of "power" together. But it's already the best we can do to conduct the comparison.

In our method, we defined R as the total reads from a pilot study. To be more specific, this should be the total reads that could be mapped to gene regions. Consequently, when considering the sequencing cost, one should divide B by the average alignment rate to be realistic.

As is briefly mentioned in chapter 3, when we have a relatively some parameter space of (N',R') for optimal experiment design, we can compute EDR of all possible design($EDR_{(N,R)}(N', R')$) under the constraints. In this scenario, we do not have to fit two-way inverse power law model. In cases when we have a much larger parameter space to search through, for example, N_{max} is 100 or there is very weak signal for DE gene detection, we should instead estimated EDR at selected (N',R') to reduce computation time and apply the inverse power law model.

Our method can predict EDR hyperplane when a pilot data is available. If there's no pilot data, one can seek previous public datasets with similar platforms. Another possible solution is to design RNA-Seq experiment in multi-phases. We can therefore estimated EDR adaptively after the completion of each phase with higher accuracy.

5.2 FUTURE WORK

Here we discuss future works and directions for SeqDEsign:

FW1: Relax common dispersion assumption: Currently, our method is based on negative binomial model with common dispersion across genes. To accommodate all possible data structure, the next step will be to evaluate our method in data with tag-wise dispersion and extend our method if necessary, for example, we can apply empirical Bayes methods. We will provided test to determine whether common dispersion model is correct.

FW2: Further investigation in semi-parametric method: In the real data example of chapter 2(Figure 21), we observed that when differential expression has very weak signal, p-value distribution of data with small pilot sample size(N) will not satisfy our mixture model assumption well. In this case, the performance of our method is not good due to model mis-specification. One solution is to apply non-parametric method. We will have further investigation in non-parametric approach under this scenario in the future.

FW3: Design of pilot study: So far, we have illustrated the impact of sample size and read depth to power by simulation studies and real data examples. The influence of different pilot data to the prediction of EDR surface, however, is not investigated sufficiently. For example, if we specify a cost upper bound of $24,000 for pilot study. The following three setting of pilot study design have same cost: (a) N=4, R=6M; (b) N=8, R=3M; (c) N=16, R=1.5M. By the three different pilot study, we will have different EDR surface predictions. In future study, we will provide suggestions for the design of pilot study.

FW4: More real data applications: In the future, we will apply our method to TCGA RNA-Seq data `https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp`. It is a rich database with more than hundreds of tumor samples for each cancer. We can start with various smaller pilot sample size, generate the predicted EDR hyperplane and compared with the "true" EDR hyperplane generated by using data with larger sample size and evaluate their performance.

FW5: Software preparation: We will prepare an R package "SeqDEsign" and a web interactive tool based on java script, which could facilitate the application of our tool in real world RNA-Seq experiment.

6.0 CONCLUSION

Power analysis is of great importance in study design phase. Especially, with accruing popularity of next generation sequencing technology, there's an increasing need for statistical solid and easy-to-implement power calculation method. Some existing power calculation tools for microarray and NGS ignore genome-wide false discovery rate control and only perform per-gene power calculation. Some others have utilized naïve modelling without adequately borrowing information from pilot data.

In this thesis, we have proposed a new approach: SeqDEsign to predict genome-wide power based on a RNA-Seq pilot study. Simulation studies and real data application showed the superiority of our methods. Our approach provides several unique advantages over all existing methods: (1) higher statistical reliability: our model is based on negative binomial assumption of count data instead of poisson or gaussian assumption; (2) genome-wide power(EDR): we define genome-wide power(EDR) which considers the DE gene detection sensitivity in the realm of whole genome, instead of single gene level; (3) better accuracy: simulation and real data analysis reveals the high accuracy of our methods; (4) optimal experiment design: we consider the influence of both sample size and read depth on genome-wide power. Consequently, given the cost constraints, one can predict the optimal experiment design (N*,R*) after EDR surface was constructed; (5) easy to implement: our method tends to be model-based compared with existing methods. We don't need to specify fold change for DE gene detection or number of true rejections.

To our knowledge, SeqDEsign is the first statistical tool that address the power calculation and experimental design for RNA-Seq data with proper model assumptions. Considering the superior performance and capability in answering various research questions, we believe it will provide researchers valuable suggestions in the experiment design of RNA-Seq data in the future.

BIBLIOGRAPHY

David B. Allison, Gary L. Gadbury, Moonseong Heo, Jose R. Fernandez JR, Cheol-Koo Lee, Tomas A. Prolla, and Richard Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, 39:1–20, 2002.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(R106), 2010.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach for multiple testing. *Journal of the Royal Statistical Society. Series B.*, 57(1):289–300, 1995.

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), 2009.

James H Bullard, Elizabeth Purdom, Kasper D, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11(94), 2010.

J. Lawrence Burg, Christopher M. Grover, Philippe Pouletty, and John C. Boothroyd. Direct and sensitive detection of a pathogenic protozoan, toxoplasma gondii, by polymerase chain reaction. *Journal of Clinical Microbiology*, 27(8):1787–1792, 1989.

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*, 30(11):1–2, 2013.

Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P. Speed. Statistical tests for differential expression in cdna microarray experiments. *Genome Biology*, 4(4): 210, 2003.

Bradley Efron. Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377, 2007.

Zhide Fang and Xiangqin Cui. Design and validation issues in rna-seq experiments. *Briefings in bioinformatics*, 12:280–287, 2011.

Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. Predicting sample size required for classification performance. *BMC Medical Informatics & Decision Making*, 12(8), 2012.

Gary L Gadbury, Grier P Page, Jode Edwards, Tsuyoshi Kayo Wisconsin, 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*, 13:325–338, 2004.

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.

Kangxia Gu, Hon Keung Tony Ng, Man Lai Tang, and William R. Schucany. Testing the ratio of two poisson rates. *Biometrical Journal*, 50:283–298, 2008.

Thomas J Hardcastle and Krystyna A Kelly. bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11:422, 2010.

Steven N. Hart, Terry M. Therneau, Yuji Zhang, Gregory A. Poland, and Jean-Pierre Kocher. Calculating sample size estimates for rna sequencing data. *Journal of computational biology*, 20(12):970–8, 2013.

M. Kathleen Keer and Gary A. Churchill. Statistical design and the analysis of gene expression microarray data. *Genetics Research*, 77:123–128, 2001.

M. Kathleen Kerr and Gary A. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201, 2001.

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*, 67:555–572, 2005.

Mei-Ling Ting Lee and G. A. Whitemore. Power and sample size for dna microarray studies. *Statistics in Medicine*, 21:3543–3570, 2002.

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 Computational Biology and Drug Design*, 6(4):358–75, 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*, 14:357, 2013b.

Jun Li, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13: 523–38, 2012.

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), 2013c.

Yuwen Liu, Jie Zhou, and Kevin P. White. Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(11):1–4, 2013.

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.

Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acid Research*, 42(9):1–10, 2012.

Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlen, 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, 2011.

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, 2009.

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(84), 2006.

Elena Perelman, Alexander Ploner, Stefano Calza, and Yudi Pawitan. Detecting differential expression in microarray data: comparison of optimal procedures. *BMC Bioinformatics*, 8(28), 2007.

Stan Pounds and Stephan W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.

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.

Mark Reimers. Making informed choices about microarray data analysis. *PLOS Comput Biology*, 6:e1000786, 2010.

Davide Risso, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. Gc-content normalization for rna-seq data. *BMC Bioinformatics*, 12(1):480, 2011.

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.

Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–40, 2010.

Charles M. Roth. Quantifying gene expression. *Current issues molecular biology*, 4(93): 93–100, 2002.

Frederick Sanger and Alan Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, 94(3):441–8, 1975.

Thomas D. Schmittgen, Eun Joo Lee, and Jinmai Jiang. High-throughput real-time pcr. *Methods in Molecular Biology*, 429:89–98, 2008.

Almut Schulze and Julian Downward. Navigating gene expression using microarrays a technology review. *Nature Cell Biology*, 3:E190–E195, 2001.

Jay Shendure. The beginning of the end for microarrays. *Nature Methods*, 5:585–587, 2008.

B. W. Silverman. *Density estimation for statistics and data analysis.* 1986.

Richard M. Simon and Kevin Dobbin. Experimental design of dna microarray experiments. *Bio Techniques*, 34:S16–S21, 2003.

Donna K. Slonim and Itai Yanai. Getting started in gene expression microarray analysis. *PLOS Comput Biol*, 5:e1000543, 2009.

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), 2004.

John D. Storey. A direct approach to false discovery rates. *Journal of Royal Statistical Society Series B*, 64(3):479–498, 2002.

John D. Storey and Robert Tibshirani. Estimating false discovery rates under dependence, with applications to dna microarrays. *Technical report*, 2001.

John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16): 9440–9445, 2003.

Peter A. C. ’t Hoen, Yavuz Ariyurek, Helene H. Thygesen, Erno Vreugdenhil, Rolf H. A. M. 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 Acid Research*, 36(21):e141, 2008.

Sonia Tarazona, Fernando Garca-Alcalde, Joaqun Dopazo, Alberto Ferrer, , and Ana Conesa. Differential expression in rna-seq: A matter of depth. *Genome Research*, 21:2213–2223, 2011.

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, 2009a.

Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, 10(1):57–63, 2009b.

Haiyuan Zhu and Hassan Lakkis. Sample size calculation for comparing two negative binomial rates. *Statistics in Medicine*, 2013.