Power analysis for RNA sequencing and mass spectrometry-based proteomics data

Master of Science Thesis
University of Turku
Department of Future Technologies
Master's Degree Programme in Bioinformatics
2018
Xu Qiao

Supervisors:
Tomi Suomi
Laura L. Elo
Martti Tolvanen

UNIVERSITY OF TURKU
Department of Future Technologies

XU QIAO: Power analysis for RNA sequencing and mass spectrometry-based proteomics data

Master of Science Thesis, 56 p.
Master's Degree Programme in Bioinformatics
November 2018

RNA-sequencing and mass spectrometry technologies have facilitated the differential expression discoveries in transcriptome and proteome studies. However, the determination of sample size to achieve adequate statistical power has been a major challenge in experimental design. The objective of this study is to develop a power analysis tool applicable to both RNA-seq and MS-based proteomics data. The methods proposed in this study are capable of both prospective and retrospective power analyses. In terms of the performance, the benchmarking results indicated that the proposed methods can give distinct power estimates for both differentially and equivalently expressed genes or proteins without prior differential expression analysis and other assumptions, such as, expected fraction of differentially expressed features, minimal fold changes and expected mean expressions. Using the proposed methods, not only can researchers evaluate the reliability of their acquired significant results, but also estimate the sufficient sample size for a desired power. The proposed methods in this study were implemented as an R package, which can be freely accessed from Bioconductor project at http://bioconductor.org/packages/PowerExplorer/.

Keywords: Power analysis, Sample size, Differential expression, RNA sequencing, Microarray, Mass spectrometry, Transcriptomics, Proteomics

Acknowledgement

Foremost, I would like to express my sincere gratitude to my supervisors MSc. Tomi Suomi, Dr. Mikko S. Venäläinen and Dr. Laura L. Elo for their valuable remarks and patient engagements during my long learning process of this Master's thesis project.

I would also like to show my gratitude to my loving friends in Finland and the fellow colleagues in Turku Centre for Biotechnology, who have offered me helpful suggestions regarding my studies and new career in Finland.

I am especially grateful to my dearest friends Cindy McElvaney, William Eccleshall and Nataliia Petruk for their emotional support and encouragement during my worst time when my grandfather, the mentor of my life, recently got diagnosed with lung cancer. Additionally, I would like to thank my friend Julia Walter for her constructive criticism on this thesis.

Most importantly, I wish to thank my family, whose unfailing love and faith are always with me for whatever I pursue.

Thank you!

Xu Qiao

Turku, Finland, October, 2018

Contents

List of Figures

List of Tables

List of Abbreviations

cDNA Complementary DNA

ChIP-Seq Chromatin Immunoprecipitation Sequencing

CPM Counts Per Million

DE Differentially Expressed

DEA Differential Expression Analysis

EE Equivalently Expressed

ERCC External RNA Control Consortium

FDR False Discovery Rate

FN False Negative

FP False Positive

GLM Generalized Linear Model

HTS High-throughput Sequencing

IRLS Iteratively Reweighted Least Squares

LFC Log Fold Change

LFQ Label-free Quantification

ML Maximum Likelihood

mRNA Messenger RNA

MS Mass Spectrometry

NB Negative Binomial

PCR Polymerase Chain Reaction

PPA Prospective Power Analysis

RNA-seq RNA Sequencing

ROTS Reproducibility Optimized Test Statistic

RPA Retrospective Power Analysis

RPKM Reads Per Kilobase of exon per Million reads mapped

rRNA Ribosomal RNA

SEQC RNA Sequencing Quality Control

TMM Trimmed Mean of M values

TN True Negative

TP True Positive

TPM Transcripts Per Million

VSN Variance Stabilization Normalization

VST Variance Stabilization Transformation

Chapter 1

Introduction

A recent literature [1] assessed thousands of recently published statistical records in the fields of cognitive neuroscience and psychology. They reported that most of the publications had low replication and power, which resulted in inflated false positive discoveries. Additionally, many published significant results were found to be "only the occasional large deviations from real effect sizes". They suggested more attention was needed for the numerous reproduction failures in psychology and cognitive neuroscience research.

Furthermore, an on-line survey was conducted to assess the public opinion regarding the reproducibility of academic research [2]. The given responses from 1,536 researchers have drawn attention of the research community. More than 70% of the researchers indicated that they had attempted and failed to reproduce the published work of other research (**Figure 1.1**). A poorly designed experiment, such as one with inadequate sample size, can lead to erroneous results [3]. Due to insufficient knowledge in statistics, researchers in many fields have been overly trusting P-value as a reliable reference to significant discoveries.

A simulation-based study indicated that repeated experiments with small effect size substantially showed variation in P-values [4]. Uniformly distributed P-values were expected in reproduced experiments. However, in their study, the stability of P-value was not yet significantly improved by moderately increasing the sample size. Only when the

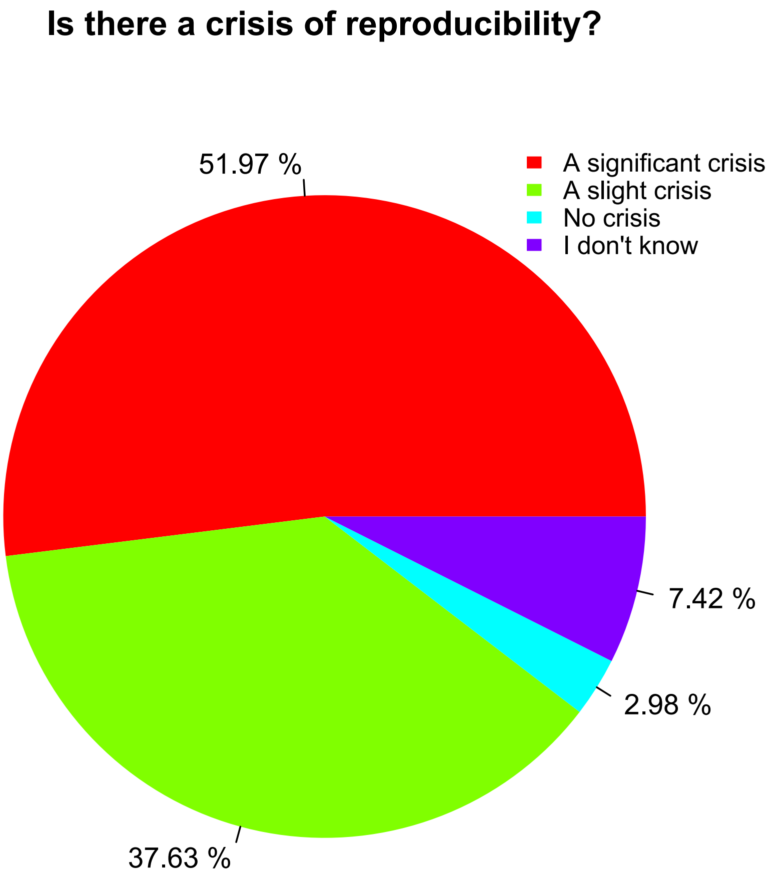


Figure 1.1: Summary of responses from the survey. More than 80% of the researchers think there is a crisis of reproducibility.

statistical power reached 90% did the P-values show much more stability among all simulations. The demonstrated results indicated that a P-value can be credited as the index of significance only when it is complemented by sufficiently high statistical power. Most importantly, power calculation has currently been required as a compulsory component of a research proposal to many funding applications.

A review paper [5] has categorized a wide range of power analysis methods for various omics studies, such as DNA sequencing, RNA Sequencing (RNA-seq), microbiome sequencing and chromatin immunoprecipitation sequencing (ChIP-Seq). The diversity of technologies results in different statistical assumptions and definitions of statistical power, which brings great difficulties in downstream statistical analysis due to the high dependence of suitable model selections. Most of the existing power estimation tools depend on the number of differentially expressed (DE) features detected by some differential expression analysis (DEA) methods, such as edgeR [6] and DESeq [7]. However, the differential expressions occurring in an experiment are usually unknown. Furthermore, few power analysis methods are found to be directly available for mass spectrometry (MS)-based proteomics data. Instead of depending on DEA methods, the proposed methods estimate the power at single gene or protein level using parametric resamplings from the estimated probability distribution of each gene or protein. However, this thesis only discusses about power analysis methods that are currently available for both RNA-seq and MS-based quantitative proteomics experiments. For more simplicity, the terms "genes" or "proteins" will be referred to as features.

Chapter 2

Background

This chapter introduces some essential concepts leading to a better understanding of this study. It includes the brief introductions to a few contemporary transcriptomics and proteomics technologies, as well as the concepts of hypothesis testing and power calculation.

2.1 Transcriptomic Analysis

A transcriptome consists of the complete set of expressed messenger RNA (mRNA) molecules in a cell or a population of cells. Because of its dynamic states, a studied transcriptome also includes the quantity or concentration of each detected transcript. One of the key aims of transcriptomics technologies is to quantify and compare the expression levels of each transcript between different tissues, time points or physiological conditions. The quantification process discovers gene expression patterns potentially reasoning for the causes of the biological changes, such as the pathological mechanism of a disease. Transcriptomics analysis may help with understanding the mechanisms behind biological changes by discovering the DE genes, but the proportion of available biomolecules in samples is usually extremely low. Thus, methods based on signal amplification, such as using fluorescence labeling, are developed to convert the nanometer-scale information into discrete optical measurements [8].

Most transcriptomics technologies generally require the process of RNA isolation which uses various RNA extraction methods aiming to achieve similar goals [9]: cellular disruption, total inhibition of RNase activity, separation of RNA from other biomolecules (mainly protein and DNA) and concentration of RNA. However, most of the isolated RNA are ribosomal RNA (rRNA) whereas only <5% are mRNA [10], hence mRNA enrichment is necessarily needed to amplify the signals. Widely used mRNA enrichment methods are generally based on rRNA capturing using sequence-specific probes, polyadenylation of mRNA or degradation of processed RNA [11].

Transcriptomics technologies are currently based on two major approaches: hybridization-based and sequence-based. Hybridization-based technologies quantify a set of interested transcripts for a specific experiment, the main contemporary hybridization-based technology is microarrays. Sequence-based technologies are currently dominated by RNA-seq, which detects all transcripts from input RNA samples using high-throughput sequencing (HTS) technologies.

2.1.1 Microarray

Microarrays are complementary DNA (cDNA) chips designed to measure the abundance of a predetermined list of transcripts, each chip has a solid surface with arrays consisting of short nucleotide oligomers, known as probes, targeting specific transcripts [12, 13]. After the hybridization of the fluorescence-labeled transcripts to the probes, the detection of the fluorescence intensities deduces the abundance of the targeted transcripts. However, microarray technology has become less popular due to its limitations: strong dependence on prior knowledge of genomes of interest; difficulty in analysis of closely related sequences owing to cross-hybridization; complexity of normalizing and comparing expression levels across experiments owing to the analog nature of fluorescence-based detection [14, 15].

2.1.2 RNA Sequencing

Microarray technologies usually depend on an annotated genomic sequence to generate a limited number of probes. In contrast, not only can RNA-seq determine most of the transcripts in an RNA extract by mapping the acquired sequences to a reference genome, but it also can assemble the sequences without a reference genome, which is often called *de novo* sequence assembly [16]. In addition to the quantification of the transcripts present in the RNA extract, it also directly determines the identity of the transcripts and connectivity between transcripts [15, 17].

RNA-seq utilizes the recently developed HTS technologies, an acquired population of isolated RNA are fragmented and reverse-transcribed to a library of cDNA fragments with length typically ranging from 30 to 10,000 base pairs. Before being delivered for actual sequencing, the whole library of created cDNA copies is amplified by polymerase chain reaction (PCR) method, which aims to generate millions of copies of the cDNA fragments in order to amplify the signals, since some RNA with small input amount may be undetectable without amplification [18]. After PCR amplification, the fragments are sequenced in single or both directions, i.e., single-end or pair-end sequencing. Using alignment algorithms, the fragment sequences are aligned based on a reference genome, depending on the sample species. The number of aligned fragments gives the read count of the corresponding transcript. The read counts represent the relative quantity of the mapped transcripts. Eventually, a long list of identified transcripts and the corresponding read counts are produced. Based on the acquired RNA-seq count data, various types of statistical analysis can be performed, for instance, quantification of transcriptomes [19], gene regulatory network analysis based on RNA-seq time-series data [20] and identification of DE genes between treatment groups [21].

Apart from mRNA, RNA-seq can also process other RNA populations including total RNA and non-coding RNA, such as rRNA, transfer RNA and micro RNA [22]. Recently, RNA-seq has also evolved with more capabilities, such as sequencing transcripts isolated from each single cell, which has led to an increasing understanding of cellular structures [23–25]. However, this study only discusses about the gene expression analysis of RNA-seq experiments for bulk populations. The expression data contain the gene read counts corresponding to each replicate of samples in two or more groups.

2.2 Proteomic Analysis

2.2.1 Mass Spectrometry-based Quantifications

Proteomics technologies have evolved rapidly from identifying the presence of a few proteins to quantifying a large amount of proteins. Quantitative proteomics technologies have enhanced our comprehension of protein expression and the underlying changes between organism samples collected from various conditions. MS-based proteomics approaches are mainly categorized as labeling-based and label-free quantification [26].

Labeling-based strategies are based on stable isotope dilution which assumes that a peptide labeled with stable isotopes should have the same physiochemical properties and different peptides can be identified with the unique isotope labels. Hence, a mass spectrometer can distinguish between the labeled and unlabeled peptides. Peptides can be quantified by comparing the respective signal intensities between the labeled and unlabeled editions of peptides. Used labels, which usually are heavy/light isotope pairs of the same element, can be introduced into the proteins or peptides using various labeling methods. As a result, the proteins or peptides will have either a heavy mass label or a light label. Over the past decade, a wide range of labeling methods have been introduced including metabolic labeling, isobaric mass tagging and isotope-coded reagents tagging, etc [27, 28].

Label-free quantification, mainly based on ion intensity and spectral counting, provides faster and cheaper strategies to determine the relative abundance of proteins from unlabeled peptide mixtures. Approaches based on ion intensity estimate the peptide abundance by the measurements and comparisons of chromatographic peak areas. Whereas spectral counting approaches determine peptide identities and protein abundance based on the number of the acquired tandem mass spectra and theoretical peptide spectra from a protein database [26, 27]. Similar to RNA-seq gene read count data, the MS-based proteomics data contain the protein abundance levels corresponding to each replicate in sample groups.

2.3 Power Analysis

2.3.1 Hypothesis Testing

Typically, a hypothesis testing involves a pair of relevant null and alternative hypotheses. Null hypothesis, usually denoted by H_0, is the hypothesis to be tested by a hypothesis testing model. Commonly, two groups of instances are compared, a null hypothesis often gives a statement that no associations exist, usually in terms of mean values, between the two groups. In contrast to null hypothesis, alternative hypothesis, denoted by H_1, is the rival statement. When a null hypothesis is rejected, i.e., the data cannot reinforce the statement of null hypothesis, the corresponding alternative hypothesis, as a result, is proven and accepted. A hypothesis testing model usually aims to seek the evidence that leads to the rejection of a null hypothesis, which results in a so-called positive detection [29]. However, a hypothesis testing model cannot give absolutely correct decisions. When testing a null hypothesis, assuming an underlying fact can be either true H_0 or false H_0, there are four possible outcomes (**Table 2.1**) respectively defined as

1. True negative (TN): H_0 is accepted when H_0 is true.
2. False negative (FN): H_0 is accepted when H_0 is false.
3. False positive (FP): H_0 is rejected when H_0 is true.
4. True positive (TP): H_0 is rejected when H_0 is false.

Among the aforementioned four outcomes, there exist two types of hypothesis testing errors:

1. Type I Error (denoted by α), equivalent to FP, the probability of which is often referred to as the significant level (α) of a test outcome.

2. Type II Error (denoted by β), equivalent to FN, the probability of which is usually used to determine the statistical power, equivalent to $1 - \beta$, of a test outcome.

Table 2.1: Possible outcomes of hypothesis testing

	H_0 is true	H_0 is false
Accept H_0	TN	FN
Reject H_0	FP	TP

Hypothesis testing measures the strength of the evidence against the null hypothesis, as well as the effect size between the samples. The degree of deviation toward null hypothesis can be reflected as the resulting statistics from a test model. In this study, the determination of true positive and power calculation is based on the degree of deviation from the null hypothesis resulting from various hypothesis testing models.

2.3.2 Statistical Power

Statistical power is the probability that a hypothesis testing model successfully rejects a null hypothesis when the null hypothesis is actually false. In other words, it is equivalent to the probability of obtaining a statistically significant result, that is, a TP, as defined in **Section 2.3.1**. Hence, a statistical power value is complementary to the Type II Error rate (β). An experiment with high statistical power ensures a hypothesis testing model making correct decisions, either significant or insignificant results, with high reproducibility.

For an RNA-seq or MS-based proteomics experiment, researchers often focus on seeking the DE features buried in an enormous amount of other features. Before performing hypothesis testing, an alternative hypothesis needs to be defined. It describes the distribution corresponding to the occasion that has enough effect or difference to reject a null hypothesis, which supports the assumption that no effect or difference is present between the studied populations [30].

Since the significant features of an experiment cannot be identified by simple observations, it often requires differential expression analysis to determine a true significance. However, when an experiment has low statistical power, most of the significant detections by hypothesis testing models may have high false discovery rates (FDRs) [4, 30]. The most common cause of low power is the lack of sample replicates. With a small number of sample replicates, each observation contains an inadequate number of data points that can credibly describe the true distribution. Hence, based on data with ambiguous specificity of the sourced distribution, the hypothesis testing model may give inaccurate reports of significance and insignificance.

Chapter 3

Materials and Methods

This chapter will introduce the implemented probability distributions, hypothesis testing models and the key components of the power analysis method in this study.

3.1 Data Modeling

Before performing statistical analysis, it is essential to determine the probability distribution that well describes the acquired data. In this study, Gaussian distribution was used to model log-transformed RNA-seq or proteomics data, and a negative binomial distribution model was implemented for the raw RNA-seq read counts.

In this study, for MS-based protein data, the protein abundance levels were transformed into log scale and modeled following normal distribution

$$\mathbf{X_{ijn} \sim N(\mu_{ij}, \sigma^2_{ij})}$$

where i is the index of proteins, j is the index of experimental groups and n is the sample id. In addition, X_{ijn} contains the abundance levels, μ_{ij} describes the average abundance level of protein i in group j and σ^2_{ij} captures the errors resulting from biological and technical variations. In addition, for some cases in which RNA-seq count data were already transformed into log scale, the count data were also modeled as normal distribution.

RNA-seq measures the expression based on the number of fragments mapped to the corresponding transcripts in a reference genome, which results in discrete counts. Count data are usually modeled using Poisson distribution which assumes the equivalence between mean and variance. However, RNA-seq often has a much wider range of measurements. As a result, the variance is usually larger than the mean, which is often referred to as overdispersion [31, 32]. The Poisson model, however, cannot account for such information. To take into account the additional information, negative binomial (NB) distribution is commonly used to capture the relationship between mean and variance [33, 34].

In this study, RNA-seq reads were modeled using the NB distribution

$$\mathbf{X_{ijn} \sim NB(\mu_{ij}, \phi_{ij})}$$

where i is the index of genes, and j is the index for experimental groups and n is the index for sample replicates within each group for gene i. Additionally, the μ_{ij} describes the average read counts of gene i in group j, and ϕ_{ij} captures the overdispersion due to biological and technical variations.

3.2 Statistical Models

3.2.1 Maximum Likelihood Methods

The generalized linear model (GLM) method is one of the most popular approaches for seeking an approximate distribution that can describe a set of acquired data (Refer to Chapter 8 in book [35] for detailed concepts). Maximum likelihood (ML) estimation requires a pre-determined assumption that a random variable X follows a selected distribution $p(x \mid \theta)$. A vector θ contains k parameters $(\theta_1, ..., \theta_k)$ that describe the selected distribution. Each data point of an observation $(x_1, ..., x_n)$ has a probability

$$p(x_i \mid \theta_1, ..., \theta_k) = P(X = x_i) \tag{3.1}$$

The aim of ML estimation is to iteratively find a set of parameters θ that give the most plausibility that the acquired data points belong to the distribution described by parameters θ, which is the maximization of the product of the distribution probabilities of all data points, i.e., a likelihood estimation

$$L(\theta) = \prod_{i=1}^{n} p(x_i \mid \theta_1, ..., \theta_k) \tag{3.2}$$

where $L(\theta)$ is the probability to get an observation $(x_1, ..., x_n)$ from a distribution with k parameters $(\theta_1, ..., \theta_k)$, assuming the samples are independent. Usually the maximum of $L(\theta)$ can be obtained by maximizing a logarithmic likelihood function

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln p(x_i \mid \theta_1, ..., \theta_k) \tag{3.3}$$

Commonly, it is more computationally convenient to maximize the logarithmic likelihood function, which is also referred to as log-likelihood function, because only the sum up of the logarithmic probabilities needs to be calculated. In terms of computational complexity, multiplication is relatively more expensive than summation.

Additionally, log-likelihood function increases monotonically with its arguments due to the nature of summation, which simplifies the subsequent analysis. For example, for a Gaussian model, log-likelihood function can avoid exponential calculations, which will be illustrated in the following paragraph.

In this study, the ML method is used for estimating the Gaussian parameters of features in log-transformed RNA-seq or MS-based proteomics data. For Gaussian-distributed observations $(x_1, ..., x_n)$, the ML estimation aims to search for a Gaussian model with parameters μ and σ that describes the acquired data. Hence, the probability function of each data point x_i in Gaussian distribution $N(\mu, \sigma^2)$ is

$$p(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2} \tag{3.4}$$

The likelihood function is the product of the probabilities of all data points

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2} \tag{3.5}$$

The log-likelihood function can be finally simplified as

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.6}$$

3.2.2 Simple Linear Regression Model

Before the introduction of the GLM, consider a simple linear regression model which corresponds to the Gaussian family and has a single explanatory variable. Suppose here it has protein expression data from two experimental conditions (Refer to Chapter 3 in book [29] for detailed concepts). Typically y is defined as a response variable consisting of all observations y_i. Each observation y_i is believed to be drawn from a normal distribution with mean μ_i depending on the experimental condition x_i, which is also known as the explanatory variable. Hence, the response of each sample y_i has a linear relationship with its experimental condition/case as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3.7}$$

where β_0 and $\beta_1$are the coefficients to be estimated, x_i are explanatory variables, which are the experimental conditions, and ϵ_i are the errors that occurred in the course of measurements, also assumed to follow a normal distribution with zero mean and a constant variance

$$\epsilon_i \sim N(0, \sigma^2) \tag{3.8}$$

Thus, the expected response is

$$E(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.9}$$

In this linear relationship, $\hat{\beta}_1$ determines the slope of the linear trend and $\hat{\beta}_0$ is the intersect. In other words, $\hat{\beta}_1$ describes the effect size of the experimental conditions to the responses. Covariate x_i is usually a categorical vector, since it corresponds to experimental conditions. For a two-case experiment, coefficient $\hat{\beta}_1$ is a binary vector, which can be simplified as a vector with only 0 and 1 elements. Suppose the two coefficients β_0 and β_1 are known, there will be two expected responses:

1. When the observation is from a control sample, no effect is present, which can be denoted as $\beta_1 = 0$, the expected measurements would be the mean expression of the control group
$$E(y_i) = \hat{\beta}_0$$

2. When the observation is from a treatment sample, the effect is present, which can be denoted as $\beta_1 = 1$, the expected measurements would be the mean expression of the treatment group
$$E(y_i) = \hat{\beta}_0 + \hat{\beta}_1$$

3.2.3 Generalized Linear Model

GLM is the extension of the ordinary linear regression model, GLM gives more flexibility to model the error distribution of responses using various model families, whereas the linear regression model is only based on the Gaussian model. To build a GLM, three essential components need to be specified (Refer to Chapter 2 in book [36] for detailed concepts).

Firstly, the distribution of outcome vector y can be determined based on the characteristics of data. The chosen distribution usually belongs to the exponential family, which includes a wide range of distributions, such as continuous distributions (normal, Gamma) for continuous data and discrete distributions (Bernoulli, binomial, Poisson) to model binary and count data.

Suppose there are response variables $y_i (i = 1, ..., n)$ from n samples and k unknown parameters β to be estimated, a design matrix can be constructed by all observation vectors x_i as
$$X = [x_1^T, ..., x_n^T]^T$$

Secondly, a linear predictor η is defined to model the relationship between the location of response variables and the explanatory variables

$$\eta = X\beta \tag{3.10}$$

Lastly, a link function G is chosen to model the relationship between linear predictor η and the mean $E(y)$ as

$$E(y) = G^{-1}(\eta) = G^{-1}(X\beta) \tag{3.11}$$

Furthermore, to estimate the parameter vector β, the method of iteratively reweighted least squares (IRLS) is used to obtain the maximum likelihood estimates of a general linear model. The estimation is a L^p norm linear regression, which aims to find a vector of parameters β that can minimize the square sum of errors

$$\underset{\beta}{\arg\min} \|y - X\beta\|_2 = \underset{\beta}{\arg\min} \sum_{i=1}^{n} |y_i - X_i\beta|^2 \tag{3.12}$$

where vector β has initial values and is updated by iteratively solving

$$\beta_{new} = (X^T W X)^{-1} X^T W y \tag{3.13}$$

where W is a diagonal matrix of weights, the initial weights are set to 1, and the weights are also updated with β_{new}

$$w_i = |y_i - X_i\beta_{new}|^{-1} \tag{3.14}$$

3.2.4 t-test

For log-transformed mass spectrometry proteomics data and RNA-seq data, GLM was used to fit the expression data. The outcome variables were assumed to be normally distributed, and the link function between the expected value $E(y)$ and the linear predictor $X\beta$ was an identity function

$$E(y) = X\beta$$

In an experiment consisting of two sample groups, the estimated parameter was equivalent to the fold change of the two groups. The obtained t-statistics were the measurements of the goodness of fit, which was equivalent to the estimated parameter divided by the standard deviation of itself.

$$t_i = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)} \tag{3.15}$$

3.2.5 Wald Test

For raw RNA-seq read counts of each gene, the distribution used to model response variable Y was NB model. In addition, a log-link function was chosen to link the expected observation to the linear predictor as

$$ln(E(y)) = X\beta$$

the Wald statistics was the parameter estimate $\hat{\beta}_i$ divided by its standard error $s.e.(\hat{\beta}_i)$

$$z_i = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)} \tag{3.16}$$

3.2.6 Reproducibility-Optimized Test Statistic

In addition to Gaussian and negative binomial models, a modified t-test, namely Reproducibility Optimized Test Statistic (ROTS), was used for all types of data involved in this study. ROTS has been successfully implemented for various types of data, including RNA-seq, mass spectrometry proteomics and single-cell genomics data [37, 38]. A ROTS statistic is formulated as

$$d = \frac{|\bar{x}_a - \bar{x}_b|}{\alpha_1 + \alpha_2 s} \tag{3.17}$$

where $|\bar{x}_a - \bar{x}_b|$ is the absolute mean expression difference between groups A and B, α_1 and α_2 are the optimization parameters estimated by ROTS R package, and s is the pooled standard error. For example, when $\alpha_1 = 0$ and $\alpha_2 = 1$, the resulting ROTS statistics are equivalent to the standard t-statistics. The two additional parameters α_1 and α_2 aim to provide a more data-driven estimation of the effect sizes by offering two additional degrees of freedom.

3.3 Data Transformation and Normalization

Because the variable factors introduce biases in the experimental measurements, the obtained data often show variances that are dependent on their mean values. The mean-variance dependence, implying the absence of homoscedasticity, often induces difficulties in the downstream statistical analyses [39].

3.3.1 Systematic Bias in RNA-seq Data

Compared to other hybridization-based technologies, such as microarray, RNA-seq has become more popularized owing to its advantages [40]: less sequencing cost, increasing sequencing depth and higher reproducibility, etc. However, RNA-seq quantifies the reads of a transcript based on the quantities of detected mRNA fragments, each mapped read count depends on the sequence coverage of the whole transcript. In other words, a longer transcript usually tends to have more mapped reads due to higher coverage, whereas a shorter transcript with similar expression has lower coverage which results in fewer reads and higher variance [40–42]. In other words, RNA-seq only gives relative quantifications other than absolute measurements.

In order to make the read counts comparable across experiments, normalization methods are often required, such as Counts Per Million (CPM), Transcripts Per Million (TPM) [43] and Reads Per Kilobase of exon per Million reads mapped (RPKM), and Trimmed Mean of M values (TMM) [44]. For instance, TMM is implemented in DEA tool edgeR [6], it normalizes the gene read counts accounting for sequencing depth, RNA composition and gene length.

In this study, DESeq2 [45] is employed to process raw RNA-seq count data. It handles the read counts of each gene as variables that are proportional to the mapped numbers of corresponding cDNA fragments in the samples. Each gene read count was scaled by a normalization factor s, which was estimated using median-of-ratio method. To normalize gene counts of an experiment using median-of-ratio method, the normalization factors s were estimated in three main steps:

1. Calculate the pseudo-reference sample x_i^{pseudo} for each gene i, which was the geometric mean of the read counts of all k samples

   $$x_i^{pseudo} = \sqrt[\frac{1}{k}]{\prod_{n=1}^{k} x_{in}}$$

   where x_{in} is the read count of gene i in sample n, and k is the total number of samples.

2. Calculate the count ratio of each sample to the reference sample x_i^{pseudo} for each gene i as $\dfrac{x_{in}}{x_i^{pseudo}}$

3. The normalization factor for each gene i in sample s_i was then calculated using the median of the ratios across all genes within each sample n

   $$s_n = \underset{i}{\text{median}}\left(\frac{x_{in}}{\sqrt[\frac{1}{k}]{\prod_{n=1}^{k} x_{in}}}\right)$$

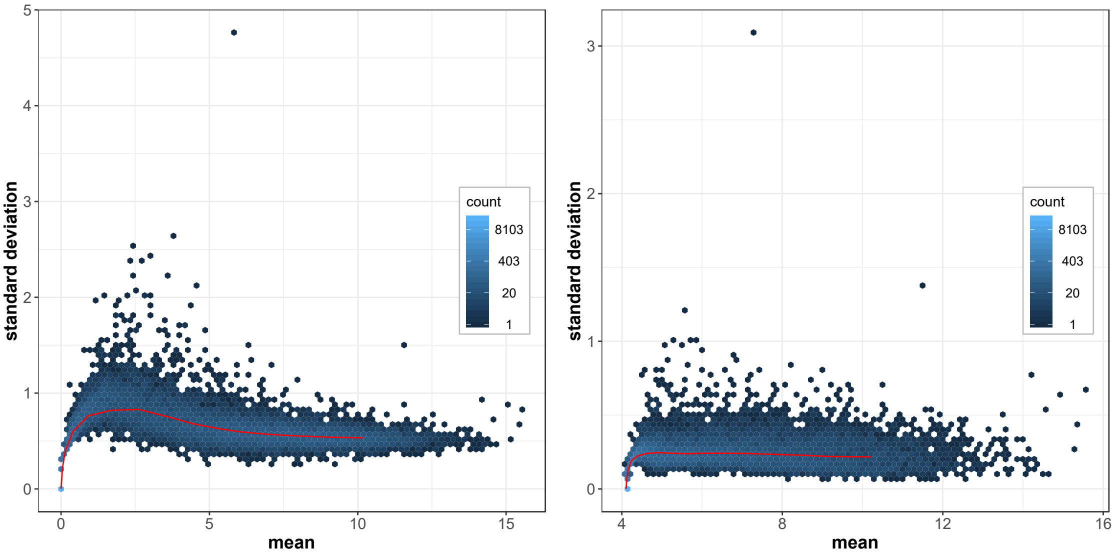
The read count x_{in} of each gene i in sample n is then scaled as $\dfrac{x_{in}}{s_n}$. For a single gene, a sample with more mapped reads had a larger normalization factor than those with less mapped reads [7]. As a result, the reads between samples could be brought into a similar scale, which made the sample groups adequately comparable.

In addition, DESeq2 uses variance stabilization transformation (VST) [46] to regulate the dispersion and fold change estimates by sharing information between genes. VST is initially designed to calibrate and transform microarray data, and it considers the variance-mean dependence into the model and uses maximum-likelihood methods to calibrate parameters of the transformation model. The VST transformation is similar to regular logarithm for large expression levels, but an error model is used to treat low expression levels by introducing bias, as genes with low expressions tend to show noisy signals in the measurements.

Fold changes between biological conditions are usually estimated based on log-ratios of gene expression levels, i.e., log fold change (LFC). However, log-ratios of weakly expressed genes can be highly variable. Evaluations [47] have shown that VST transformation can stabilize the variance across the gene expression data (**Figure 3.1**), improve the fold change estimates (**Figure 3.2**) and differential expression detection by reducing false positives resulting from measurement noise [45].

3.3.2 Systematic Bias in MS-based Proteomics Data

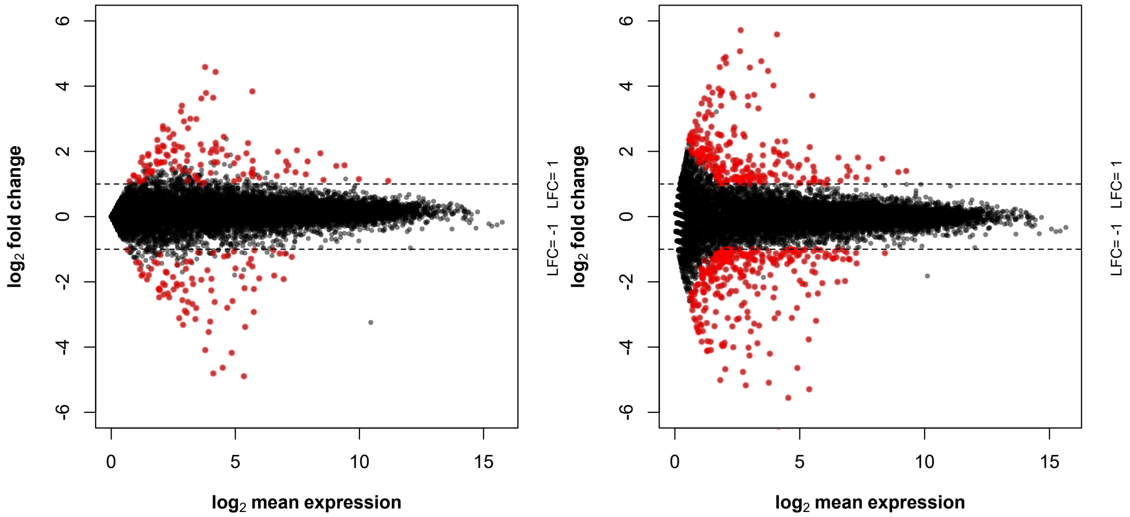
During the complex sample preparations, procedures and measurements in biological experiments, systematic biases can intrinsically exist due to many fickle environmental factors, such as biological conditions, instrument calibrations and temperature. However, the sources of the biases cannot be specifically targeted, unresolved biases may lead to incorrect conclusions in the downstream quantitative analyses [50–52]. In order to account for the biases and improve the data comparability between samples, many normalization methods are available [53]. In this study, variance stabilization normalization (VSN) [46] was the default method to normalize the MS-based proteomics data, as VSN has been tested to perform well with proteomics spike-in data [52], which were also used in the later benchmarking sections.



(a) mean-sd plot before using VST

(b) mean-sd plot after using VST

Figure 3.1: Demonstration of variance stabilization transformation. To demonstrate the effect of variance stabilization, VST was tested on an RNA-seq mice dataset [48]. The scope of a point on the red line indicated the local mean-variance dependence, variance-stabilized data were expected to have an approximately horizontal line. (a) Some genes with weak expression had larger standard deviation. (b) After using VST, the mean-variance dependence was much weaker, the red line indicated an approximately horizontal trend.



(a) LFC estimated using common approach

(b) LFC estimated using DESeq2

Figure 3.2: Improved LFC estimation using DESeq2. The LFC estimates and p-values of the mouse dataset [48] were estimated using both DESeq2 and ordinary t-test. The significant (p-value $<$ 0.05) genes were highlighted as red points. DESeq2 was able to detect more genes with weak differential expressions as significant compared to the ordinary method. (a) The LFC estimates were calculated as the log_2 mean ratios between two groups and the p-values were obtained from ordinary t-test and adjusted using BH (Benjamini and Hochberg) method [49]. (b) The LFC estimates and adjusted p-values were obtained from DESeq2.

3.4 Data Distribution Estimation

3.4.1 Log-transformed RNA-seq and MS-based Proteomics Data

Regarding log-transformed RNA-seq and MS-based proteomics data, the log-transformed gene read counts and protein abundance levels were modeled as random normal variables. A random normal variable x_{ijn} has a probability density function (Also see function 3.1 in Section 3.2.1)

$$p(x_{ijn} \mid \mu_{ij}, \sigma_{ij}^2) = \frac{e^{-\frac{1}{2\sigma_{ij}^2}(x_{ijn}-\mu_{ij})^2}}{\sqrt{2\pi\sigma_{ij}^2}} \tag{3.18}$$

where i is the index of features, j is the index for experimental conditions, n is the index for samples and x_{ijn} contains the gene read counts or protein abundance levels, μ_{ij} is the mean and σ_{ij} is the standard deviation.

Parameters μ_{ij} and σ_{ij} of a normal distribution were estimated using maximum likelihood approach, which was aimed to find a set of parameters (μ_{ij} and σ_{ij}) that maximize the likelihood function (Also see function 3.2 in Section 3.2.1)

$$L(x_{ijn} \mid \mu_{ij}, \sigma_{ij}^2) = \frac{e^{-\frac{1}{2\sigma_{ij}^2}\prod_{n=1}^{n}(x_{ijn}-\mu_{ij})^2}}{\sqrt[\frac{n}{2}]{2\pi\sigma_{ij}^2}} \tag{3.19}$$

In addition, the log-likelihood function (Also see function 3.3 in Section 3.2.1) is

$$LL(x_{ijn} \mid \mu_{ij}, \sigma_{ij}^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma_{ij}^2) - \frac{1}{2\sigma_{ij}^2}\sum_{n}^{n=1}(x_{ijn}-\mu_{ij})^2 \tag{3.20}$$

The normal distribution parameters (μ_{ij} and σ_{ij}) of each gene or protein i in condition j were estimated by maximizing $LL(x_{ijn} \mid \mu_{ij}, \sigma_{ij}^2)$.

3.4.2 Raw Gene Read Counts

To model RNA-seq data as NB distribution, the lack of replicates is the common challenging issue, which usually results in unstable estimates of dispersion. However, with the enormous amount of genes, the stability of estimates can be improved by borrowing the information from other genes by accounting for the relationship between mean and dispersion [33]. DESeq [7] and edgeR [44] are the popular methods, they model the gene-wise dispersion estimates to obtain a global reference estimate. This study utilized the successive version of DESeq, namely DESeq2, which provides moderated estimation methods for LFC and dispersion of RNA-seq data.

DESeq2 provides a hierarchical model to estimate dispersion and LFC. It first uses ML to obtain the gene-wise dispersion estimates ϕ_{ij}^{gw} depending on the data of each individual gene. Next, the obtained ϕ_{ij}^{gw} are fitted to a smooth curve based on the global dependency relationship between dispersion and mean. Based on an empirical Bayes method, the model then shrinks the gene-wise dispersion estimates ϕ_{ij}^{gw} toward the smooth curve to obtain final estimates. The adjustment is based on the assumption that genes with similar average expression strength should have approximately equivalent dispersions. As shown in **Figure 3.3**, most of the estimated ϕ_{ij}^{gw} (black dots) were shifted toward the red smooth curve and became the final estimates (blue dots). Some ϕ_{ij}^{gw}, which had extremely large dispersions, were eventually kept as the original ϕ_{ij}^{gw} because they were distinctly different (more than 2 residual standard deviations away from the curve) from other genes with similar mean expression level [45].

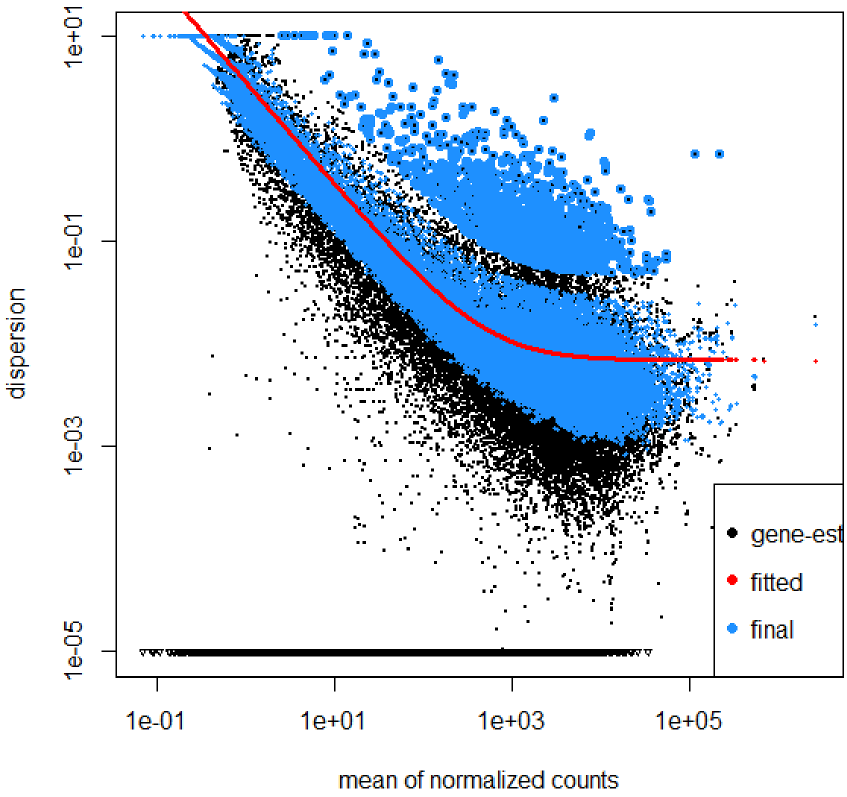


Figure 3.3: Dispersion estimation using DESeq2. After the gene-wise dispersions (black dots) were estimated using maximum likelihood method, the data points were fitted onto a smooth curve, and the gene-wise dispersions were adjusted toward the smooth curve (red curve) to achieve final estimates (blue dots). Some extreme dispersion estimates (black dots circled with blue), which were 2 residual standard deviations away from the fitted curve, were kept as the gene-wise estimates, which can be caused by biological or technical factors.

3.5 Simulation and Power Calculation

3.5.1 Monte Carlo Simulations

The data simulation was based on the Monte Carlo method [54], which is a simulation procedure that iteratively samples new data points multiple times from a probability distribution. The utilization of Monte Carlo method was based on the assumption that the data collection of an experiment with sample size n is an action of randomly sampling n possible observations from the nature following a certain order. After distribution estimation, the collected data were represented by the estimated distribution parameters. Re-sampling new data from the estimated distribution created the outcomes of the same experiment for unlimited times without requiring the actual biological experiments.

In the simulations, each feature was simulated T times both under null and alternative hypotheses. To achieve better clarity, the data simulated under null hypothesis will be referred to as "null data", whereas the term "alternative data" stands for the data simulated under alternative hypothesis. After performing hypothesis tests on the simulated data, statistics for null and alternative data were produced. Statistics resulting from both hypotheses were respectively given terms "null statistics" and "alternative statistics". Note that the absolute values of statistics were used because this study only accounted for the degrees of up- or down-regulated expression changes to estimate the statistical power.

As mentioned in Chapter 1, even if the true effect size in an experiment is assumed to be a constant value, the actual measured effect size by the statistical model will have some levels of variability due to the random sampling error. Experiments with low sample power may give measurements that conclude significant results but only because of the occasionally occurred measurements that led to a large deviation [1]. To avoid the false discovery of significant results at a desired level, the repeated simulations and tests for null hypothesis were intended to obtain a data-specific null distribution accounting for the variances in the input data.

Furthermore, a threshold statistics c was calculated from the null distribution to control a fraction of FP detections, which represent the scenarios when the null hypothesis is rejected even if it is true. As shown in **Figure 3.4**, based on a user-specified FDR ($\alpha = 5\%$ by default), the threshold statistics c was yielded as $(1-\alpha) \times 100$th percentile of null statistics. The threshold statistics c divided the null distribution into two parts, which were the null hypothesis acceptance and rejection regions. The null statistics within the acceptance region were consistent with the null hypothesis, whereas the rejection region contained the test statistics that were opposite to the null hypothesis. Based on the same threshold c, the acceptance and rejection regions of alternative hypothesis were also determined, as when the null hypothesis is rejected, the paired alternative hypothesis will be accepted. Moreover, the alternative statistics resulting in the acceptance region of the alternative distribution were marked as TPs. The proportion of alternative statistics in the rejection region of alternative hypothesis, denoted as β, was also the FN rate, which is complementary to the statistical power $(1 - \beta)$.

For instance, after the parameters of probability distribution were estimated from a two-group dataset, new data were simulated in following steps:

1. Null data were simulated under null hypothesis in scenarios of both groups, between which no fold changes were added.

2. Simulated null data went through hypothesis tests, which will result in two null statistics for both scenarios, the maximum of the two null statistics was kept as the final null statistics for the current feature.

3. Similar to Step 2, alternative data were simulated following alternative hypothesis. The estimated fold changes of the original dataset were applied between two groups.

4. The alternative statistics were obtained from the hypothesis tests.

5. Step 1 - 4 were repeated for T times to produce two vectors of null statistics S_0 and alternative statistics S_1. Both had length T.

6. A threshold null statistics c, by default, was calculated as the 95th percentile of the S_0 to determine the acceptance and rejection regions of both null and alternative distributions.

7. The power estimates were then calculated by the proportion of alternative statistics S_1 falling in the acceptance region of alternative distribution.

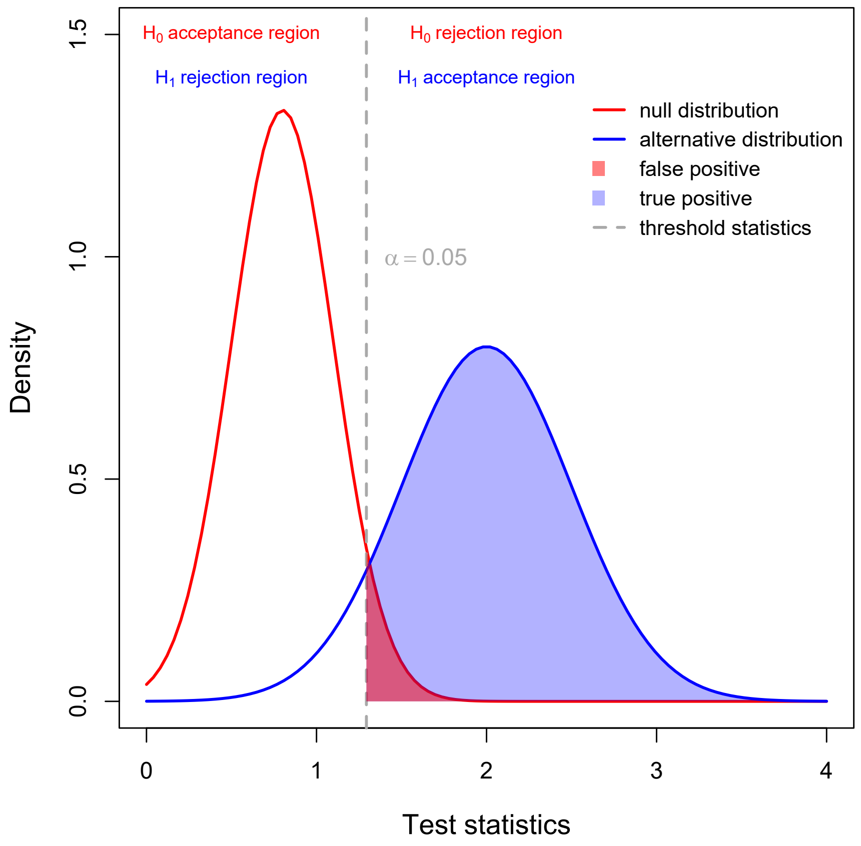


Figure 3.4: Interpretation of power calculation. The statistics in vectors S_0 and S_1 resulting from the simulations under null (H_0) and alternative (H_1) hypotheses were the null (red curve) and alternative (blue curve) distributions. By default, a threshold statistics c_α was determined as the 95th percentile of S_0 to yield the intervals of acceptance and rejection regions of null hypothesis. Null statistics that exceeded threshold c_α, i.e., null statistics in rejection region (red fill), were marked as FPs. Furthermore, the alternative statistics falling in the acceptance region of alternative hypothesis were marked as TPs. Additionally, the fraction of alternative statistics that were larger than c_α (blue fill) was equivalent to the statistical power.

3.5.2 Power Calculation

For a null hypothesis, the obtained null statistics S_0 followed a null distribution. Furthermore, there was a probability, often denoted as α, that the null hypothesis would be rejected even if it was true, which resulted in FPs. The null statistics representing FPs are determined by comparing with the $(1-\alpha) \times 100$th percentile of null statistics in S_0

$$FP = \{X \mid x \geq c_\alpha \cap x \in S_0\} \tag{3.21}$$

where c_α was a threshold value c for a user-specified FDR ($\alpha \in (0, 1]$) to yield the acceptance and rejection regions of both null and alternative distributions.

Furthermore, the threshold value c_α determined the TPs among the alternative statistics vector S_1 as

$$TP = \{X \mid x > c_\alpha \cap x \in S_1\} \tag{3.22}$$

Whereas the FNs were determined among alternative statistics S_1 as

$$FN = \{X \mid x \leq c_\alpha \cap x \in S_1\} \tag{3.23}$$

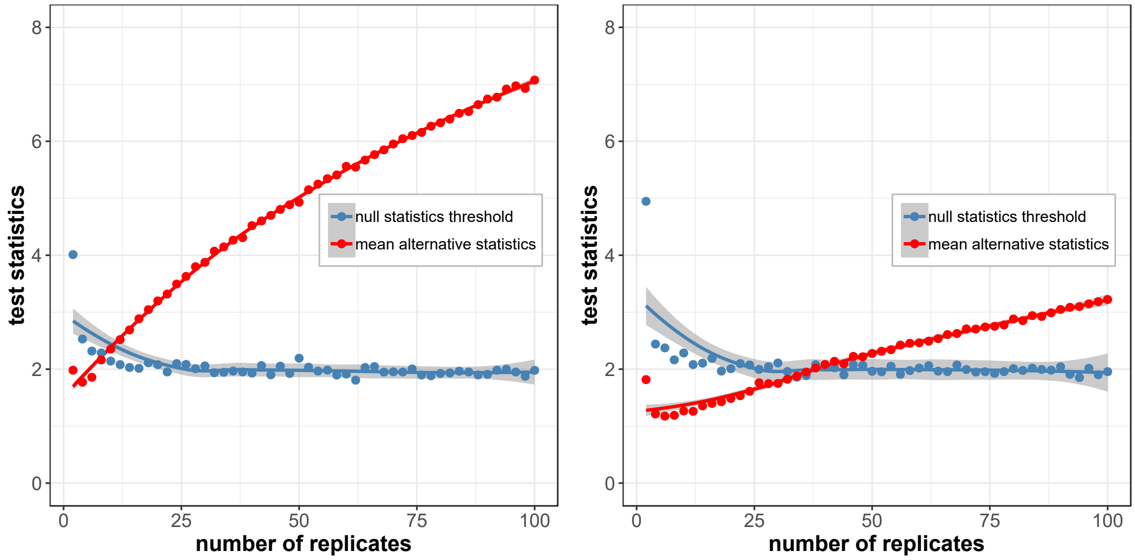
Finally, the fraction of alternative statistics in S_1 that were greater than c_α concluded the power as

$$Power = \frac{|TP|}{|S_1|} \times 100\% \tag{3.24}$$

3.5.3 Demonstration of Power Estimation

For a more interpretable demonstration, this section presents the power estimation in a few examples. In addition, the effect of data variance on power estimates is also illustrated. Some features, which have similar mean expressions, can have distant power estimates because of the differences in variations. In the demonstrations, normally distributed data were generated respectively from Gaussian distributions with small and large variations.

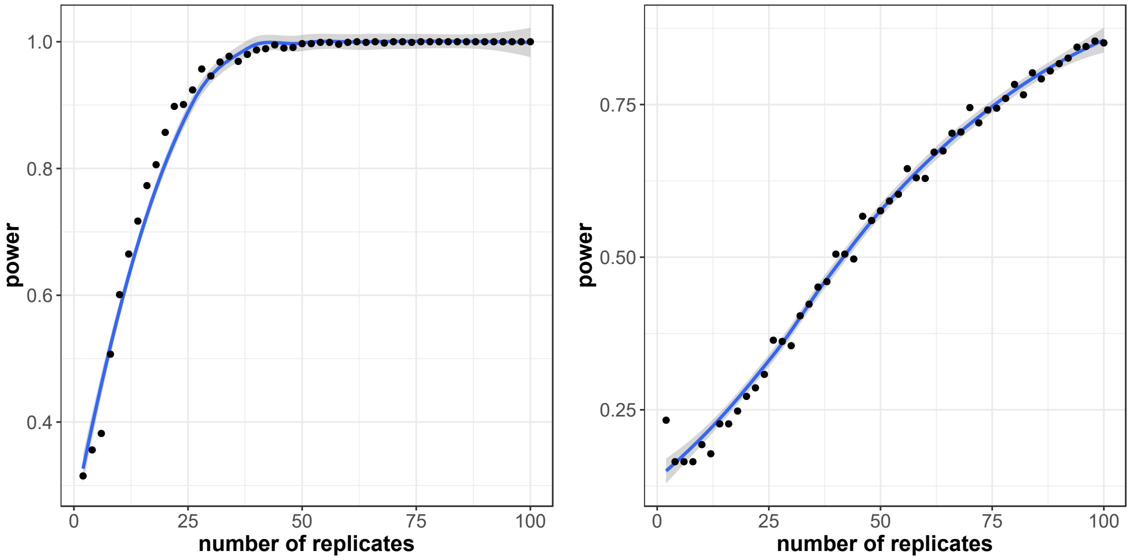
As shown in **Figure 3.5a**, the generated data with small variance resulted in clear differences between threshold statistics and the alternative statistics, which led to a rapid increase rate in power estimates (**Figure 3.6a**). When the data had a larger variance, the distances between null and alternative statistics became closer as illustrated in **Figure 3.5b**. As a result, the power estimates had a moderate increase rate with the increasing number of replicates (**Figure 3.6b**).



(a) t-statistics from data with small variance

(b) t-statistics from data with large variance

Figure 3.5: The effect of feature variance sizes on the resulting test statistics. (a) t-statistics were obtained from simulations both under null and alternative hypotheses. The data points were randomly generated from two normal distributions $N(\mu = 10, \sigma^2 = 1)$ and $N(\mu = 11, \sigma^2 = 1)$ 1,000 times. The threshold null statistics (green dots and curve) substantially remained at the same level for the increased number of replicates, whereas the alternative statistics (red dots and curve), for more sample replicates, had a large increase rate compared to the threshold values. (b) Similar to the previous case, the data points were randomly generated from distributions with the same means but larger variance ($\sigma^2 = 3$). Overall, the alternative statistics increased with larger number of replicates, but exceeding the threshold null statistics required much more replicates, compared to the previous case with small variance.



(a) Power estimates for data with small variance

(b) Power estimates for data with large variance

Figure 3.6: The effect of feature variance sizes on resulting power estimates. (a) Power estimates for the t-statistics obtained from the small-variance case. The power estimates increased rapidly with larger number of replicates. (b) Power estimates for the t-statistics obtained from the large-variance case. The power estimates increased with a smaller scope while the number of replicates was increased.

3.6 Data Descriptions

For the performance assessment of the proposed methods in this study, simulated datasets for both RNA-seq and MS-based proteomics cases were created, in which the known DE and equivalently expressed (EE) features were added. In addition, the found public datasets from spike-in RNA-seq and protein mixture experiments were used for the benchmarking with biological data.

3.6.1 Simulated RNA-seq Read Counts

Six RNA-seq datasets (A-F) were simulated by randomly generating data from NB distributions. Each dataset contained 10,000 genes, of which 80% were EE and the other 20% genes were DE with two-fold change between two sample groups. The mean expression and the number of replicates were different between datasets, as summarized in **Table 3.1**. The dispersions were calculated based on the mean-dispersion relationship estimated by using public mice dataset [48]. These settings allowed the capability to observe the dependency between power estimates and the data with changed attributes.

Table 3.1: Simulated RNA-seq datasets

Dataset ID	Mean expression (μ)	Fold change (λ)	# of replicates (n)
A	10	2	10
B	100	2	10
C	500	2	10
D	1000	2	10
E	10	2	50
F	100	2	50

3.6.2 Simulated MS-based Proteomics Datasets

With similar settings as the simulated RNA-seq datasets, six protein datasets, namely A-F, were assembled with randomly generated numbers from normal distributions with different mean values as listed in **Table 3.2**. Each dataset was created with 2,000 DE and 8,000 EE proteins randomly mixed. The DE proteins were specified to have two-fold change between two groups. Proteins with four different $\log_2$ mean abundance levels and two different numbers of replicates ($n = 10$ and $n = 50$) were generated to investigate the performance of MS-based proteomics power analysis.

Table 3.2: Simulated MS-based proteomics datasets

Dataset ID	**Log2 mean expression (μ)**	**Fold change (λ)**	**# of replicates (n)**
A	5	2	10
B	10	2	10
C	15	2	10
D	20	2	10
E	5	2	50
F	10	2	50

3.6.3 Dataset of RNA-seq Read Counts with Known Spike-in Genes

To assess the performance of the power analysis method on biological data, the assessment used a spike-in RNA-seq dataset sourced from RNA Sequencing Quality Control (SEQC) project. The SEQC project was designed to evaluate the performances of RNA-seq technologies based on various sequencing platforms [55].

In this study, only the expression data of External RNA Control Consortium (ERCC) spike-ins in samples A and B were used to assess the performance of the proposed methods because only ERCC had exactly known concentration ratios and expected LFCs, as summarized in **Table 3.3**. Due to the limitation of sequencing technology, the number of detectable fragments decreases with lower concentration of ERCC genes. Hence, the groups with lower LFCs were expected to have much more low read counts, which introduced large variances. The ERCC spike-ins organized in four groups had different concentration ratios, the power estimates for the four groups were expected to show four noticeably distinct ranges. As shown in **Table 3.3**, for example, group I had the highest concentration ratio at 4 (The expected LFC was 2). The number of significant power estimates in group I was expected to be the greatest.

Table 3.3: Summary of ERCC spike-in genes utilized in SEQC project

Spike-in groups	Concentration ratio	Expected LFC	Total # of spike-ins
I	4	2	23
II	1	0	23
III	0.67	-0.58	23
IV	0.5	1	23

3.6.4 MS Measurements of Heterogeneous Protein Mixtures

Additionally, an MS-based proteomics dataset [56] was used to benchmark the performance of proteomics power analysis. The proteomics dataset was derived from a comparison experiment designed to assess the performance of various label-free quantification (LFQ) tools. The dataset contained protein abundances from two hybrid proteome samples and each sample had five technical replicates. Both samples were the mixtures of tryptic digests sourced from *E. coli*, human, and yeast proteomes. The concentration ratios between the two samples were precisely known (1:1 for human, 2:1 for yeast and 1:4 for *E. coli* proteins), as shown in **Figure 3.7**. Similar to the RNA-seq case, most high power estimates were expected to appear among the known DE proteins. Hence, in this case, most of high power estimates were expected to be in both yeast and *E. coli* proteins, whereas the background human proteins should be much less powerful because of the equal concentrations between two groups.

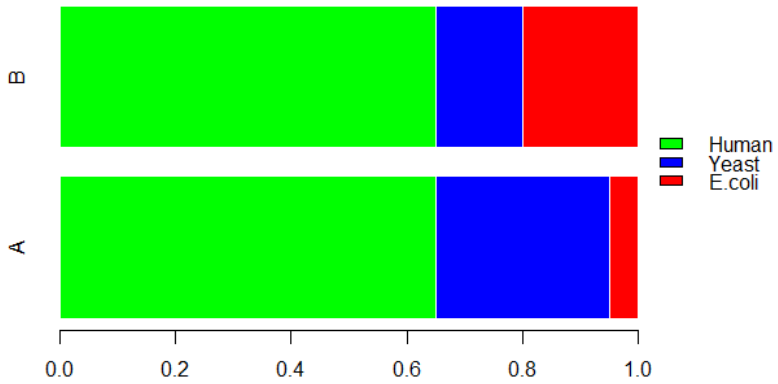


Figure 3.7: Mixture ratios of heterogeneous proteins in samples A and B. Sample A and B were prepared with background human proteins in equal concentration, and then proteins derived from non-human organisms were added into both samples A and B in concentration ratios 2:1 for yeast and 1:4 for *E. coli*.

Chapter 4

Implementation and Results

This chapter will demonstrate the benchmarking results based on both simulated and real biological data. The power estimates were accessed by referring to the known DE and EE features, expecting most of the high power estimates would appear among DE features. The proposed methods have been implemented as an R package called PowerExplorer which is openly available in Bioconductor project. The R package and illustrative manuals can be downloaded at http://bioconductor.org/packages/PowerExplorer/.

4.1 Data preprocessing

4.1.1 Raw RNA-seq Read Counts

By default, the RNA-seq raw read counts were processed by the function `vst()` in R package DESeq2 [45], which is an openly available in the Bioconductor project. For each of the six datasets, DESeq2 normalized the read counts using method VST and estimated the dispersion and LFC of each gene using Empirical Bayes approach (Refer to **Section 3.3.1** for more method details).

4.1.2 Log-scale RNA-seq and MS-based Proteomics Data

For the cases where RNA-seq or proteomics data were transformed into logarithmic form, the data were assumed to follow normal distribution. VSN [46] was used to transform and normalize the data. Additionally, the distribution parameters were estimated using ML approach (Refer to **Section 3.4.1** for method details).

4.2 Performance assessments

When using R package PowerExplorer, for both retrospective power analysis (RPA) and prospective power analysis (PPA), parameters, such as minimal log fold change (LFC), false discovery rate (FDR) and simulation times, need to be specified. For benchmarking purpose, the minimum LFC was set to zero so that genes and proteins within all ranges of fold changes were included into the calculation. The simulations were repeated for 1000 times for each dataset and the FDR was the default value 0.05. For PPA, power estimates for increased amount of replicates ($n = 5, 10, 15, ..., 50$) were also calculated.

The performance of PowerExplorer was assessed using the six simulated RNA-seq and proteomics datasets described in **Section 3.6**, which was aimed at observing the effects on power estimates from the three main factors, i.e., fold change, mean expression level and the number of replicates.

4.2.1 Assessment Results for Simulated RNA-seq Data

As expected, genes with stronger expression levels had higher power estimates as shown in **Table 4.1**. When each sample had ten replicates (datasets *A*-*D*), the datasets with larger mean expression levels had substantially larger proportion of DE genes estimated with high power (power $\geq$ 0.8), compared to the dataset with lower mean expression level. For instance, with the same fold change (FC=2) and number of replicates (n=10), there were many more DE genes estimated with high power in dataset *D* as compared to dataset A, which has the lowest mean expression level (*A: 1052/2000, B:1784/2000, C: 1853/2000 and D: 1869/2000*). In addition, the proportions of DE genes estimated with high power increased to 100% when the number of replicates was 50 (datasets E and F), whereas only a negligible fraction (<0.5%) of EE genes showed unexpected high power estimates (*A: 15/8000, B:21/8000, C: 24/8000, D: 19/8000, E: 41/8000 and F: 28/8000*) due to the random variation resulting from data simulation.

Table 4.1: Summary of power estimates for simulated RNA-seq datasets

Dataset ID	# of replicates	Mean expression	# of DE high power genes	# of EE high power genes
A	10	10	1052	15
B	10	100	1784	21
C	10	500	1853	24
D	10	1000	1869	19
E	50	10	2000	41
F	50	100	2000	28

4.2.2 Assessment Results for Simulated MS-based Proteomics Data

Similarly, proteins with larger $\log_2$ mean expression level were estimated to be more powerful. With ten replicates (datasets A-D), the power estimates continuously increased with higher $\log_2$ mean expression level (*A: 568/2000, B: 1936/2000, C: 1990/2000, D: 1998/2000*). When the the number of replicates increased to 50, almost all the DE proteins were estimated with high power as expected (*E: 1994/2000, F: 2000/2000*). In contrast, the number of EE proteins with high power (false estimates) consistently remained low (*A: 10/8000, B:18/8000, C: 10/8000, D: 14/8000, E: 28/8000 and F: 26/8000*).

Table 4.2: Summary of power estimates for simulated proteomics datasets

Dataset ID	# of replicates	$\log_2$ mean expression	# of DE high power proteins	# of EE high power proteins
A	10	5	568	10
B	10	10	1936	18
C	10	15	1990	10
D	10	20	1994	14
E	50	5	1998	28
F	50	10	2000	26

4.2.3 Assessment Results for Spike-in RNA-seq Data

Using the found public RNA-seq dataset with 92 known spike-in genes, the performance of the proposed methods were evaluated based on the power estimates of the ERCC spike-in genes. The power estimates of the ERCC spike-ins were summarized in four *groups I-IV* by the expected fold changes between samples A and B.

Retrospective power analysis (RPA)

A few ERCC genes were removed due to excessive zero counts (less than two non-zero counts). Among the four spike-in *groups I-IV*, much more high power estimates were found in *group I*, whereas *group II* had the least significant power estimates (*I: 17/21, II: 3/22, III: 6/22, IV: 13/22*). As expected, the ERCC genes with larger concentration ratios were estimated to be more powerful. Due to the gradient concentration, ERCC genes in *group II and III*, which had relatively small or no effect sizes between sample A and B, were estimated with the lowest number of high power genes (**Table 4.3**).

Table 4.3: Retrospective power estimates of ERCC spike-ins

Spike-in groups	Expected LFC	# of filtered spike-ins	# of high power spike-ins
I	2	2	17
II	0	1	3
III	-0.58	1	6
IV	-1	1	13

Prospective power analysis (PPA)

For more amount of replicates ($n = 5, 10, 15, ..., 50$), PPA was performed expecting obvious increases of power estimates among ERCC spike-in genes, while the noticeable distances between groups still remained. As summarized in **Figure 4.1**, the increased number of replicates resulted in more ERCC genes estimated with significant power. *group I* and *IV* with larger fold changes showed faster increases of power. When $n = 15$, *group I* and *IV* already had more than 90% genes with significant power, by contrast, only around 50% of genes in *group II* and *III* were powerful.

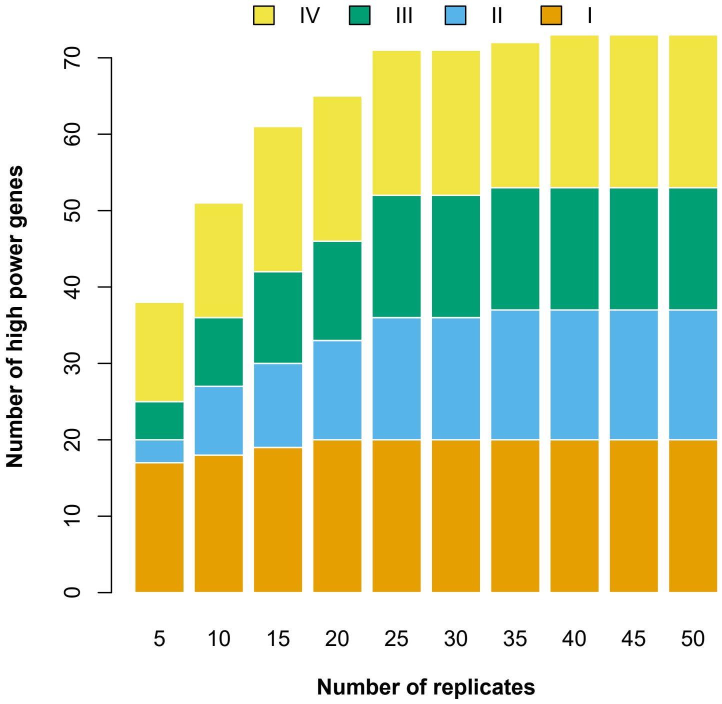


Figure 4.1: Prospective power estimates of ERCC spike-in genes. Overall, the number of high power genes increased with more replicates. ERCC groups *I* and *IV*, which had the largest LFC (*IV: -1 and I: 2*), remained with a large fraction of high power genes. Whereas groups *II* and *III* with the lowest LFC (*II: 0 and III: -0.58*) had a moderate increase in power estimates.

4.2.4 Assessment Results for Heterogeneous Protein Mixtures Data

To benchmark the performance on proteomics data, as described in **Section 3.6.2**, a public MS-based proteomics dataset was used. The non-human proteins were expected to have more high power estimates than the human background proteins, which were prepared in equal concentrations between two sample groups.

Retrospective power analysis (RPA)

As expected, human background and non-human (yeast and *E. coli*) proteins were estimated with distinct results as shown in **Figure 4.2**. Among non-human proteins, large proportions of proteins were estimated with high power (*Yeast: 766/864, E.coli: 666/699*). In contrast, 74.84% (1029/1375) of the estimates for human background proteins were insignificantly powerful.

Prospective power analysis (PPA)

For the increased number of replicates $n = (5, 10, 15, ..., 50)$, power estimates of human background proteins and non-human proteins were shown to be distinct, as displayed in **Figure 4.3**. With the fewest replicates ($n = 5$), 89.00% of yeast proteins and 95.99% of *E. coli* proteins were significantly powerful, whereas the percentage of human background proteins estimated with higher power was only 26.61%. Moreover, the percentages of high power for both non-human proteins had exceeded 95% with 15 replicates, whereas the high power percentage of human proteins only increased to about 50% (Yeast: 95.83%, *E. coli*: 98.71% and Human: 50.66%).

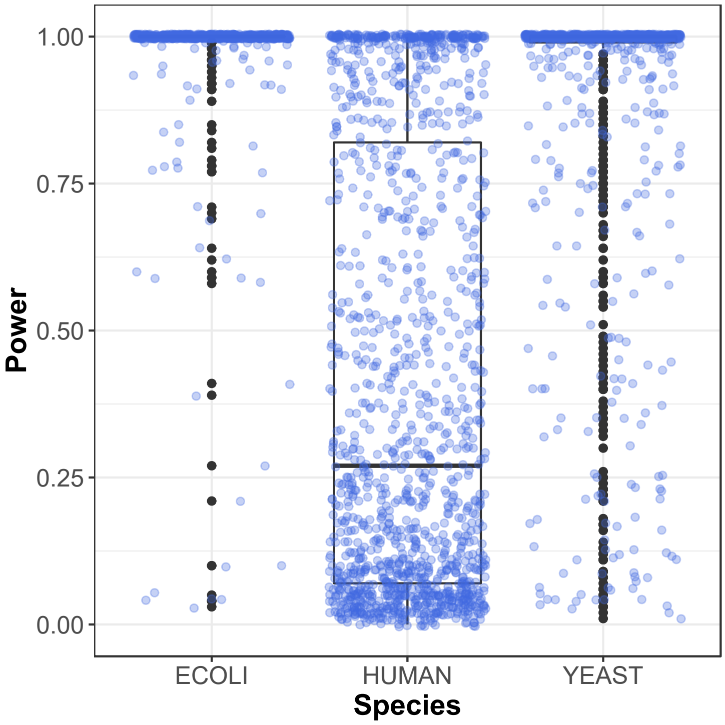


Figure 4.2: Retrospective power estimates for protein mixtures dataset. For human proteins, there was a large amount of low power estimates, the mean estimate is lower than 0.3. In contrast, non-human proteins were mostly with high power as expected, the mean power estimates of yeast and *E. coli* were almost 1, the highest power level.

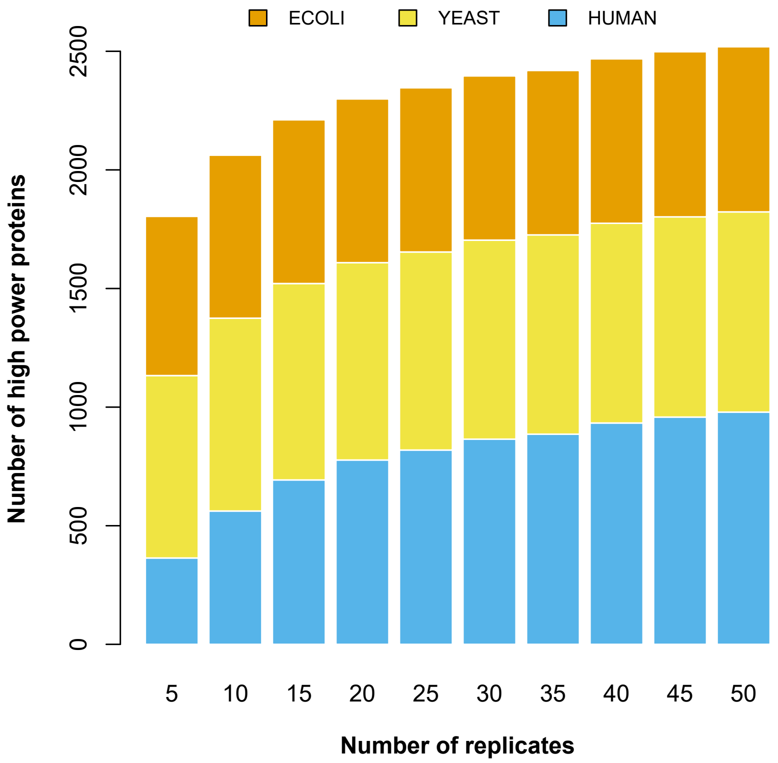


Figure 4.3: Prospective power estimates for protein mixtures dataset. For the heterogeneous protein mixtures dataset, the power estimates substantially increased with more replicates. A large fraction of non-human proteins remained with high power estimates since when having the fewest replicates ($n = 5$), whereas the number of high power human proteins gradually increased with more replicates but remained distinct to the one of non-human proteins.

Chapter 5

Discussion

In this study, the proposed methods were able to carry out power analysis for both RNA-seq and MS-based proteomics data. Additionally, they were capable of both retrospective and prospective power analyses based on acquired datasets. The work-flow contained three main components: parameter estimation, data simulation and power calculation. Firstly, the parameter estimation converted the provided dataset into parameter vectors, which approximately described the feature-wise distributions of the expression data. Secondly, the data simulation utilized the resulting parameter vectors to generate null data, where no effect sizes were introduced, and alternative data, where the actual effect sizes were added. Thirdly, hypothesis tests were performed for both null and alternative data. The statistics resulting from tests on null data were used to calculate a boundary statistics for a user-specified FDR. The boundary statistics was used to divide null statistics into representatives for TN and FP detections. Lastly, the boundary statistics was also the divisional statistics that yields the FN and TP detections among alternative statistics.

However, a few existing power analysis methods are usually reliant on DE/EE detection results from other DEA methods [57, 58], the proposed method in this study can estimate the power only based on the actual statistical characteristics of a provided dataset without the assistance of other DEA methods. In addition, some power analysis methods also require prior assumptions, such as expected mean expression, the proportion of DE features and minimal fold changes [57–60]. Usually, the significant features in an experiment are unknown for the researchers, thus an estimation based on uncertain assumptions may be misleading. In contrast, the proposed methods only required the user specify the FDR to control a proportion of false positives among hypothesis tests, as well as the desired preprocessing methods and hypothesis testing models.

The performance of the proposed methods were tested using simulated data containing known DE features and two public datasets with known concentrations of gene or protein mixtures. The evaluation was based on the comparison between the acquired and expected results. Fundamentally, it was assumed that, compared to weakly expressed or EE features, highly DE features should have higher power estimates. The benchmarking results indicated that the proposed methods were capable of labeling the predetermined DE features with high power estimates, whereas the power estimates for EE features were substantially low as expected. Additionally, the power estimates were significantly greater for features with large effect sizes, more sample replicates or stronger mean expressions.

Furthermore, with the alterable parameter of sample sizes, the simulation-based methods brought more flexibility to investigate the relationship between sample power and the sample size. Overall, the shown performance indicated that, in addition to the power investigation of the provided data, the proposed power analysis methods can be applied to assist the prospective experimental designs, where one of the main problems is to anticipate a proper sample size to achieve the adequate sample power.

Chapter 6

Conclusion and Perspective

6.1 Conclusions

Reproducibility crises took place in many academic fields, which brought a lot of questions into the published statistical results in the past literatures. Hence, sample size and power analysis have been seen as essential procedures dedicated for the evaluation of reproducible significant discoveries. In this study, a series of power analysis methods were proposed and implemented for both simulated data and biological data. The performance of the proposed methods were assessed using both simulated data and public biological data. In addition, the proposed methods were implemented as an R software package, namely PowerExplorer. The package is openly available in Bioconductor project and can be freely downloaded at http://bioconductor.org/packages/PowerExplorer/.

6.2 Future improvements

Despite the performance, the proposed methods require intensive computational power. However, the published R software package regarding this study provides the option of using parallel computation. Nevertheless, in terms of the computational consumption, the high demand can be further reduced by optimizing the simulation component of the method, such as using C or C++ programming language to potentially achieve faster executions. Moreover, the methods currently only support RNA-seq and MS-based proteomics data. The types of supported data can be diversified if allowing more optional statistical models and hypothesis tests to be appended. Furthermore, it is also possible to add parameters allowing users to attach external functions for statistical models and hypothesis tests, which can further improve the feasibility of the methods.

In terms of the performance, the benchmark results have shown that some EE or weakly expressed features can have noticeable increase in power estimates, which are often caused by the additional variations sourced from features with excessive missing values. In spite of the fact that the features with excessive missing values and overly low expressions can be removed from the estimations, the attempt of keeping most of the features in the downstream analyses, if correctly treated, may potentially lead to interesting discoveries. However, the estimation will perhaps be improved if some further penalty mechanisms can be added for the features with missing values. The penalties intend to decrease the power estimation for large amounts of missing values or extremely low expressions.

Furthermore, it is known that a small variance in an observation indicates that most of the data points are close to the mean value. When the sample size increases, the variability of the sampling distribution gets smaller, since more performed random samplings tend to get more values that are close to the true mean. Based on such empirical assumption, the prospective power analysis may be able to be improved by decreasing the variances of the simulated features when the sample size is increased. For this purpose, the experiment-specific relationships between data variability and sample size still need to be well studied.

References

1. Szucs, D. & Ioannidis, J. P. A. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology* **15** (ed Wagenmakers, E.-J.) e2000797. ISSN: 1545-7885 (Mar. 2017).

2. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454. ISSN: 0028-0836 (May 2016).

3. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1,** 0021. ISSN: 2397-3374 (Jan. 2017).

4. Halsey, L. G., Curran-Everett, D., Vowler, S. L. & Drummond, G. B. The fickle P value generates irreproducible results. *Nature Methods* **12,** 179–185. ISSN: 1548-7091 (Mar. 2015).

5. Li, C.-I., Samuels, D. C., Zhao, Y.-Y., Shyr, Y. & Guo, Y. Power and sample size calculations for high-throughput sequencing-based experiments. *Briefings in Bioinformatics,* bbx061. ISSN: 1467-5463 (June 2017).

6. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26,** 139–40. ISSN: 1367-4811 (Jan. 2010).

7. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106. ISSN: 1465-6906 (Oct. 2010).

8. Jarvius, J. *et al.* Digital quantification using amplified single-molecule detection. *Nature Methods* **3,** 725–727. ISSN: 1548-7091 (Sept. 2006).

9. Farrell, R. E. in *RNA Methodologies* 4th, 45–80 (Elsevier, 2010). ISBN: 9780123747273. doi:10.1016/B978-0-12-374727-3.00002-4. <http://linkinghub.elsevier.com/retrieve/pii/B9780123747273000024>.

10. Bryant, S. & Manning, D. L. in *Methods in molecular biology (Clifton, N.J.)* 61–64 (Humana Press, Totowa, NJ, 1998). ISBN: 1064-3745. doi:10.1385/0-89603-494-1:61. <http://link.springer.com/10.1385/0-89603-494-1:61>.

11. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLOS Computational Biology* **13,** e1005457. ISSN: 1553-7358 (May 2017).

12. Govindarajan, R., Duraiyan, J., Kaliyappan, K. & Palanisamy, M. Microarray and its applications. *Journal of pharmacy & bioallied sciences* **4,** S310–2. ISSN: 0975-7406 (Aug. 2012).

13. Clark, T. A. Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. *Science* **296,** 907–910. ISSN: 00368075 (May 2002).

14. Shendure, J. The beginning of the end for microarrays? *Nature Methods* **5,** 585–587. ISSN: 1548-7091 (July 2008).

15. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63. ISSN: 1471-0056 (Jan. 2009).

16. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29,** 644–652. ISSN: 1087-0156 (July 2011).

17. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12,** 87–98. ISSN: 1471-0056 (Feb. 2011).

18. Butler, J. M. & Butler, J. M. PCR Amplification: Capabilities and Cautions. *Advanced Topics in Forensic DNA Typing: Methodology,* 69–97 (Jan. 2012).

19. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628. ISSN: 1548-7091 (July 2008).

20. Thorne, T. Approximate inference of gene regulatory network models from RNA-Seq time series data. *BMC Bioinformatics* **19,** 127. ISSN: 1471-2105 (Apr. 2018).

21. Zhang, Z. H. *et al.* A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data. *PLoS ONE* **9** (ed Provero, P.) e103207. ISSN: 1932-6203 (Aug. 2014).

22. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harbor Protocols* **2015,** pdb.top084970. ISSN: 1940-3402 (Nov. 2015).

23. Kratz, A. & Carninci, P. The devil in the details of RNA-seq. *Nature Biotechnology* **32,** 882–884. ISSN: 1087-0156 (Sept. 2014).

24. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics* **33,** 155–168. ISSN: 01689525 (Feb. 2017).

25. Levitin, H. M., Yuan, J. & Sims, P. A. Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. *Trends in Cancer* **4,** 264–268. ISSN: 24058033 (Apr. 2018).

26. Megger, D. A., Bracht, T., Meyer, H. E. & Sitek, B. Label-free quantification in clinical proteomics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1834,** 1581–1590. ISSN: 15709639 (Aug. 2013).

27. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry* **389,** 1017–1031. ISSN: 1618-2642 (Sept. 2007).

28. Chahrour, O., Cobice, D. & Malone, J. Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *Journal of Pharmaceutical and Biomedical Analysis* **113,** 2–20. ISSN: 07317085 (Sept. 2015).

29. James, G., Witten, D., Hastie, T. & Tibshirani, R. in *An Introduction to Statistical Learning* (eds James, G., Witten, D., Hastie, T. & Robert, T.) 59–126 (Springer New York, New York, NY, 2013). ISBN: 978-1-4614-7138-7. doi:10.1007/978-1-4614-7138-7. <http://link.springer.com/10.1007/978-1-4614-7138-7>.

30. Krzywinski, M. & Altman, N. Power and sample size. *Nature Methods* **10,** 1139–1140. ISSN: 1548-7091 (Dec. 2013).

31. COX, D. R. Some remarks on overdispersion. *Biometrika* **70,** 269–274. ISSN: 0006-3444 (Apr. 1983).

32. Dean, C. B. Testing for Overdispersion in Poisson and Binomial Regression Models. *Journal of the American Statistical Association* **87,** 451. ISSN: 01621459 (June 1992).

33. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14,** 232–243. ISSN: 1465-4644 (Apr. 2013).

34. LEE, J.-H., HAN, G., FULP, W. J. & GIULIANO, A. R. Analysis of overdispersed count data: application to the Human Papillomavirus Infection in Men (HIM) Study. *Epidemiology and Infection* **140,** 1087–1094. ISSN: 0950-2688 (June 2012).

35. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning* (eds Hastie, T., Friedman, J. & Tibshirani, R.) 2nd ed., 261–267 (Springer New York, New York, NY, 2009). ISBN: 978-0-387-84857-0. doi:10.1007/978-0-387-84858-7. <https://www.springer.com/gp/book/9780387848570>.

36. McCullagh, P. & Nelder, J. A. *Generalized Linear Models, Second Edition (Monographs on Statistics and Applied Probability)* in (Chapman and Hall, 1989). ISBN: 978-1-4503-3779-3.

37. Elo, L., Filen, S., Lahesmaa, R. & Aittokallio, T. Reproducibility-Optimized Test Statistic for Ranking Genes in Microarray Studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5,** 423–431. ISSN: 1545-5963 (July 2008).

38. Suomi, T., Seyednasrollah, F., Jaakkola, M. K., Faux, T. & Elo, L. L. ROTS: An R package for reproducibility-optimized statistical testing. *PLOS Computational Biology* **13** (ed Poisot, T.) e1005562. ISSN: 1553-7358 (May 2017).

39. Motakis, E. S., Nason, G. P., Fryzlewicz, P. & Rutter, G. A. Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. *Bioinformatics* **22,** 2547–2553. ISSN: 1367-4803 (Oct. 2006).

40. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18,** 1509–1517. ISSN: 1088-9051 (July 2008).

41. Zheng, W., Chung, L. M. & Zhao, H. Bias detection and correction in RNA-Sequencing data. *BMC bioinformatics* **12,** 290. ISSN: 1471-2105 (July 2011).

42. Raabe, C. A., Tang, T.-H., Brosius, J. & Rozhdestvensky, T. S. Biases in small RNA deep sequencing data. *Nucleic Acids Research* **42,** 1414–1426. ISSN: 1362-4962 (Feb. 2014).

43. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26,** 493–500. ISSN: 1460-2059 (Feb. 2010).

44. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11,** R25. ISSN: 1465-6906 (2010).

45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550. ISSN: 1474-760X (Dec. 2014).

46. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18,** S96–S104. ISSN: 1367-4803 (July 2002).

47. Lin, S. M., Du, P., Huber, W. & Kibbe, W. A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Research* **36,** e11–e11. ISSN: 1362-4962 (Feb. 2008).

48. Bottomly, D. *et al.* Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *PLoS ONE* **6** (ed Zhuang, X.) e17820. ISSN: 1932-6203 (Mar. 2011).

49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57,** 289–300. ISSN: 00359246 (1995).

50. Ting, L. *et al.* Normalization and Statistical Analysis of Quantitative Proteomics Data Generated by Metabolic Labeling. *Molecular & Cellular Proteomics* **8,** 2227–2242. ISSN: 1535-9476 (Oct. 2009).

51. Karpievitch, Y. V., Dabney, A. R. & Smith, R. D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **13,** S5. ISSN: 1471-2105 (Nov. 2012).

52. Välikangas, T., Suomi, T. & Elo, L. L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics* **19,** bbw095. ISSN: 1467-5463 (Oct. 2016).

53. Chawade, A., Alexandersson, E. & Levander, F. Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. *Journal of Proteome Research* **13,** 3114–3120. ISSN: 1535-3893 (June 2014).

54. Fang, Z. & Cui, X. Design and validation issues in RNA-seq experiments. *Briefings in Bioinformatics* **12,** 280–287. ISSN: 1467-5463 (May 2011).

55. Su, Z. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology* **32,** 903–914. ISSN: 1087-0156 (Sept. 2014).

56. Kuharev, J., Navarro, P., Distler, U., Jahn, O. & Tenzer, S. In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *Proteomics* **15,** 3140–3151. ISSN: 16159853 (Sept. 2015).

57. Wu, H., Wang, C. & Wu, Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* **31,** 233–241. ISSN: 1460-2059 (Jan. 2015).

58. Zhao, S., Li, C.-I., Guo, Y., Sheng, Q. & Shyr, Y. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. *BMC Bioinformatics* **19,** 191. ISSN: 1471-2105 (Dec. 2018).

59. Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. & Kocher, J.-P. Calculating Sample Size Estimates for RNA Sequencing Data. *Journal of Computational Biology* **20,** 970–978. ISSN: 1066-5277 (Dec. 2013).

60. Bi, R. & Liu, P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics* **17,** 146. ISSN: 14712105 (Dec. 2016).