

University of Louisville

## ThinkIR: The University of Louisville's Institutional Repository

---

Electronic Theses and Dissertations

---

8-2019

### Designing and sample size calculation in presence of heterogeneity in biological studies involving high-throughput data.

Sudhir Srivastava  
*University of Louisville*

Follow this and additional works at: <https://ir.library.louisville.edu/etd>



Part of the [Bioinformatics Commons](#), [Computer Sciences Commons](#), and the [Statistics and Probability Commons](#)

---

#### Recommended Citation

Srivastava, Sudhir, "Designing and sample size calculation in presence of heterogeneity in biological studies involving high-throughput data." (2019). *Electronic Theses and Dissertations*. Paper 3261. <https://doi.org/10.18297/etd/3261>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact [thinkir@louisville.edu](mailto:thinkir@louisville.edu).

DESIGNING AND SAMPLE SIZE CALCULATION IN PRESENCE OF  
HETEROGENEITY IN BIOLOGICAL STUDIES INVOLVING  
HIGH-THROUGHPUT DATA

By

Sudhir Srivastava

B.Sc. (Agriculture), Banaras Hindu University, 2008  
M.Sc. (Agricultural Statistics), Indian Agricultural Research Institute, 2010

A Dissertation

Submitted to the Graduate School  
of the University of Louisville  
in Partial Fulfillment of the Requirements  
for the Degree of

Doctor of Philosophy

in Interdisciplinary Studies: Bioinformatics

Graduate School  
University of Louisville  
Louisville, Kentucky

August 2019

Copyright 2019 by Sudhir Srivastava

All rights reserved



DESIGNING AND SAMPLE SIZE CALCULATION IN PRESENCE OF  
HETEROGENEITY IN BIOLOGICAL STUDIES INVOLVING  
HIGH-THROUGHPUT DATA

By

Sudhir Srivastava

B.Sc. (Agriculture), Banaras Hindu University, 2008  
M.Sc. (Agricultural Statistics), Indian Agricultural Research Institute, 2010

A Dissertation Approved on

July 17, 2019

by the following Dissertation Committee:

---

Shesh N. Rai, Ph.D., Principal Advisor

---

Eric C. Rouchka, D.Sc.

---

Ted Kalbfleisch, Ph.D.

---

Michael L. Merchant, Ph.D.

---

Riten Mitra, Ph.D.

## DEDICATION

This dissertation is dedicated

to my parents

Mr. Sheo Pratap Lal and Mrs. Sidhi Devi

And to my wife

Arpita Srivastava

for their love, endless support and encouragement

in all my endeavors.

## ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Shesh N. Rai for his excellent guidance and continuous support throughout my doctoral journey. Dr. Rai provided me unique opportunities and created foundation of the research. Dr. Rai encouraged independent thinking and further exploration of the research problems. Dr. Rai is a true mentor who has included me in several projects with colleagues from the University of Louisville, USA. This further strengthens my thinking skills and research capabilities. I would like to thank my diverse group of mentors, Dr. Eric C. Rouchka, Dr. Michael Merchant, Dr. Ted Kalbfleisch and Dr. Riten Mitra for their continuous guidance, motivation and support. Each of them provided significant contributions in various aspects of my dissertation work.

I would like to thank the Indian Council of Agricultural Research (ICAR), India for giving me opportunity to pursue Ph.D. through ICAR-International fellowship. I would like to thank the University of Louisville, USA for providing me a great environment and facilities for conducting my Ph.D. research. It is my pleasure to do Ph.D. at the University of Louisville, USA.

I would like to thank my lab members at the University of Louisville, USA and Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute (IASRI), India for giving me support and encouragement. I

would like to thank Dr. Anil Rai, Dr. K. K. Chaturvedi and Dr. D. C. Mishra for their continuous support and motivation during my Ph.D. research work. I would also like to give thanks to my cohort mates for making the course work more enjoyable.

Last not but the least, I would like to thank my parents for the great education and support they provided me during my entire life. I would also like to express thanks to my wife, Arpita, for keeping patience and understanding me. She gave me continuous support and motivation during my Ph.D. work. Finally, I would like to thank my family members – Anil, Priti, Rekha, Raksha and Priya – for giving me moral and loving support throughout my life.



## ABSTRACT

### DESIGNING AND SAMPLE SIZE CALCULATION IN PRESENCE OF HETEROGENEITY IN BIOLOGICAL STUDIES INVOLVING HIGH-THROUGHPUT DATA

Sudhir Srivastava

July 17, 2019

The designing and determination of sample size are important for conducting high-throughput biological experiments such as proteomics experiments and RNA-Seq expression studies, thus leading to better understanding of complex mechanisms underlying various biological processes. The variations in the biological data or technical approaches to data collection lead to heterogeneity for the samples under study. We critically worked on the issues of technical and biological heterogeneity.

The quantitative measurements based on liquid chromatography (LC) coupled with mass spectrometry (MS) often suffer from the problem of missing values (MVs) and data heterogeneity. We considered a proteomics data set generated from human kidney biopsy material to investigate the technical effects of sample preparation and the quantitative MS. We studied the effect of tissue storage methods (TSMs) and tissue extraction methods (TEMs) on data analysis. There are two TSMs: frozen (FR) and FFPE (formalin-fixed paraffin embedded); and three TEMs: MAX, TX followed by MAX and SDS followed by MAX. We

assessed the impact of different strategies to analyze the data while considering heterogeneity and MVs. We found that the FFPE is better than that of FR for tissue storage. We also found that the one-step TEM (MAX) is better than those of two-steps TEMs. Furthermore, we found the imputation method is a better approach than excluding the proteins with MVs or using unbalanced design.

We introduce a web application, PWST (Proteomics Workflow Standardization Tool) to standardize the proteomics workflow. The tool will be helpful in deciding the most suitable choice for each step and studying the variability associated with technical steps as well as the effects of continuous variables. We have used the special cases of general linear model - ANCOVA and ANOVA with fixed effects to study the effects due to various sources of variability. We introduce an interactive tool, "SATP: Statistical Analysis Tool for Proteomics", for analyzing proteomics expression data that is scalable to large clinical proteomic studies. The user can perform differential expression analysis of proteomics data either at the protein or peptide level using multiple approaches. We have developed statistical approaches for calculating sample size for proteomics experiments under allocation and cost constraints. We have developed R programs and a shiny app "SSCP: Sample Size Calculator for Proteomics Experiment" for computing sample sizes.

We have proposed statistical approaches for calculating sample size for RNA-Seq experiments considering allocation and cost. We have developed R programs and shiny apps to calculate sample size for conducting RNA-Seq experiments.

## TABLE OF CONTENTS

	PAGE
DEDICATION.....	iii
ACKNOWLEDGEMENTS.....	iv
ABSTRACT.....	vi
LIST OF TABLES.....	xi
LIST OF FIGURES.....	xiv
CHAPTER 1	
INTRODUCTION.....	1
Design and sample size calculation for high-throughput experiments.....	1
Contributions.....	2
Layout of the dissertation.....	4
CHAPTER 2	
PRELIMINARIES: SAMPLE SIZE CALCULATION, EXPERIMENTAL DESIGN AND BIOLOGICAL EXPERIMENTS.....	5
Sample size calculation.....	5
Experimental design and heterogeneity issues.....	6
Proteomics experiments.....	9
RNA-Sequencing experiments.....	10
CHAPTER 3	
STANDARDIZING PROTEOMICS WORKFLOW FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY: TECHNICAL AND STATISTICAL CONSIDERATIONS.....	13

Introduction.....	13
Methods.....	16
Results and discussion.....	23
Conclusion.....	30
CHAPTER 4	
INTERACTIVE WEB TOOL FOR STANDARDIZING PROTEOMICS	
WORKFLOW FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY	
DATA.....	32
Introduction.....	32
Methods.....	34
Demonstration and discussion.....	37
Conclusion.....	49
CHAPTER 5	
INTERACTIVE TOOL FOR STATISTICAL ANALYSIS OF LABEL-FREE LC-MS	
PROTEOMICS DATA CONSIDERING MISSING VALUES AND	
HETEROGENEITY.....	50
Introduction.....	50
Methods.....	51
Demonstration and results.....	57
Discussion.....	72
Conclusion.....	74
CHAPTER 6	
SAMPLE SIZE ESTIMATION FOR HETEROGENEOUS PROTEOMICS	
EXPERIMENTS USING STATISTICAL AND COMPUTATIONAL	
APPROACHES.....	75
Introduction.....	75
Sample size calculation for detecting differentially expressed features	
between two classes.....	76
Sample size calculation using pilot data.....	87

Web app for calculating sample size using pilot data.....	90
Impact of technical components on the study design and sample size estimation of LC-MS proteomics workflow.....	95
CHAPTER 7	
SAMPLE SIZE CALCULATION FOR RNA-SEQ EXPERIMENTS CONSIDERING HETEROGENEITY.....	98
Introduction.....	98
Modeling the count data in RNA-Seq experiments.....	99
Estimation of parameters based on negative binomial distribution.....	102
Power and sample size calculation based on negative binomial distribution.....	107
A shiny app for sample size estimation based on Poisson-log normal distribution.....	116
CHAPTER 8	
DISCUSSION AND CONCLUSION.....	120
REFERENCES.....	122
APPENDIX A: ACRONYMS USED.....	136
CURRICULUM VITAE.....	138

## LIST OF TABLES

TABLE	PAGE
Table 2.1. An example of a biological experiment showing response outcome for multiple features in samples across different conditions.....	8
Table 3.1. Table showing different groups under study.....	20
Table 3.2. Summary of number of proteins and missing values in different groups.....	21
Table 3.3. Summary of CV (in %) using 9 statistical approaches among TSM and TEM.....	24
Table 3.4. Summary of CV (in %) using 9 statistical approaches among six groups of TSM×TEM.....	25
Table 3.5. Summary of the contribution of % SS due to TSM, TEM and TSM×TEM.....	26
Table 3.6. The summary of proportion of proteins showing effects due to the variables: TSM, TEM and TSM×TEM.....	27
Table 5.1. Comparison among the tools: SATP, RepExplore and MSqRob.....	73
Table 6.1. The outcomes of testing null hypothesis vs. alternative hypothesis.....	77
Table 6.2. Sample sizes ( $N_1$ , $N_2$ ) and power computed when $r$ is fixed with parameters $\mu_d = 4$ , $\alpha = 0.05$ , $1 - \beta = 0.90$ for different combinations of $r$ and $\sigma_1: \sigma_2$ .....	81
Table 6.3. A hypothetical example of various costs involved in quantitative proteomics experiment.....	82

Table 6.4. Sample sizes ( $N_1, N_2$ ), total number of samples ( $N_1 + N_2$ ) and the power obtained for fixed cost $C$ with parameters $\mu_d = 4, \alpha = 0.05, C_1: C_2 = 1: 1/3$ and different values of $\sigma_1: \sigma_2$ using our method and the original method.....	84
Table 6.5. The possible outcomes of testing multiple null hypotheses.....	85
Table 6.6. Summary of estimated standard deviations for two groups.....	88
Table 6.7. Sample sizes ( $N_1, N_2$ ), total number of samples ( $N_1 + N_2$ ) and the exact power obtained with $\log_2$ FC 1 for fixed sample size ratio 1 and 2.....	89
Table 6.8. Sample sizes ( $N_1, N_2$ ), total number of samples ( $N_1 + N_2$ ) and the maximum power obtained with $\log_2$ FC 1 for fixed cost (5000, 10000 and 15000).....	89
Table 6.9. Sample sizes ( $N_1, N_2$ ), total number of samples ( $N_1 + N_2$ ) and exact power obtained for a minimum cost of experiment with $\log_2$ FC 1.....	90
Table 6.10. Computed sample sizes for different technical approaches.....	97
Table 7.1. Sample sizes ( $N_1, N_2$ ) and power computed with parameters $\mu_1 = \mu_2 = 20, \phi_1 = \phi_2 = 0.4, \alpha = 0.05, 1 - \beta = 0.90$ for different combinations of $r$ and $\Delta$ .....	112
Table 7.2. A hypothetical example of various costs involved in RNA-Seq experiment.....	113
Table 7.3. Sample sizes ( $N_1, N_2$ ) and the power obtained with parameters $\mu_1 = \mu_2 = 20, \phi_1 = \phi_2 = 0.4, \alpha = 0.05, C_1 = 1000, C_2 = 1100$ for different values of fixed cost $C$ and fold change. ....	114

Table 7.4. Sample sizes ( $N_1, N_2$ ) and power computed with parameters $m_1 = 100, m = 10000, \mu_1 = \mu_2 = 20, \phi_1 = \phi_2 = 0.4, \alpha = 0.05, 1 - \beta = 0.90$ for different combinations of $r$ and $\Delta$ .....	115
Table 7.5. Sample sizes ( $N_1, N_2$ ) and the power obtained with parameters $m_1 = 100, m = 10000, \mu_1 = \mu_2 = 20, \phi_1 = \phi_2 = 0.4, \alpha = 0.05, C_1 = 1000, C_2 = 1100$ for different values of fixed cost $C$ and fold change.....	116
Table 7.6. The inputs provided for sample size and power calculation.....	119



## LIST OF FIGURES

FIGURE	PAGE
Figure 3.1. Flowchart of the proteomics experiment.....	19
Figure 3.2. Plot of CV (in %) versus the proteins with increasing order of p-values for TSM (FR and FFPE) .....	29
Figure 3.3. Plot of CV (in %) versus the proteins with increasing order of p-values for TSM (MAX, TX.MAX and SDS.MAX).....	29
Figure 4.1. Webpage of the tool “PWST” .....	38
Figure 4.2. A portion of proteomics expression data.....	39
Figure 4.3. Upload the proteomics expression data.....	39
Figure 4.4. Choose the feature type – “Protein” or “Peptide” .....	40
Figure 4.5. Choose the aggregation method (Mean/Median/Sum/Max).....	40
Figure 4.6. A portion of additional information of data.....	41
Figure 4.7. Upload the additional information of data.....	41
Figure 4.8. Selection of categorical variables.....	42
Figure 4.9. Selection of numeric variables.....	42
Figure 4.10. Selection of analysis method.....	42
Figure 4.11. Selection of normalization method.....	43
Figure 4.12. Specify the level of significance.....	43
Figure 4.13. Specify the adjustment method.....	43
Figure 4.14. Inputs selected.....	44
Figure 4.15. Box plot of preprocessed expression data.....	44

Figure 4.16. Density plot of preprocessed expression data.....	45
Figure 4.17. Interactive correlation heatmap of preprocessed expression data.....	45
Figure 4.18. Contribution of SS due to each variable, the p-values and the adjusted p-values corresponding to each variable for each protein.....	46
Figure 4.19. Summary of % SS contribution due to each variable.....	46
Figure 4.20. The box plot showing % contribution of SS due to each variable.....	47
Figure 4.21. The CV of different groups of each categorical variable for all the proteins.....	47
Figure 4.22. Summary of CV (in %) for all the proteins.....	48
Figure 4.23. Box plot showing CV under the various groups of categorical variables.....	48
Figure 4.24. Summary of significant proteins.....	48
Figure 5.1. Webpage of the tool “SATP” .....	58
Figure 5.2. A portion of proteomics expression data.....	59
Figure 5.3. Upload the proteomics expression data.....	59
Figure 5.4. Choose the type of feature– “Protein” or “Peptide” .....	60
Figure 5.5. Choose the aggregation method.....	60
Figure 5.6. Additional information of data.....	61
Figure 5.7. Upload the file with additional information of data.....	62
Figure 5.8. Selection of categorical fixed effects.....	62
Figure 5.9. Selection of continuous variables.....	63
Figure 5.10. Selection of random effects.....	63
Figure 5.11. Selection of categorical fixed effect of interest.....	64

Figure 5.12. Selection comparison of interest.....	64
Figure 5.13. Select analysis method.....	64
Figure 5.14. Selection of normalization method.....	65
Figure 5.15. Selection of statistical testing method.....	65
Figure 5.16. Specify the level of significance.....	66
Figure 5.17. Specify the desired log fold change.....	66
Figure 5.18. Specify the adjustment method.....	66
Figure 5.19. Inputs selected.....	67
Figure 5.20. Box plots of preprocessed expression data for different groups.....	67
Figure 5.21. Density plots of preprocessed expression data for different groups.....	68
Figure 5.22. Interactive correlation heatmap of preprocessed expression data.....	68
Figure 5.23. Interactive MDS plot.....	68
Figure 5.24. Summary of result.....	69
Figure 5.25. Result of differential expression analysis.....	70
Figure 5.26. Volcano plot without adjustment.....	70
Figure 5.27. Volcano plot with adjustment.....	71
Figure 6.1. Screenshot of the tool “SSCP” .....	91
Figure 6.2. Uploading the two input files and selecting feature type and class name under comparison.....	93
Figure 6.3. Specifying expected proportion of significant features, false discovery rate, average power and log <sub>2</sub> fold change.....	93
Figure 6.4. The input parameters under each method of sample size calculation.....	94

Figure 6.5. Example of inputs and output obtained for sample allocation with maximum power for a fixed cost.....	94
Figure 7.1. Shiny application to calculate sample size for RNA-Seq experiments using pilot data.....	117

## CHAPTER 1

### INTRODUCTION

#### **Design and sample size calculation for high-throughput experiments**

The designing and determination of sample size are important for conducting high-throughput biological experiments such as proteomics experiments and RNA-Seq expression studies, thus leading to better understanding of complex mechanisms underlying various biological processes. These experiments undergo various steps such as choosing appropriate experimental design, proper selection and collection of samples from various sources, choice of platform, data generation, data preprocessing, data analysis and interpretation. Various experiments are being conducted but there is lot of variation at each of these steps. There are various discrepancies observed with the result and conclusions obtained from these experiments. Sometimes these results cannot be reproduced, and this failure may be derived from one or more technical variables. So, we studied the data variability and developed statistical approaches for sample size determination for these experiments while taking various heterogeneity issues in account.

This research work has focused on the development of statistical methods for designing and sample size calculation covering a wide range of high throughput biological experiments such as proteomics and RNA-Seq

experiments. This is an innovative work that will be helpful to researchers/experimenters in the design of their study applicable to different areas such as proteomics and genomics. This will bring out clarity to the experimenters in conducting their study with a specific goal.

## **Contributions**

Identification, quantification and characterization of peptides and proteins in cells, are necessary to understand the molecular process governing the cell physiology. Liquid chromatography (LC) coupled with mass spectrometry (MS) is generally used for identifying and quantifying proteins and peptides in complex mixtures. With the introduction of high throughput technologies such as MS, proteomics data can be reliably generated from samples that can be further analyzed using various statistical approaches. Sometimes, variations in the biological data or technical approaches to data collection lead to heterogeneity for the samples under study. Furthermore, the proteomics data obtained from proteomics experiments have a lot of missing values (MVs) and are highly heterogeneous. We investigated the technical effects of sample preparation and the quantitative MS resulting in heterogeneity for low abundant protein quantification (Chapter 3). We developed statistical approaches and a web-application for standardizing proteomics experiment work flow (Chapters 3 and 4). We discussed and developed a shiny app for differential expression analysis of proteomics data using multiple statistical approaches in the presence of heterogeneity and MVs (Chapter 5). We devised various approaches for sample size calculation for conducting proteomics experiments under allocation and

budget constraints (Chapter 6). Furthermore, we developed shiny apps for estimating sample size for proteomics studies based on different constraints with and without using pilot data (Chapter 6). We studied the impact of technical variability (using data from Chapter 3) on the study design and sample size estimation in Chapter 6.

Next-generation sequencing (NGS) of mRNA (RNA-Seq) has become the standard for measuring gene expressions in biological experiments. The determination of sequencing depth, number of replicates and power calculation are important while designing an RNA-Seq experiment. Various methods exist for estimating sample size for differential expression analysis of RNA-Seq data under the assumption of different models. The RNA-seq experiments are complex in nature, and still there is requirement of advanced method to calculate sample size for differential expression analysis using RNA-Seq data. Therefore, we devised statistical approaches for designing and sample size calculation considering allocation and cost constraints required to carry out the RNA-Seq experiments under the assumptions of various models (Chapter 7).

We have implemented all the methods in R [1] and used various Bioconductor packages [2], applicable to these experiments. We have developed all the apps using “shiny” R package [3]. It will be easier for the experimenters to calculate the sample size required for conducting the experiments according to the budget. These programs can be used by the researchers for writing grants and conducting research projects, that will save resources in terms of cost and time.

## **Layout of the dissertation**

The layout of the dissertation is as follows:

Chapter 2 provides the basic background of various topics such as sample size calculation, heterogeneity issues, proteomics and RNA-Seq experiments.

Chapter 3 provides the technical and statistical considerations for standardizing proteomics workflow for LC-MS proteomics expression data [4].

Chapter 4 provides the interactive web tool for standardizing proteomics workflow for LC-MS Data [5].

Chapter 5 provides the various approaches for differential expression analysis of proteomics expression data and an interactive tool for statistical analysis of label-free LC-MS proteomics data considering MVs and heterogeneity.

Chapter 6 provides the sample size estimation methods for proteomics experiments under various constraints.

Chapter 7 provides the statistical methods of sample size calculation for RNA-Seq experiments considering allocation and cost.

Chapter 8 provides the discussion and conclusion.



## CHAPTER 2

### PRELIMINARIES: SAMPLE SIZE CALCULATION, EXPERIMENTAL DESIGN AND BIOLOGICAL EXPERIMENTS

#### **Sample size calculation**

Sample size calculation is an important process of choosing the number of replicates in a study with the goal to make inferences about a population from a sample. The sample size used in a study depends on various constraints such as data availability, budget, support facilities, time requirement, etc. The basic principles underlying the method of sample size calculation are the same, but these methods are not universal. So, the methods of sample size calculation depend on the type of experiment. In complicated studies, there may be several different sample sizes involved. For example, in an experiment where a study may be divided into different treatment groups/ conditions, there may be different sample sizes for each group/ condition.

Methods for sample size calculations begin with an understanding of the type of data and its distribution. In most of the experiments, the data can be broadly divided into quantitative (numerical) and categorical (qualitative) data. Let us consider there are two groups for comparison. Let the number of samples in each group are  $N_1$  and  $N_2$ , respectively. In general, the following factors must be known or estimated to calculate sample size [6-8]:

- (i) The desired fold change: It is the difference between mean responses in the two groups, i.e., the difference between  $\mu_1$  and  $\mu_2$  for quantitative data.
- (ii) The population standard deviations (SDs): It is the variability or spread associated with quantitative data. We require either common population SD ( $\sigma$ ) for the two groups or SDs ( $\sigma_1$  and  $\sigma_2$ ) for each group. The population SD of the variable of interest can be estimated from a pilot study or data obtained from an experiment or from the scientific literature.
- (iii) The level of significance: The probability that a positive finding is due to chance is denoted as  $\alpha$ , the significance level. It is usually chosen to be 0.05 or 0.01.
- (iv) The desired power of the experiment: The power of an experiment is the probability that the effect will be detected. It is usually set to 0.8 or 0.9.

### **Experimental design and heterogeneity issues**

*Experimental design:* The purpose of experimental design is to plan experiment in an effective way so that it can answer the biological question under consideration. The major points to be considered while documenting experimental plan are as follows:

- (i) Biological aspects: Any biological experimental plan starts with a biological question or hypothesis generating/ hypothesis testing. The experimenter might have some prior knowledge of the question under study before conducting the experiments, e.g., expression levels of some known genes, proteins, etc., that may be helpful. Later, the question arises about the samples such as:

- whether enough samples are available for experiment;
- there may be samples available before hand in some situations;
- availability of enough RNA, DNA or proteins from samples;
- whether pooling of samples is required or not;
- biopsies collected from same part of tissue or other tissues;
- whether the cell type is expressing the feature such as gene of interest;
- number of replicates required;
- effect size, etc.

(ii) Technical aspects: These include the choice of platform and avoiding systematic errors. If the experiment has systematic errors, then the result obtained for comparative analysis will be biased, irrespective of the precision of measurement and the number of experimental units.

For the two above aspects, biological replicates are used to answer biological questions and technical replicates are required to answer technical questions.

(iii) Economic aspects: These include cost of experiment and analysis, budget available, time required to complete the experiment and its analysis, whether pilot study is required or not, etc.

*Example of a biological experiment:* Let us consider an experimental study in which there are  $I$  conditions/groups denoted by  $G_i (i = 1, 2, \dots, I)$  and there are  $N_i$  individuals/samples denoted by  $S_{i,j} (j = 1, 2, \dots, N_i)$  corresponding to group  $G_i$  (Please see Table 2.1). Therefore, there is a total  $N = \sum_{i=1}^I N_i$  samples in the experimental study. Now suppose, there are  $K$  features (e.g., transcripts, genes, peptides, proteins, etc.) under study (e.g., testing for differential expression,

testing association with trait, etc.) denoted by  $F_k$  ( $k = 1, 2, \dots, K$ ). Let  $y_{i,j,k}$  is a response outcome corresponding to sample  $S_{i,j}$  of condition  $G_i$  for feature  $F_k$  as shown below in Table 2.1.

**Table 2.1.** An example of a biological experiment showing response outcome for multiple features in samples across different conditions

	$G_1$					...	$G_i$					...	$G_I$				
	$S_{1,1}$	...	$S_{1,j}$	...	$S_{1,N_1}$		$S_{i,1}$	...	$S_{i,j}$	...	$S_{i,N_i}$		$S_{I,1}$	...	$S_{I,j}$	...	$S_{I,N_I}$
$F_1$	$y_{1,1,1}$	...	$y_{1,j,1}$	...	$y_{1,N_1,1}$		$y_{i,1,1}$	...	$y_{i,j,1}$	...	$y_{i,N_i,1}$		$y_{I,1,1}$	...	$y_{I,j,1}$	...	$y_{I,N_I,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$F_k$	$y_{1,1,k}$	...	$y_{1,j,k}$	...	$y_{1,N_1,k}$		$y_{i,1,k}$	...	$y_{i,j,k}$	...	$y_{i,N_i,k}$		$y_{I,1,k}$	...	$y_{I,j,k}$	...	$y_{I,N_I,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$F_K$	$y_{1,1,K}$	...	$y_{1,j,K}$	...	$y_{1,N_1,K}$		$y_{i,1,K}$	...	$y_{i,j,K}$	...	$y_{i,N_i,K}$		$y_{I,1,K}$	...	$y_{I,j,K}$	...	$y_{I,N_I,K}$

There may be a variety of purposes for the experiments such as detection of differentially expressed features, detecting association of quantitative or qualitative trait associated features, etc.

*Heterogeneity:* A heterogeneous sample or population means that every observed data has different value for the corresponding characteristic of interest. For example, in gene expression studies, transcriptional variation is characterized with respect to measured variables of interest such as different conditions, different treatments, different points of time, etc. The major sources of variations in gene expression studies are due to technical, genetic, demographic and environmental factors [9]. There may be various factors responsible for influencing expression in any feature (genes, proteins, etc.), some of which cannot be measured, or some may be unknown.

*Example of heterogeneity in gene expression data:* Expression in a gene can be influenced by interaction with other genes, presence of external stimulus or signal, etc. A gene may be highly expressed in one condition and it may be less expressed in other condition or even sometimes it may not be expressed. In RNA-Seq data, the biological source of RNA are tissue samples which may be highly heterogeneous. The accuracy of the transcript quantification will depend on the purity of samples [10]. Therefore, failure to detect such heterogeneity will lead to false data interpretation and the result will be irreproducible.

### **Proteomics experiments**

Proteins are important biological macromolecules performing a wide variety of functions. The term “proteome” is defined as the entire set of proteins produced or modified by a living organism [11, 12]. Proteomics generally refers to the large-scale quantitative/ qualitative study of proteins for a given cell type. Now it has emerged as a powerful tool across various fields such as biomedicine mainly applied to diseases, agriculture and animal sciences [13-16]. It is becoming increasingly important for the study of different aspects of plant functions, such as identification of candidate proteins involved in the defensive response of plants to insects, effect of global climate changes on crop production, etc. [17-19]. The practical application of proteomics includes expression proteomics, structural proteomics, biomarkers, interaction proteomics, protein networks, etc.

Proteomic expression data are generated by using high throughput technologies usually involving a mass spectrometer [20-24]. Liquid chromatography (LC) coupled with mass spectrometry (MS) is used in

proteomics as a method for identification and quantification of proteins and peptides in complex mixtures. The intensity of the resulting LC-MS features is used for relative quantification of peptides and proteins. LC-MS/MS (tandem-MS) experiments are used to derive the sequence of peptides and deduce the protein underlying a subset of the features. Various software tools have been developed to extract and quantify LC-MS features from the acquired spectra, annotate the features with sequence identity, and align the features across runs [25-32]. Samples of mixtures can be analyzed using modern LC-MS/MS systems which are capable of identifying and quantifying thousands of peptides simultaneously. For most of these types of experiments, the raw intensity data is summarized for each of the replicates for each feature. Here, the feature can be either at protein level or peptide level. Further, the data obtained from LC-MS experiments can be used for differential expression analysis between sample groups (e.g., testing peptides/proteins for differential abundance between subjects in a case-control study), or to analyze protein abundance of individual biological subjects (e.g., unsupervised clustering or supervised classification of individuals, based on their quantitative protein profiles).

### **RNA-Sequencing experiments**

RNA-Sequencing, also called whole transcriptome shotgun sequencing, uses NGS technology to reveal the presence and quantity of RNA in a biological sample at a given moment. It has been the most productive research area from the computational and statistical point of view that can provide an insight into the roles of genes at transcriptomic level. It allows transcript quantification and

differential gene expression analysis, including identification of alternative splicing events and post-transcriptional RNA editing events. Several machines/protocols are available for generating RNA-Seq data, namely, Illumina (MiSeq, NextSeq, HiSeq, NovaSeq), Ion Torrent (Proton, Personal Genome Machine), etc. The basic steps for summarizing a typical RNA-Seq experiment are as follows:

- Purified RNA is converted to cDNA, sequencing library is prepared, and sequencing is done on an NGS platform.
- Millions of short read sequences are generated from one end (single-end) or both ends (paired-end) of the cDNA fragments.
- These sequences are aligned to a reference genome.
- The number of reads mapped to known features are recorded and summarized in a table. The features can be either genes, transcripts (alternative transcripts), allele specific expression or exon level expression. For example, if there are  $F$  features and  $N$  samples, then a table of read counts is a  $F \times N$  matrix of non-negative integers

In RNA-Seq experiments, the samples are sequenced and resulting reads are aligned with a reference genome. The numbers of reads mapped to each of the reference gene are calculated. Then, normalization techniques are used to account for the within library and between library variability. For differential expression analysis in RNA-Seq data, the number of reads mapped to a reference genome (read counts) are generally modelled by assuming Poisson distribution or negative binomial distribution.

The challenges associated with application of RNA-Seq experiments are the problems in library construction, bioinformatics problems (storage, retrieval and processing of large data sets, mapping and assembly problem), sequence/transcriptome coverage versus cost, transcriptomic analysis (mapping gene for identifying introns and exon boundaries as well as discovery of novel transcribed genes, detection of splicing events, quantification of transcriptome/RNA expression levels to study gene expression in complex experiments) [33].



## CHAPTER 3

# STANDARDIZING PROTEOMICS WORKFLOW FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY: TECHNICAL AND STATISTICAL CONSIDERATIONS

### **Introduction**

Proteins are important biological macromolecules performing a wide variety of functions. The proteome can be defined as the entire set of proteins translated and/or modified within a living organism [11, 12]. Proteomics more generally refers to large-scale LC-MS based discovery studies designed to address both quantitative and qualitative aspects of the proteome in question. Now proteomics has emerged as a powerful tool across various fields such as biomedicine mainly applied to diseases, agriculture and animal sciences [13-16, 34-38]. The practical application of proteomics includes expression proteomics, structural proteomics, biomarker discovery, interaction proteomics, protein networks, etc. [39, 40]. Here, we are dealing with proteomic expression data that are generated by using high throughput technologies usually involving MS [20-24, 41].

LC-MS is used in proteomics as a method for identification and quantification of peptides and proteins in complex mixtures [42, 43]. There are two basic proteomics approaches, namely bottom-up and top-down [38, 44]. The most common proteomics approach is the bottom-up in which proteins in a

sample are enzymatically digested into peptides and subjected to chromatographic separation, ionization and mass analysis. In the top-down approach, intact proteins are introduced into MS where they are subjected to fragmentation. Further, the quantification of peptides/proteins may be either label-free or labelled (metabolic, enzymatic, or chemical) to detect differences in protein abundances among different conditions [45-48]. In label-free quantification, MS ion intensity (peak area) and spectral counting of features are the major approaches. Conversely, top-down proteomics addresses the study of intact proteins and consequently is most often used to address purified or partially purified proteins [49]. Here, we are dealing with the bottom-up approach in which peak area values have been used in label-free quantification of proteins.

Various approaches exist for proteomics data analysis in which the first step is to summarize the intensities of all features using a quantitative summary followed by some transformation such as log transformation to approximate it to normal distribution. However, each of these methods has several drawbacks which can be studied by examining the statistical properties of these methods [50-52]. When a data set contains an equal number of subjects in each group, and when features have no missing observations, the data set is called balanced. It is not always the condition; sometimes the data can be unbalanced, having an unequal number of subjects, or missing observations, or both. Missing values (MVs) in proteomics data can occur due to biological and/or technical issues. These are of three types: (i) missing completely at random (MCAR) in which MVs are independent of both unobserved and observed data; (ii) missing at random

(MAR) if conditional on the observed data, the MVs are independent of the missing measurements; and (iii) missing not at random (MNAR) when data is neither MCAR nor MAR [53]. The data with missing observations can be analyzed either by excluding the features having missing observations, by using statistical methods that can handle unbalanced data, or by using imputation methods. If the features having missing observations are excluded, then there is loss of information from the experiment. Therefore, the use of methods that can handle MVs, such as imputation methods, are generally preferred. However, the use of imputation methods may lead to wrong interpretation and still these methods are questionable in statistical terms [54, 55].

The data set usually consists of biological replicates only or both biological and technical replicates. Biological variability arises from genetic and environmental factors; it is intrinsic to all organisms. The technical approaches include sample collection and storage, sample preparation, extraction, LC separation and MS detection [43]. Sometimes, variations in the biological data or technical approaches to data collection lead to heterogeneity for the samples under study [56, 57]. We performed analysis of laser capture microdissection (LCMD)-LCMS high-resolution proteomics dataset using multifactor ANOVA model. We studied the variability in the data based on different tissue storage methods (TSMs) and tissue extraction methods (TEMs). We estimated the contribution of various sources of variation to the overall variability. The study of data variability was done using various analysis methods and transformation and/or normalization techniques.

We investigated the technical effects of sample preparation and the quantitative MS resulting in heterogeneity for low abundant protein quantification. This will improve the biomarker discovery studies utilizing limited bioreposited tissue resources.

## **Methods**

### **Proteomics experiment**

Data for the methods used in the collection, extraction, and proteomic analysis have previously been published under Hobeika L., et al. [58]. Individual data files for MS data (.RAW), peak lists (.mgf), and compressed search results (.mzIdentML) files can be downloaded from the MassIVE data repository (<http://massive.ucsd.edu/>; MassIVE ID: MSV000079914) and ProteomeXchange data repository [59] (<http://www.proteomexchange.org/>; ID: PXD004601). For consideration of variability of the feature detection and MVs, the abbreviated methods for these studies are provided below.

**Tissue collection:** Frozen (FR) and formalin-fixed paraffin embedded (FFPE) tissue from the same human kidney unsuitable for transplant were cut into 10  $\mu$ m sections on polyethylene terephthalate membrane frame slides, stained with Mayer's hematoxylin and glomerular tissue compartments isolated using a Leica LMD6500 Laser Microdissection System.

**Protein extraction:** Experiments were conducted to compare a single tissue solubilization step using an acid labile surfactant to approaches for tissue decellularization. The single step method used the acid-labile surfactant Protease MAX surfactant with heating (MAX). Two tissue decellularization methods

incorporated sequential decellularization with solubilization of the residual pellet with MAX. First tissue decellularization approach used 0.4% SDS + HALT protease/phosphatase inhibitor cocktail (Thermo Fisher) followed by solubilization of residual extracellular matrix (ECM) pellet using MAX (SDS.MAX). Second tissue decellularization approach used sequential decellularization with 25mM NH<sub>4</sub>OH/ 0.5% TritonX-100 (TX) followed by solubilization of residual ECM pellet using MAX (TX.MAX).

**Liquid Chromatography:** Peptide separation was achieved using a Dionex Acclaim PepMap 100 75µm x 2cm, nanoViper (C18, 3µm, 100Å) trap, and a Dionex Acclaim PepMap RSLC 50µm x 15cm, nanoViper (C18, 2µm, 100Å) separating column. An EASY n-LC (Thermo) UHPLC system was used to resolve peptide separation using a 140min linear gradient from 2% v/v acetonitrile / 0.1% v/v formic acid to 40% v/v acetonitrile / 0.1% v/v formic acid. Peptides were introduced into the Orbitrap ELITE MS using a 40mm stainless steel emitter (Thermo P/N ES542) and a Nanospray Flex source (Thermo) was used to position the end of the emitter near the ion transfer capillary of the mass spectrometer.

**Mass Spectrometry Data Acquisition:** MS data was collected using an Nth Order Double Play with or Electron-transfer dissociation (ETD) Decision Tree method created in Xcalibur v2.2. Scan event one of the method obtained a Fourier transform MS MS1 scan (normal mass range; 60,000 resolution, full scan type, positive polarity, profile data type) for the range 300-2000m/z. Scan event two obtained ion trap MS MS2 scans (normal mass range, rapid scan rate,

centroid data type) on up to twenty peaks that had a minimum signal threshold of 5,000 counts from scan event one. A decision tree was used to determine whether collision-induced dissociation (CID) ETD activation was used. An ETD scan was triggered if any of the following held: an ion had charge state 3 and  $m/z$  less than 650, an ion had charge state 4 and  $m/z$  less than 900, an ion had charge state 5 and  $m/z$  less than 950, or an ion had charge state greater than 5; a CID scan was triggered in all other cases. The lock mass option was enabled (0% lock mass abundance) using the 371.101236 $m/z$  polysiloxane peak as an internal calibrant.

***Data Analysis with Proteome Discoverer v1.4.1.14 and Scaffold Q+S v4.4.3:***

Proteome Discoverer v1.4.1.114 was used to analyze the data collected by the mass spectrometer. The database used in Mascot v2.5.1 and SequestHT searches was a 4/7/2015 version of the UniprotKB *Homo sapiens* reference proteome canonical and isoform sequences with the 1/1/2012 version of the common Repository of Adventitious Proteins (cRAP) database (thegpm.org) appended to it (the cRAP database contains common contaminant proteins observed in MS experiments). To estimate the false discovery rate (FDR), a Target Decoy Peptide-Spectrum Match Validator node was included in the Proteome Discoverer workflow.

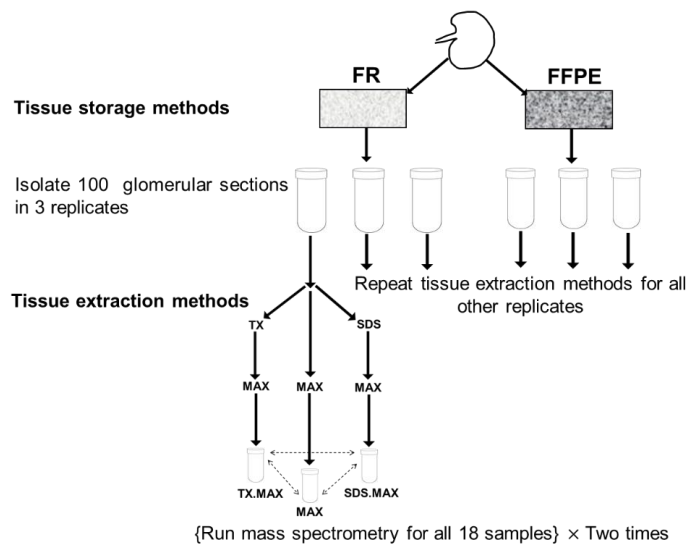
The Proteome Discoverer was used for extraction of MS2 scan data from the Xcalibur RAW file, separate searches of CID and ETD MS2 scans in Mascot and Sequest, and collection of the results into a single file (.msf extension) prior to loading into Scaffold Q+S v4.4.3. The FDR for peptides was calculated using

the Scaffold Local FDR algorithm. Protein probabilities were calculated using the Protein Prophet algorithm. Proteins were grouped by the Scaffold protein cluster analysis to satisfy the parsimony principle. Label-free quantification of identified proteins were exported as total precursor ion area values to an excel sheet for analysis of proteomics data (Please see the file “ProteomicsData\_Kidney.xlsx”). We analyzed the data for comparing statistical methods with MVs in the presence of heterogeneity.

**Proteomics data analysis**

The purpose of this study is to (1) compare variability between (a) tissue storage methods and (b) tissue extraction methods; (2) compare various statistical approaches of analysis and normalization methods.

We have two TSMs (FR and FFPE) and three TEMs (MAX, TX.MAX, SDS.MAX) with three replicates and two MS runs leading to 36 samples (total number of samples =  $2 \times 3 \times 3 \times 2 = 36$ ). A flow chart of the experiment is given below in Figure 3.1.



**Figure 3.1.** Flowchart of the proteomics experiment

In the flowchart, we have shown the basic steps of carrying out the experiment involving TSMs and TEMs. The MS was repeated twice to get more reliable results for estimating experimental variability. We obtained the following six groups as given below in the Table 3.1.

**Table 3.1.** Table showing different groups under study

		TSM →	
		FR	FFPE
TEM ↓	Direct	1 (FR_MAX)	4 (FFPE_MAX)
	Sequential Extraction	TX.MAX	2 (FR_TX.MAX)
SDS.MAX		3 (FR_SDS.MAX)	6 (FFPE_SDS.MAX)

There are three replicates for each of the six groups thus leading to 18 samples. Then, we have repeated the MS two times for the 18 samples and we obtained six samples for each of the six groups.

***Data preprocessing***

Initially, there were 728 proteins identified in both runs, 380 proteins identified in run 1 only and 342 proteins identified in run 2 only. There was a total of 1450 identified proteins out of which 1376 proteins were unique, and 37 proteins were redundant and duplicate entries were removed from the data. Furthermore, there were 111 proteins for which all the samples have MVs (NA values). Therefore, we are left with protein data with 1302 proteins that correspond to 1178 gene symbols. The percentage of NA values within each sample (36 samples) ranges from 41.3%-78.3% with a median value of 49.5%.

As we have a greater number of groups, therefore it is difficult to perform analysis with this data having MVs. If we discard the proteins having any MVs in



any of the samples in a group, then there will be only 26 proteins available. Another way is to retain the proteins having at least one or two observations in each group. A summary of number of proteins available in each group is given below in Table 3.2.

**Table 3.2.** Summary of number of proteins and missing values in different groups

Groups	No. of proteins with no MVs	No. of proteins with MVs in all samples	No. of proteins with at least one observation	No. of proteins with at least two observations
FR_MAX	448	205	1097	995
FR_TX.MAX	357	324	978	881
FR_SDS.MAX	170	678	624	454
FFPE_MAX	373	295	1007	874
FFPE_TX.MAX	353	261	1041	890
FFPE_SDS.MAX	381	237	1065	920

If we use the number of proteins having at least one observation in a group, then we can assess a greater number of proteins. However, we need at least two observations in each group to calculate the coefficient of variation (CV) for a protein in each group. Therefore, we used 372 proteins which have at least two observations in each of the six groups for further analysis.

### ***Statistical approaches***

The analysis of proteomics data becomes more complex due to non-normality behavior of the data, and greater proportion of MVs within and across the samples. To get a better insight of proteomics data analysis while dealing with these problems, we have performed the analysis using three methods which are as follows:

A1. Method for data excluding missing values: Proteins having complete observations for all the samples, i.e., no MVs, were used for comparison. Proteins having MVs were discarded from the analysis.

A2. Method for data including missing values: The proteins with MVs across the samples were analyzed using unbalanced ANOVA method [60].

A3. Method for data using imputation: The MVs were imputed after applying the normalization methods to the data [61] as given in next section. We have used the “impute.MAR” function of the R package “imputeLCMD” [62] for imputing the MVs. Three different types of imputation under the assumption of MAR or MCAR, namely, MLE [63], SVD [64] and KNN [65, 66] are available in this package. We have used only the SVD method (A3) for imputation.

We applied three different data transformation and/or normalization methods:

N1. Logarithmic transformation: The raw data is transformed by using logarithmic base 2.

N2. Quantile normalization (QN): It is done by using log base 2 transformation of raw data followed by “normalize.quantiles” method [67] available in R package “preprocessCore” [68].

N3. Variance stabilizing normalization (VSN): It is done by applying “justvsn” function available in R package “vsn” [69] to the raw data.

Therefore, by using three methods of analysis (A1, A2 and A3) based on three transformation and/or normalization methods (N1, N2 and N3), we have 9 different combinations (statistical approaches): excluding MVs (A1.N1, A1.N2, A1.N3); including MVs (A2.N1, A2.N2, A2.N3); imputing MVs (A3.N1, A3.N2,

A3.N3). We preprocessed the data using these methods to get 9 different datasets (preprocessed data) for 6 groups having 6 samples in each group.

We calculated the CV for each protein in the groups: TSM (FR vs. FFPE), TEM (MAX vs. TX.MAX vs. SDS.MAX) and TSM×TEM (FR\_MAX, FR\_TX.MAX, FR\_SDS.MAX, FFPE\_MAX, FFPE\_TX.MAX, FFPE\_SDS.MAX). It has two purposes: (i) Which TSM/ TEM/ TSM×TEM have the minimum CV based on different statistical approaches; (ii) Which statistical approach leads to the minimum CV. We have used ANOVA model as given below for studying the contribution of variability due of TSM, TEM and the interaction term TSM×TEM:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (3.1)$$

where,  $y_{ijk}$  is the transformed and/or normalized data for a protein,  $\alpha_i$  ( $i = 1, 2$ ) is the  $i^{\text{th}}$  TSM effect,  $\beta_j$  ( $j = 1, 2, 3$ ) is the  $j^{\text{th}}$  TEM effect and  $(\alpha\beta)_{ij}$  is the interaction effect, TSM×TEM. The term  $\varepsilon_{ijk}$  is the normally distributed error component and  $\varepsilon_{ijk} \sim N(0, \sigma^2)$ . The mapping of the above model to the experimental design allows us to estimate the contribution due to each source of variation for each protein.

## Results and discussion

### Comparison of CV among various groups

We have 141, 372 and 372 proteins obtained by using the analysis methods A1, A2 and A3, respectively. The summary of CV (in %) using 9 different statistical approaches for comparisons among TSMs and TEMs is shown in Table 3.3.

**Table 3.3.** Summary of CV (in %) using 9 statistical approaches among TSM and TEM

		TSM		TEM		
		FR	FFPE	MAX	TX.MAX	SDS.MAX
MV Excluded	A1.N1	6.92 (2.23, 12.77)	2.76 (2.00, 9.49)	3.25 (1.93, 9.64)	3.26 (2.05, 15.90)	7.40 (2.38, 15.24)
	A1.N2	6.29 (0.52, 12.50)	1.30 (0.55, 5.10)	1.94 (0.34, 9.02)	1.91 (0.32, 12.67)	6.74 (0.59, 14.99)
	A1.N3	6.25 (0.95, 12.51)	1.28 (0.48, 8.31)	2.03 (0.28, 9.33)	1.95 (0.26, 15.21)	6.81 (1.05, 15.01)
MV Included	A2.N1	7.08 (1.23, 12.77)	2.92 (0.83, 11)	3.50 (0.65, 12.21)	3.49 (0.73, 15.90)	7.53 (0.23, 16.95)
	A2.N2	6.62 (0.39, 12.51)	1.75 (0.52, 9.13)	2.71 (0.16, 12.16)	2.49 (0.32, 14.42)	7.17 (0.42, 16.48)
	A2.N3	6.68 (0.80, 12.49)	1.73 (0.47, 11.16)	2.71 (0.28, 11.61)	2.55 (0.20, 15.22)	7.21 (0.76, 15.54)
MV Imputed	A3.N1	7.72 (2.23, 17.47)	3.29 (1.70, 15.28)	4.03 (1.79, 15.49)	3.87 (1.72, 15.90)	8.03 (2.38, 18.01)
	A3.N2	7.10 (0.39, 15.96)	2.15 (0.52, 13.60)	3.10 (0.45, 14.64)	2.98 (0.38, 14.42)	7.35 (0.56, 19.25)
	A3.N3	7.07 (1.01, 18.34)	2.13 (0.47, 13.64)	3.10 (0.33, 16.02)	3.04 (0.28, 15.22)	7.35 (1.08, 18.68)

**Note:** The first figure is the median value and the figures inside the parenthesis are respectively, minimum and maximum value.

The summary of CV (in %) using 9 different statistical approaches for comparisons among six groups of TSMxTEM is shown in Table 3.4.

**Table 3.4.** Summary of CV (in %) using 9 statistical approaches among six groups of TSMxTEM

		FR_ MAX	FR_ TX.MAX	FR_ SDS.MAX	FFPE_ MAX	FFPE_ TX.MAX	FFPE_ SDS.MAX
MV Excluded	A1.N1	2.64 (1.34, 8.62)	2.71 (0.83, 9.95)	4.73 (2.25, 12.90)	3.00 (1.96, 7.14)	2.87 (2.08, 13.80)	2.34 (0.75, 8.34)
	A1.N2	0.87 (0.12, 6.26)	1.05 (0.18, 9.09)	2.32 (0.22, 10.55)	0.87 (0, 5.12)	0.96 (0, 7.28)	0.85 (0, 8.13)
	A1.N3	0.77 (0.17, 7.53)	1.01 (0.13, 9.87)	2.37 (0.32, 11.92)	0.84 (0.12, 6.14)	0.95 (0.18, 11.75)	0.83 (0.10, 8.10)
MV Included	A2.N1	2.64 (0.05, 11.71)	2.81 (0.14, 10.93)	4.49 (0.03, 19.81)	2.97 (0.09, 13.33)	3.01 (0.15, 13.8)	2.41 (0.17, 17.14)
	A2.N2	1.08 (0, 10.62)	1.47 (0, 9.33)	2.88 (0.07, 16.32)	1.28 (0, 10.50)	1.32 (0, 12.62)	1.14 (0, 13.32)
	A2.N3	1.09 (0.04, 9.67)	1.39 (0.04, 9.87)	2.44 (0.02, 17.52)	1.28 (0.01, 9.55)	1.41 (0.12, 12.45)	1.19 (0.07, 17.72)
MV Imputed	A3.N1	2.94 (0.95, 16.56)	3.26 (0.83, 15.27)	5.06 (2.25, 17.75)	3.40 (1.34, 16.87)	3.33 (0.62, 15.62)	2.86 (0.69, 16.21)
	A3.N2	1.59 (0.24, 17.06)	1.83 (0.06, 14.28)	2.77 (0.20, 19.86)	1.78 (0.02, 15.03)	1.70 (0.02, 14.08)	1.75 (0.03, 14.23)
	A3.N3	1.57 (0.14, 19.00)	1.82 (0.19, 15.69)	2.48 (0.32, 17.28)	1.74 (0.07, 14.88)	1.7 (0.21, 14.28)	1.63 (0.16, 15.38)

**Note:** The first figure is the median value and the figures inside the parenthesis are respectively, minimum and maximum value.

TSM: We found that median value of CV is lowest in FFPE using all the statistical approaches. Furthermore, within FFPE, the normalization method N3 has the minimum value of median CV for each analysis method. Overall, the minimum median CV is for A1.N3 in FFPE.

TEM: We have the minimum median value of CV in TX.MAX. We found A1.N2 has the minimum value of median CV.

TSM×TEM: We have the minimum median value of CV in FR\_MAX followed by FFPE\_SDS.MAX using all the approaches. We found A1.N3 has the minimum value of median CV in all the groups except for A1.N2 in FR\_SDS.MAX. Overall, the minimum median CV is for A1.N3 in group FR\_MAX.

Based on median CV, FFPE is a better choice than FR using all the statistical approaches. Similarly, among TEMs, TX.MAX has the least CV and can be a better choice. However, based on the maximum value of CV, MAX is a better choice for TEM. If we consider approaches (A2 & A3) having greater number of proteins and TEM within FFPE, we see that A3.N3 in FFPE\_SDS.MAX is having the least median CV (1.63).

### Contribution of Sum of Squares (SS) due to each component

The percent contribution of SS due to each variable to the total SS was computed for each protein. A summary of contribution of each variable to the total variability is given below in Table 3.5.

**Table 3.5.** Summary of the contribution of % SS due to TSM, TEM and TSM×TEM

		<b>SS<sub>TSM</sub></b>	<b>SS<sub>TEM</sub></b>	<b>SS<sub>TSM×TEM</sub></b>
<b>MV Excluded</b>	<b>A1.N1</b>	9.86 (0, 68.98)	20.9 (0.47, 36.32)	32.87 (0.29, 54.41)
	<b>A1.N2</b>	14.71 (0, 78.88)	27.49 (1.35, 48.44)	43.21 (0.92, 64.54)
	<b>A1.N3</b>	15.05 (0, 73.78)	26.7 (2.31, 44.92)	41.88 (0.59, 65.23)

<b>MV Included</b>	<b>A2.N1</b>	10.84 (0, 83.65)	20.97 (0.08, 49.47)	33.46 (0.29, 78.05)
	<b>A2.N2</b>	12.59 (0, 85)	25.56 (0.06, 54.68)	39.37 (0.08, 80.29)
	<b>A2.N3</b>	12.84 (0, 88.18)	25.72 (0.04, 53.37)	40.32 (0.06, 77.54)
<b>MV Imputed</b>	<b>A3.N1</b>	8.52 (0, 73.76)	18.83 (0, 40.46)	29.86 (0.09, 57.77)
	<b>A3.N2</b>	11.07 (0, 85.67)	23.53 (0.03, 50.93)	37.33 (0.05, 65.75)
	<b>A3.N3</b>	11.18 (0, 85.31)	23.32 (0, 49.68)	37.26 (0.14, 65.32)

**Note:** The first figure is the median value and the figures inside the parenthesis are respectively, minimum and maximum value.

We found that the TSM has the least contribution to the total variability whereas interaction term has the maximum contribution ( $SS_{TSM} < SS_{TEM} < SS_{TSM \times TEM}$ ). The imputation method leads to decrease in the SS contribution due to each variable.

The proportion of proteins showing significant effects due to TSM, TEM and TSM $\times$ TEM using 9 different approaches are given in Table 3.6.

**Table 3.6.** The summary of proportion of proteins showing effects due to the variables: TSM, TEM and TSM $\times$ TEM

		<b>N<sub>TSM</sub></b>	<b>N<sub>TEM</sub></b>	<b>N<sub>TSM<math>\times</math>TEM</sub></b>
<b>MV Excluded</b>	<b>A1.N1</b>	0.65/ 0.62/ 0.33	0.77/ 0.76/ 0.5	0.77/ 0.77/ 0.65
	<b>A1.N2</b>	0.84/ 0.84/ 0.72	0.91/ 0.91/ 0.77	0.89/ 0.88/ 0.78
	<b>A1.N3</b>	0.82/ 0.82/ 0.71	0.87/ 0.87/ 0.72	0.87/ 0.85/ 0.77

<b>MV Included</b>	<b>A2.N1</b>	0.61/ 0.57/ 0.25	0.72/ 0.72/ 0.28	0.79/ 0.79/ 0.49
	<b>A2.N2</b>	0.75/ 0.73/ 0.48	0.83/ 0.82/ 0.58	0.87/ 0.87/ 0.68
	<b>A2.N3</b>	0.74/ 0.74/ 0.52	0.81/ 0.81/ 0.6	0.85/ 0.84/ 0.67
<b>MV Imputed</b>	<b>A3.N1</b>	0.58/ 0.53/ 0.24	0.69/ 0.67/ 0.35	0.78/ 0.77/ 0.52
	<b>A3.N2</b>	0.71/ 0.68/ 0.48	0.81/ 0.8/ 0.58	0.86/ 0.85/ 0.69
	<b>A3.N3</b>	0.7/ 0.69/ 0.49	0.8/ 0.78/ 0.58	0.84/ 0.83/ 0.67

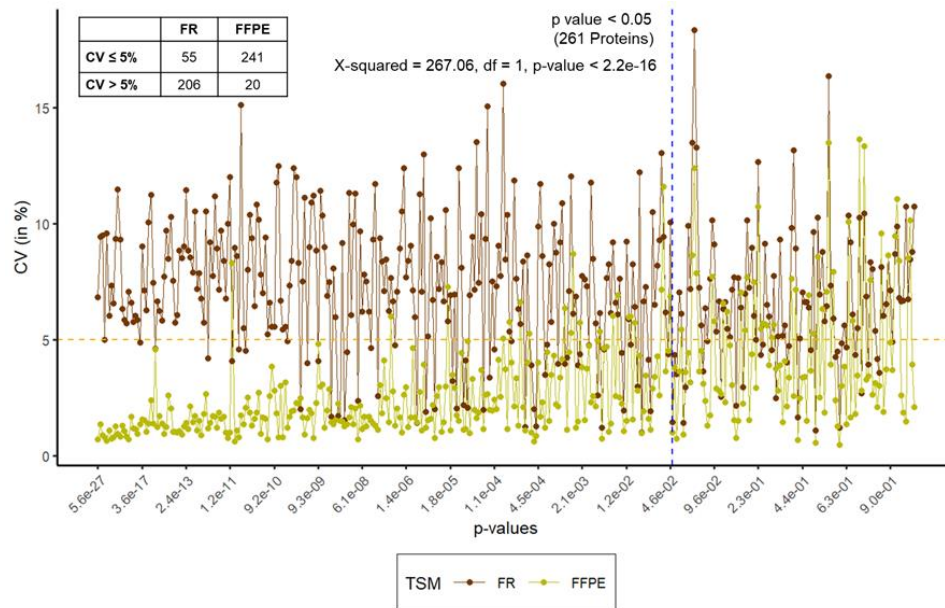
**Note:** The result obtained using p-values corresponding to without adjustment, BH adjusted and Bonferroni adjusted are separated serially by slash “/” in the table.

The proportion of proteins showing significant effects due to TSM and TEM and their interaction vary with each statistical approach. The TSM has the least proportion of significant proteins as compared to those of TEM and TSMxTEM. This shows that TSM has the least influence. Furthermore, the imputation approach has the least proportion of significant proteins. This shows that imputation of MVs is a better approach for analysis as it leads to reduction in variability and increase in the number of proteins assessed for analysis.

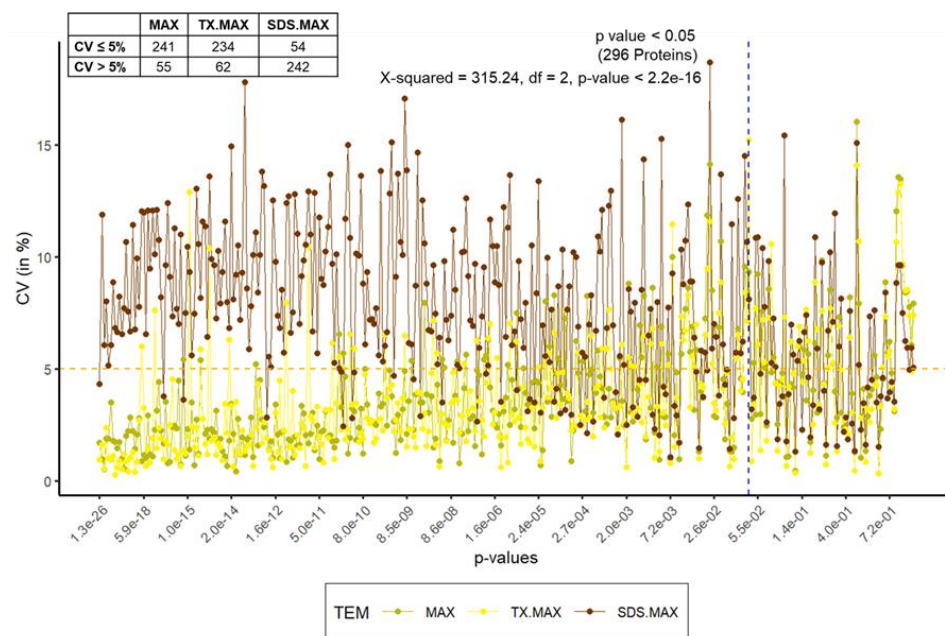
### **Analysis for imputed data using VSN**

We used ANOVA to test the significance of proteins based on TSM and TEM. The plot of CV (in %) of the proteins in increasing order of p-values based on A3.N3 for TSM and TEM are respectively given in Figures 3.2 and 3.3.





**Figure 3.2.** Plot of CV (in %) versus the proteins with increasing order of p-values for TSM (FR and FFPE)



**Figure 3.3.** Plot of CV (in %) versus the proteins with increasing order of p-values for TSM (MAX, TX.MAX and SDS.MAX)

There are respectively 261 and 296 proteins showing significant effects due to TSM and TEM. From Figure 3.2, we see that FR has more CV as compared to that of FFPE for most of the proteins. From Figure 3.3, we found

SDS.MAX has more CV as compared to those of MAX and TX.MAX. We applied chi-square test for the proteins having significant effects due to TSM and TEM. We found that there is association between the TSM and the CV (p-value < 0.001). Similarly, in case of TSM, we found that there is association between the variables, TEM and CV (p-value < 0.001).

We found that the FFPE is a better method than that of the FR for tissue storage. Further, we found that MAX, the single step approach is better than those of two-step approach for tissue extraction. The maximum contribution to the total variability is due to the interaction effect TSM $\times$ TEM and TEM. The TSMs and TEMs have significant effects on the protein expression. However, the effect due to TSM is the least. In the present article, we have used different analysis and normalization methods for the proteomics data. The number of proteins for testing can be increased by either by including the MVs (A2) or by using imputed data (A3). The imputation method (A3) has the least SS contribution than those of A1 (complete data) and A2 (unbalanced data). We found the least proportion of significant proteins when using the imputation method (A3). The normalization method N1, i.e., only logarithmic transformation is not suited for analyzing the proteomics data. The other normalization methods N2 and N3 having lesser CV can be a better approach.

## **Conclusion**

Our study discussed the technical issues with a focus on the statistical analysis. It will provide better insight to the researchers while designing and executing experiments. There may be small changes caused during sample handling and

storage, different batches of buffer, electrospray, instrument components, calibration and tuning, etc. While designing any proteomics experiment, we must identify the variability associated with technical steps. The researchers involved in proteomics research area can use this data for further study. The data can further be used for planning new proteomics experiments. In the future, we will come up with a rigorous statistical approach using different proteomics dataset that could overcome the heterogeneity problem caused due to technical reasons in the proteomics data with MVs. We found that the CVs obtained using all the approaches is lesser for FFPE as compared to those of FR. Among the TEMs, we found that TX.MAX has the least value based on median CV and MAX has the least value based on maximum CV. The normalization methods N2 and N3 have lesser CV as compared to that of logarithmic transformation. Based on SS, we found that the TSM has the least contribution to the total variability. The imputation of MVs leads to reduction in variability and increase in the number of proteins assessed for analysis. Therefore, we can recommend: (i) FFPE is the better choice than FR for tissue storage, (ii) one-step TEM is better than the two-step TEM, (iii) Imputation method (A3) is the best approach, (iv) N2 or N3 method of normalization should be the preferred choice.

## CHAPTER 4

### INTERACTIVE WEB TOOL FOR STANDARDIZING PROTEOMICS WORKFLOW FOR LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY DATA

#### **Introduction**

Standardization of experimental workflow is an essential task for carrying out proteomics experiments [4, 70]. There are various technical steps involved in proteomics experiments such as sample collection, sample storage, sample preparation, extraction, liquid chromatography (LC) - mass spectrometry (MS) detection. The experimenters have various choices available for each step in the proteomics workflow. Therefore, it becomes necessary to find the most suitable choice for each step in the proteomics workflow. LC-MS is used in proteomics as a method for identification and quantification of features (peptides/proteins) in complex mixtures [42, 43]. There are several challenges associated with the proteomics data such as data heterogeneity due to technical reasons, MVs and low-abundant features. Furthermore, the proteomics data can be the balanced (equal number of observations in each group) or unbalanced (unequal number of observations in each group). The data can be unbalanced due to unequal number of subjects, or missing observations, or both. The missing values (MVs) in proteomics data can occur due to biological and/or technical issues.

The missing observations are broadly categorized as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [53].

Various studies have been done for studying the data variability, standardization and quality control of proteomics expression data [56, 71-73]. However, only a few tools exist for the standardization and quality control of proteomics expression data based on different approaches [74-76]. Therefore, we have developed a user-friendly tool for standardizing the proteomics workflow and studying the variability in proteomic expression data generated by high throughput technologies involving MS [20, 22, 24, 41]. We use the special cases of general linear model (GLM), analysis of covariance (ANCOVA) and analysis of variance (ANOVA) to study the data variability. The user can estimate the contribution of various sources of variation to the overall variability. The study of data variability can be done using various analysis methods and normalization techniques. The user can analyze the data either by excluding the features having missing observations or by imputing the MVs. Excluding the features having missing observations leads to loss of information from the experiment. Therefore, we have provided two imputation methods to include more features in the analysis. We have demonstrated the tool using simulated proteomics data comprising of 1000 peptides corresponding to 200 proteins. We implemented all the steps in R [1] and used “shiny” package [3] for developing the web application. The PWST tool can be accessed freely by the users from <https://ulbbf.shinyapps.io/pwst/>.

## Methods

The steps and various options available under each step are described below. More details about all the steps are given in the next section.

1. Upload the proteomics expression data
2. Feature type: The analysis can be done either at protein level or peptide level. After uploading the expression data file, the user has to select the feature type.
3. Aggregation method: We have provided four options for data aggregation: (i) Mean, (ii) Median, (iii) Sum, (iv) Maximum. Data aggregation is required if the user has provided the peptide data and wants analysis at protein level. It is also applicable to other situations, such as when the features (proteins or peptides) are redundant. For example, if the user uses more than one database for searching features, there may be many redundant features.
4. Upload the additional information: The user has to upload the additional information about the data. This file contains the information of the samples and the variables under study. The variables may be categorical and/or continuous (numeric).
5. Choose the categorical variables: The user has to select the categorical variables which will automatically pop out after the file containing additional information has been uploaded. Categorical variables contain a finite number of categories/groups. Examples of the categorical variables in proteomics workflow are: storage methods, extraction methods, etc.
6. Choosing the numeric variables: After selecting the categorical variables, the user can now select the numeric (continuous) variable from the remaining

variables, if available. The variable may contain any value within some range. Examples of numeric variables are age, weight, height, etc. of the individuals.

7. Analysis method: We have provided two options for the analysis:

(i) Excluding missing values: Features having MVs in any of the samples are discarded from the analysis. The features having observations in all the samples are retained for analysis. This approach may not be appropriate as it will exclude many features. Therefore, we have provided the imputation methods.

(ii) Imputing missing values: The MVs are imputed after applying the normalization methods to the data [61] as given in next section. We have provided two imputation methods under the assumption of MAR or MCAR, namely, SVD [64] and KNN [65, 66] available from the “impute.MAR” function of the R package “imputeLCMD” [62]. We impute the data at protein level if the data is available at protein level. Otherwise, we impute the data at peptide level. In case, if the analysis is to be done at protein level for the peptide data, then we first impute the data at peptide level and then aggregate the data. By default, the imputation is done globally. However, the user can apply the imputation methods group wise by specifying additional column “Norm\_Imp\_Group” and the group numbers in the file containing additional information.

8. Transformation/Normalization method: There are four options available for data transformation and/or normalization:

(i) Logarithmic transformation: The raw data is transformed by taking log base 2.

(ii) Quantile normalization (QN): This method is applied on log base 2 transformed data using the “normalize.quantiles” method [67] available in R package “preprocessCore” [68].

(iii) Variance stabilizing normalization (VSN): This method is applied on the raw data using “justvsn” function available in R package “vsn” [69].

(iv) None: In some situations, if the user wants to use his own normalized data, then he can use the “None” option.

By default, the normalization methods (QN and VSN) are applied globally. The user can apply the normalization methods (QN and VSN) group wise by specifying additional column “Norm\_Imp\_Group” and the group numbers in the file containing additional information.

9. Level of significance: The user can specify the level of significance ( $\alpha$ ). By default, the level of significance is 0.05.

10. Method of adjustment: The user must adjust the p-values for multiple testing of features for which we have provided the following options: “BH”, “bonferroni”, “holm”, “Hochberg”, “hommel” and “BY”. The method “BH” is the default adjustment method.

The user has to hit the “Submit” button after specifying the above-mentioned inputs. The user will get the following results under different tabs:

1. Inputs selected: It shows the various inputs defined by the user for the analysis.

2. Visual plots of the preprocessed data: We provide exploratory plots of the preprocessed data such as box plot, density plot, correlation heatmap.



3. The sum of squares (SS) results: We fit the ANOVA/ ANCOVA model with fixed effects for each feature. The results are comprised of: (i) A table showing the contribution of SS due to each variable, the p-values and the adjusted p-values corresponding to each variable, (ii) summary of % contribution of SS and (iii) box plot showing % contribution of SS due to each variable.

4. The coefficient of variation (CV) analysis: We calculate the CV (in %) corresponding to the groups within each categorical variable. The results consist of: (i) A table showing the CV of different groups of each categorical variable for all the features, (ii) summary of CV and (iii) box plot showing CV under the various groups of categorical variables.

5. Number of significant features: We provided a table showing the number of features without and with adjustment which have significant effect due to each variable.

All these results can be viewed and downloaded. The complete demonstration of the tool is discussed in next section.

### **Demonstration and discussion**

We used a simulated dataset generated with the aid of the kidney proteomics expression data (used in Chapter 3) for demonstrating our tool. We generated a proteomics expression data set that consists of 200 proteins with 1000 peptides. Suppose there are two steps (M1 and M2), e.g., tissue storage method and tissue extraction method, involved in an experiment. Our purpose is to study the variability associated with the two steps/ variables/ methods. Furthermore, suppose we have respectively two approaches of M1 (A1 & A2) and three

approaches of M2 (B1, B2 & B3). Now our purpose is to select the most suitable approach for M1 and M2. In the example dataset, we have respectively two levels of M1 and three levels of M2 each with three biological replicates. The MS is repeated two times so that we have total 36 samples. We have included “Age” of the subjects (biological replicates) as a numeric (continuous) variable. The screenshot of the PWST tool is shown below in Figure 4.1.

**PWST (Proteomics Workflow Standardization Tool)**

Proteomics Data Analysis

Choose file to upload the expression data

Browse... No file selected

Select feature type

Aggregation method

Mean

Choose file to upload the additional information

Browse... No file selected

Select categorical variables

Select numeric variables

Analysis method

Excluding missing values

Imputing missing values

Normalization method

Log base 2

Quantile Normalization

Variance Stabilizing Normalization

None

Level of significance

0.05

Method of adjustment

BH

Submit

Inputs selected Box plot (Preprocessed data) Density plot (Preprocessed data)

Correlation heatmap (Preprocessed data) SS\_Features SS%\_Summary Box plot: SS (%) contribution

CV%\_Features CV%\_Summary Box plot: CV (in %) # Significant Features

Download\_Inputs\_Selected

**Figure 4.1.** Webpage of the tool “PWST”

## Inputs to be specified by the user

1. Upload the expression data: The user has to upload the proteomics expression data either in csv, tsv, txt, xls or xlsx format. The first two columns are reserved for proteins and peptides. Even if the data is available at protein level only and there is no peptide data, then the user must leave the second column blank. The expression data must start from the third column and onwards. The first row must contain the labels such as “Protein”, “Peptide” and the sample names (from third column). After the first row, we have the name of proteins and peptides in the first and second column respectively. In the remaining portion, we have the expression values of corresponding features (proteins/ peptides) and samples. A portion of input expression data is shown below.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Protein	Peptide	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
2	Protein_1	AKIWCTIDYY	67805000	181605000	117605000	NA	114605000	41405000	99405000	103605000	NA	31305000
3	Protein_1	KLQMIK	67762000	181562000	117562000	51362000	114562000	41362000	99362000	103562000	156562000	31262000
4	Protein_1	KWFHDSHKYSHFVNQ	67476000	181276000	117276000	51076000	114276000	41076000	NA	103276000	156276000	NA
5	Protein_1	KVVMAIRFMIEKKY	67425000	181225000	117225000	51025000	114225000	41025000	99025000	103225000	156225000	30925000
6	Protein_1	QDHSTCTEG	68184000	181984000	117984000	51784000	114984000	41784000	99784000	103984000	156984000	31684000
7	Protein_1	NWTCPLDYI	67297000	181097000	117097000	50897000	114097000	40897000	98897000	103097000	156097000	30797000
8	Protein_2	KSCVLFVTHWDN	619880000	1140880000	801880000	260880000	627880000	NA	432880000	390880000	627880000	156880000
9	Protein_2	VLQNNVGLHK	619232000	1140232000	801232000	NA	627232000	295232000	432232000	390232000	NA	156232000
10	Protein_2	AESPTMKVAQLLT	619772000	1140772000	801772000	260772000	627772000	295772000	432772000	390772000	NA	156772000
11	Protein_2	VHQVYSWECCT	619471000	1140471000	801471000	260471000	627471000	295471000	432471000	390471000	627471000	156471000
12	Protein_2	QPGHTMAGAQNTYWGED	619408000	NA	NA	260408000	627408000	295408000	NA	390408000	NA	156408000
13	Protein_3	FCFCAIDEDTLQTAQ	28923000	56823000	26323000	24823000	NA	NA	NA	NA	12123000	9370200
14	Protein_4	RYAQLIKKVLIRG	17235000	22235000	17135000	4326200	NA	NA	10023100	NA	17335000	5150600
15	Protein_5	ILPRFIKRHTIEFWENNQ	163491000	NA	NA	85991000	NA	67291000	158491000	180491000	NA	48191000
16	Protein_5	APYINGCHML	163013000	154013000	154013000	85513000	NA	66813000	NA	NA	239013000	47713000
17	Protein_6	YWWAQHETHDQMD	164643000	280643000	181643000	96143000	NA	NA	NA	49643000	94243000	10943000
18	Protein_7	AEMMVEIWGPHDPVKQ	NA	NA	25720000	16720000	25620000	11920000	31420000	47920000	59920000	8976400
19	Protein_7	SSYHKFHG	31134000	49834000	25934000	16934000	25834000	12134000	31634000	48134000	60134000	9190400
20	Protein_8	LVAHDPLCVMS	NA	221955000	119955000	NA	86855000	NA	67155000	109955000	173955000	37355000
21	Protein_9	WVSKSK	12439000	38639000	22839000	8514000	30039000	12839000	30039000	43039000	41539000	17439000
22	Protein_9	MSTI RWTMKPWVQVI	NA	38884000	23084000	8759000	30284000	13084000	30284000	43284000	41784000	17684000

Figure 4.2. A portion of proteomics expression data

The user has to click on the “Browse...” button and select the file to upload the expression data as given below in Figures 4.3:

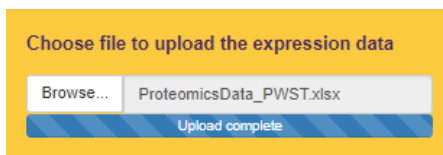
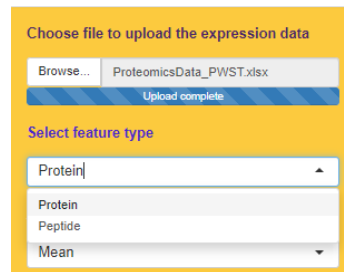


Figure 4.3. Upload the proteomics expression data

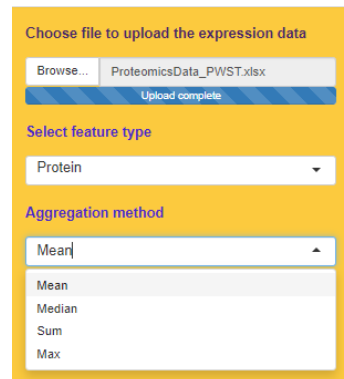
2. Feature type: After uploading the expression data file, the user has to select the feature type. The feature type available will automatically be detected. In the given example, we have the peptide data. So, there are two options available: “Protein” or “Peptide”. We selected the analysis to be done at “Protein” level as given below:



The screenshot shows a web interface for uploading expression data. At the top, it says "Choose file to upload the expression data". Below this is a file selection area with a "Browse..." button and a text box containing "ProteomicsData\_PWST.xlsx". A blue "Upload complete" button is below the text box. Underneath, there is a section titled "Select feature type" with a dropdown menu. The dropdown is open, showing "Protein" as the selected option, with "Peptide" and "Mean" listed below it.

**Figure 4.4.** Choose the feature type – “Protein” or “Peptide”

3. Aggregation method: There are four options available for data aggregation: (i) Mean, (ii) Median, (iii) Sum, (iv) Maximum. We selected “Mean” for aggregating the peptide data at protein level as given below:



The screenshot shows the same web interface as Figure 4.4. The "Select feature type" dropdown is now closed and shows "Protein". Below it is a section titled "Aggregation method" with a dropdown menu. The dropdown is open, showing "Mean" as the selected option, with "Median", "Sum", and "Max" listed below it.

**Figure 4.5.** Choose the aggregation method (Mean/Median/Sum/Max)

4. Upload the additional information: Now the user has to upload the additional information about the data either in csv, tsv, txt, xls or xlsx format. This file contains the information of the samples and the variables under study. The

variables may be categorical and/or numeric (continuous). A portion of additional data is shown below.

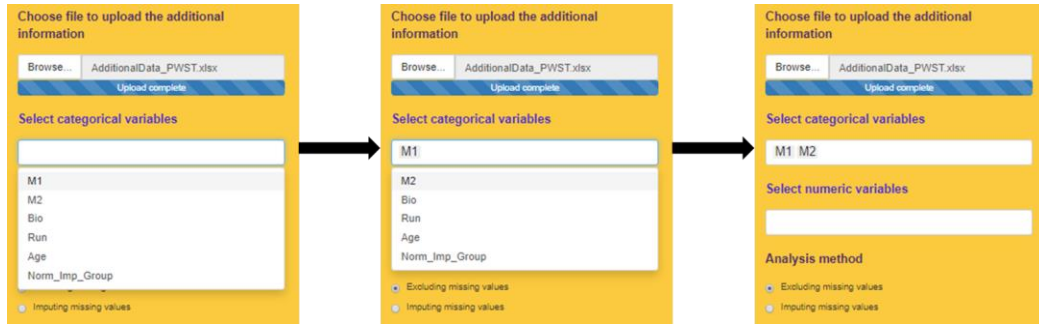
	A	B	C	D	E	F	G
1	Samples	M1	M2	Bio	Run	Age	Norm_imp_Group
2	S1	A1	B1	1	1	27	1
3	S2	A1	B1	2	1	36	1
4	S3	A1	B1	3	1	23	1
5	S4	A1	B1	1	2	27	1
6	S5	A1	B1	2	2	36	1
7	S6	A1	B1	3	2	23	1
8	S7	A1	B2	4	1	32	2
9	S8	A1	B2	5	1	33	2
10	S9	A1	B2	6	1	42	2
11	S10	A1	B2	4	2	32	2
12	S11	A1	B2	5	2	33	2
13	S12	A1	B2	6	2	42	2
14	S13	A1	B3	7	1	36	3
15	S14	A1	B3	8	1	29	3
16	S15	A1	B3	9	1	39	3
17	S16	A1	B3	7	2	36	3
18	S17	A1	B3	8	2	29	3
19	S18	A1	B3	9	2	39	3
20	S19	A2	B1	10	1	28	4
21	S20	A2	B1	11	1	35	4
22	S21	A2	B1	12	1	22	4

**Figure 4.6.** A portion of additional information of data

The user has to click on the “Browse...” button and select the file to upload the additional data as given below:

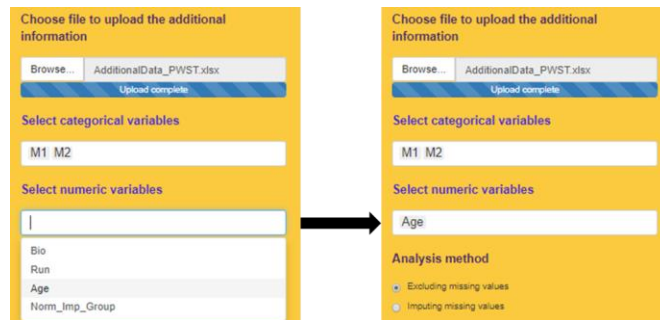
**Figure 4.7.** Upload the additional information of data

5. Choose the categorical variables: The user has to select the categorical variables one by one which will automatically pop out after the file containing additional information has been uploaded. We have selected “M1” and “M2” as the categorical variables under study as shown in Figure 4.8.



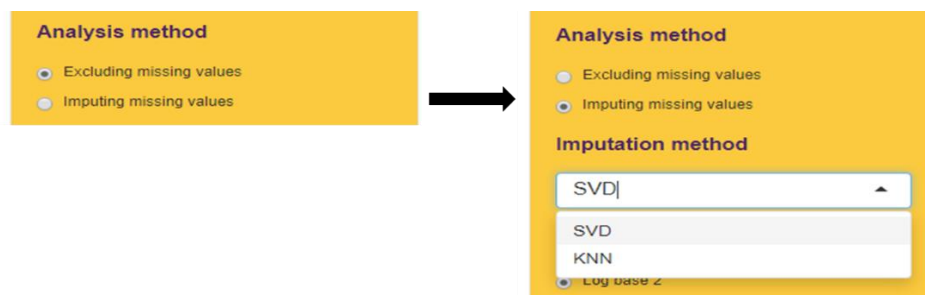
**Figure 4.8.** Selection of categorical variables

6. Choosing the numeric variables: After selecting the categorical variables, the user can now select the continuous variable from the remaining variables, if available. In this example, we have selected “Age” as given below.



**Figure 4.9.** Selection of numeric variables

7. Analysis method: We have provided two options for the analysis: (i) Excluding missing values and (ii) Imputing missing values. Further, there are two methods of data imputation available: (a) SVD and (b) KNN. We selected the radio button “Imputing missing values” and “SVD” method for data imputation. The screenshots are given below.



**Figure 4.10.** Selection of analysis method

8. Transformation/Normalization method: There are four options available for data transformation and/or normalization. We selected “Variance Stabilizing Normalization” for data normalization as given below.

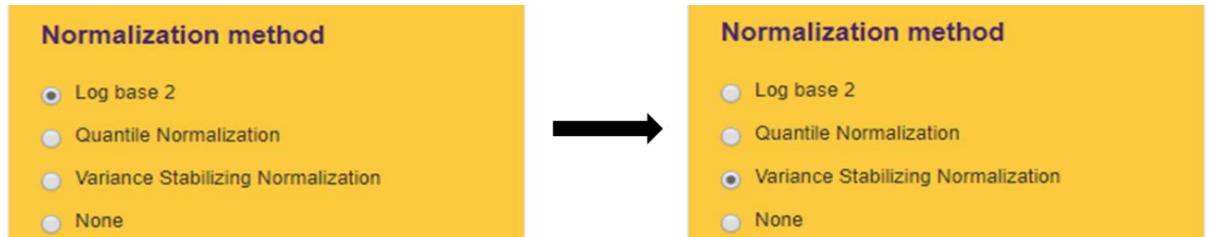


Figure 4.11 shows two panels of the 'Normalization method' selection interface. The left panel shows four radio button options: 'Log base 2', 'Quantile Normalization', 'Variance Stabilizing Normalization', and 'None'. The right panel shows the same options, but 'Variance Stabilizing Normalization' is selected with a filled radio button.

**Figure 4.11.** Selection of normalization method

9. Level of significance: The user has to specify the level of significance. We have selected the default value 0.05 as the level of significance.

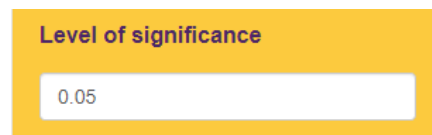


Figure 4.12 shows a yellow box titled 'Level of significance' containing a text input field with the value '0.05'.

**Figure 4.12.** Specify the level of significance

10. Method of adjustment: The user has to select the method of adjusting the p-values for multiple testing of features. We have provided six adjustment methods. We selected “BH” adjustment method.

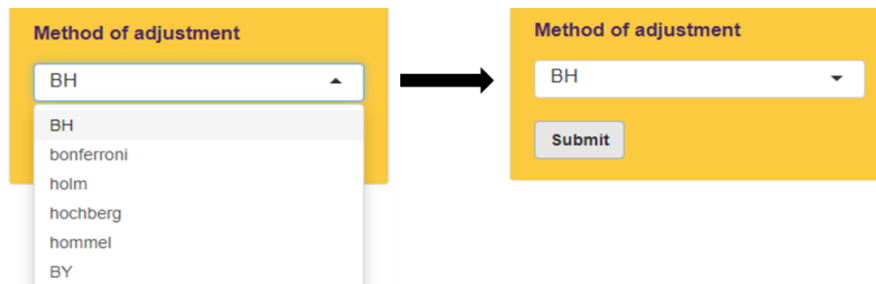


Figure 4.13 shows two panels of the 'Method of adjustment' selection interface. The left panel shows a dropdown menu with 'BH' selected and a list of options: 'BH', 'bonferroni', 'holm', 'hochberg', 'hommel', and 'BY'. The right panel shows the dropdown menu with 'BH' selected and a 'Submit' button.

**Figure 4.13.** Specify the adjustment method

After specifying all the inputs, the user has to hit the “Submit” button and wait for the results.

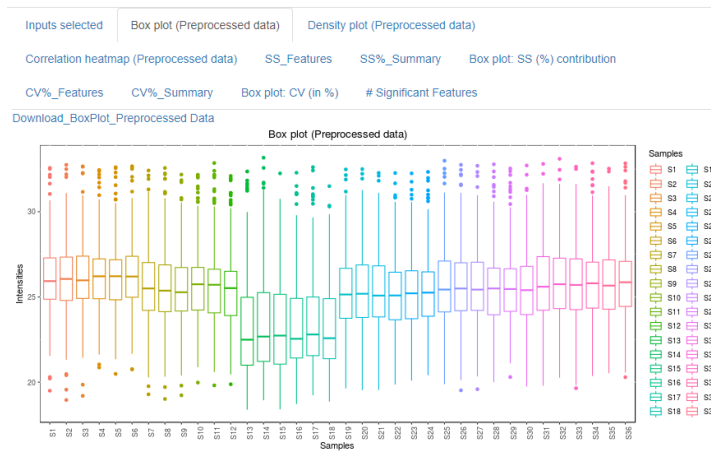
## Results obtained

1. Inputs selected: The various inputs defined by the user for the analysis can be viewed as given below.

Inputs selected	Box plot (Preprocessed data)	Density plot (Preprocessed data)	
Correlation heatmap (Preprocessed data)	SS_Features	SS%_Summary	Box plot: SS (%) contribution
CV%_Features	CV%_Summary	Box plot: CV (in %)	# Significant Features
Download_Inputs_Selected			
Input parameters	Input selected		
Categorical variable	M1, M2		
Numeric variable	Age		
Feature type	Protein		
Aggregation method	Mean		
Transformation/Normalization method	Variance Stabilizing Normalization		
Type of analysis	Imputing missing values using SVD method		
Level of significance	0.05		
Adjustment method	BH		

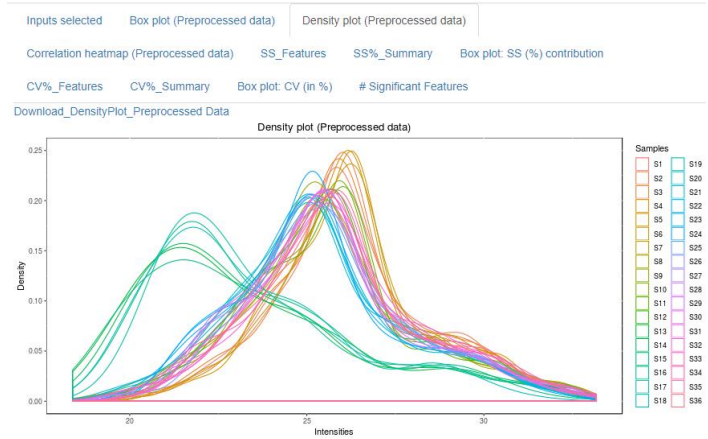
**Figure 4.14.** Inputs selected

2. Visual plots of the preprocessed data: The various exploratory plots of the preprocessed data such as box plot, density plot, correlation heatmap can be viewed under each tab as shown in Figures 4.15-4.17.

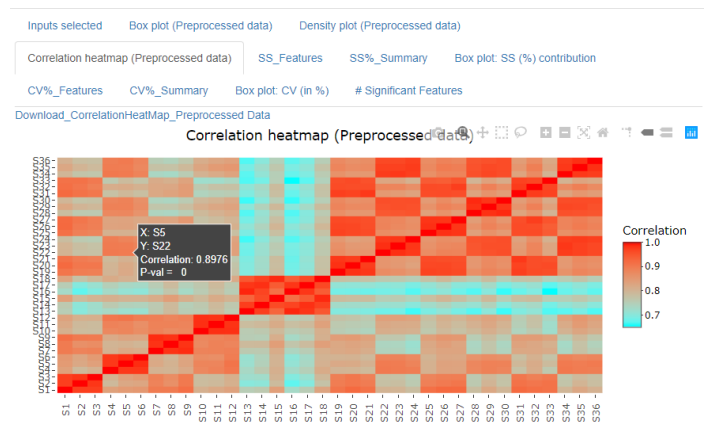


**Figure 4.15.** Box plot of preprocessed expression data





**Figure 4.16.** Density plot of preprocessed expression data



**Figure 4.17.** Interactive correlation heatmap of preprocessed expression data

From the box plot and density plot, we find that the data normalized using the “VSN” normalization method and analysis using “SVD” imputation method are normally distributed for all the samples. Here, we have normalized and imputed the data group wise.

The correlation heatmap shows correlation between the samples and the corresponding p-values.

3. The SS results: (i) The results showing the contribution of SS squares due to each variable, the p-values and the adjusted p-values corresponding to each variable are shown below. If the input is peptide data and analysis is at “Protein” level, the table will also show the number of peptides corresponding to each

protein. The complete SS result can be downloaded by clicking on the “Download\_Result\_SS\_Features” link.

Protein	# Peptides	SS_M1	SS_M2	SS_Age	SS_Residual	p-value_M1	p-value_M2	p-value_Age	adj.p-value_M1	adj.p-value_M2	adj.p-value_Age
Protein_1	6	0.9197	16.7778	0.7261	43.3444	0.4235	0.0063	0.4766	0.5261	0.0110	0.7053
Protein_10	2	12.5710	15.9741	0.7918	33.6310	0.0019	0.0024	0.3995	0.0068	0.0056	0.6793
Protein_100	5	11.3728	8.7589	0.9410	27.8581	0.0012	0.0144	0.3141	0.0049	0.0221	0.6793
Protein_101	2	9.6825	8.5872	0.3922	39.8670	0.0100	0.0486	0.5847	0.0267	0.0632	0.7259
Protein_102	2	0.4382	0.6678	0.1998	10.2582	0.2587	0.3762	0.4430	0.3449	0.4112	0.6977
Protein_103	2	68.0975	16.7087	0.5743	18.9655	0.0000	0.0001	0.3401	0.0000	0.0010	0.6793
Protein_104	2	3.5967	17.1492	1.0910	33.2955	0.0769	0.0016	0.3213	0.1292	0.0043	0.6793
Protein_105	6	14.2693	12.2656	3.9041	51.1483	0.0061	0.0357	0.1341	0.0186	0.0486	0.6793
Protein_106	8	2.1324	19.3680	2.6650	46.0743	0.2401	0.0043	0.1903	0.3289	0.0085	0.6793
Protein_107	1	0.9647	5.7582	0.1420	39.7510	0.3924	0.1228	0.7415	0.4905	0.1462	0.8104

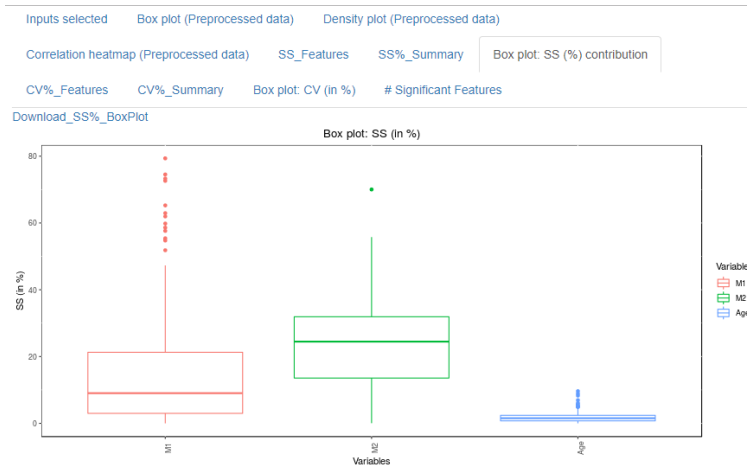
**Figure 4.18.** Contribution of SS due to each variable, the p-values and the adjusted p-values corresponding to each variable for each protein

(ii) The result summary of % SS contribution due to each variable is shown below in Figure 4.19.

Summary	M1	M2	Age
Min.	0.00	0.05	0.01
1st Qu.	2.99	13.56	0.74
Median	9.03	24.47	1.53
Mean	15.47	23.43	1.88
3rd Qu.	21.29	31.92	2.40
Max.	79.34	70.03	9.63

**Figure 4.19.** Summary of % SS contribution due to each variable

(iii) The box plot showing % contribution of SS due to each variable is given in Figure 4.20.



**Figure 4.20.** The box plot showing % contribution of SS due to each variable

From the summary and box plots, we found that the SS contribution due to the variable M2 is more than that of variable M1. The variable “Age” has the least SS contribution.

4. The CV analysis: We calculate the CV corresponding to the groups within each categorical variable. We obtained the following results: (i) CV of different groups of each categorical variable for all the proteins, (ii) Summary of CV (%) for all the proteins, and (iii) Box plot showing CV under the various groups of categorical variables. These results are shown in Figures 4.21-4.23.

Protein	# Peptides	M1_A1	M1_A2	M2_B1	M2_B2	M2_B3
Protein_1	6	7.27	1.98	2.81	2.08	7.53
Protein_10	2	7.31	2.75	2.85	1.72	9.34
Protein_100	5	6.12	1.49	2.65	4.21	6.69
Protein_101	2	4.75	4.40	4.07	4.12	6.03
Protein_102	2	2.52	2.05	2.42	1.80	2.72
Protein_103	2	6.29	0.49	2.28	7.54	9.45
Protein_104	2	6.90	1.64	2.08	0.96	7.60
Protein_105	6	7.02	5.00	7.72	5.09	5.60
Protein_106	8	7.75	2.03	3.81	1.25	7.93
Protein_107	1	1.99	5.21	0.77	6.04	2.46

**Figure 4.21.** The CV (in %) of different groups of each categorical variable for all the proteins

Inputs selected    Box plot (Preprocessed data)    Density plot (Preprocessed data)

Correlation heatmap (Preprocessed data)    SS\_Features    SS%\_Summary    Box plot: SS (%) contribution

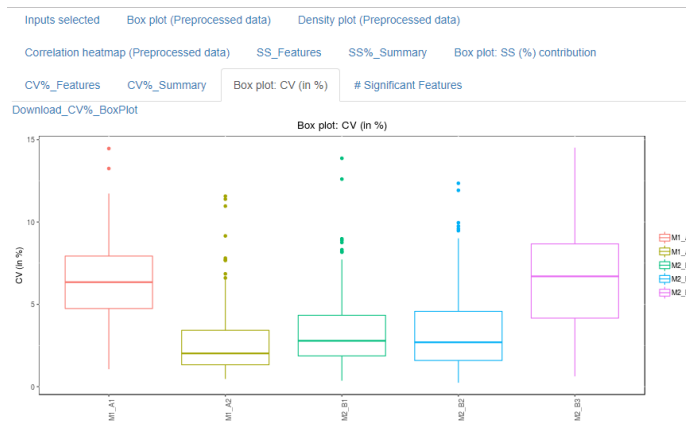
CV%\_Features    CV%\_Summary    Box plot: CV (in %)    # Significant Features

Download\_Result\_CV%\_Summary

Show 10 entries    Search:

Summary	M1_A1	M1_A2	M2_B1	M2_B2	M2_B3
Min.	1.06	0.47	0.36	0.24	0.63
1st Qu.	4.75	1.33	1.87	1.59	4.16
Median	6.35	2.02	2.79	2.70	6.70
Mean	6.34	2.64	3.33	3.26	6.61
3rd Qu.	7.93	3.42	4.34	4.57	8.67
Max.	14.46	11.57	13.87	12.35	14.52

**Figure 4.22.** Summary of CV (in %) for all the proteins



**Figure 4.23.** Box plot showing CV (in %) under the various groups of categorical variables

The summary and box plots of CV show that (i) within variable M1, A2 has lesser variability than that of A1 and (ii) within variable M2, B2 has the least variability among the three approaches of M2.

5. Number of significant features: A table showing the total number of proteins assessed and the number of proteins which have significant effect due to each variable, “M1”, “M2” and “Age”, without and with adjustment is shown below.

Inputs selected    Box plot (Preprocessed data)    Density plot (Preprocessed data)

Correlation heatmap (Preprocessed data)    SS\_Features    SS%\_Summary    Box plot: SS (%) contribution

CV%\_Features    CV%\_Summary    Box plot: CV (in %)    # Significant Features

Download\_#SignificantFeatures

Show 10 entries    Search:

Adjustment	Total # of Proteins	M1	M2	Age
Not adjusted	200	107	154	2
BH adjusted	200	82	148	0

**Figure 4.24.** Summary of significant proteins

We found greater number of proteins showing significant effects due to M2 than that of M1. This further shows that variable M2 has more significant effect than that of M1. The variable “Age” has very less effect, i.e., only two protein showed significant effect due to “Age”.

The user can download the results under each tab by clicking on the download links provided under each tab. The tables will be downloaded in “xlsx” format and the plots will be download in “png” format.

We analyzed the data at the protein level using VSN normalization and the SVD imputation method. By providing various inputs to the tool, the user gets various results. Based on the summary and box plots of SS, we found that the SS contribution due to the variable M2 is more than that of variable M1. We found that the variable “Age” has the least SS contribution. Furthermore, the summary and box plots of CV show that (i) within variable M1, A2 has lesser variability that of A1 and (ii) within variable M2, B2 has the least variability among the three approaches of M2. Therefore, we can conclude that (i) approach A2 is better that that of A1 for the method M1, (ii) approach B2 is better than those of B1 and B3 for the method M2.

## **Conclusion**

Our tool provides a user-friendly approach to standardize proteomics workflow using multiple statistical approaches. The user can identify the variable with greater variability based on SS as well as the best approach for the steps involved in the proteomics workflow based on the CV. The tool will be helpful to the researchers for designing and executing experiments.

## CHAPTER 5

### INTERACTIVE TOOL FOR STATISTICAL ANALYSIS OF LABEL-FREE LC-MS PROTEOMICS DATA CONSIDERING MISSING VALUES AND HETEROGENEITY

#### **Introduction**

Identification, quantification, and characterization of peptides and proteins from biological samples are important for understanding the molecular processes governing the cell physiology and pathophysiology [11]. With the introduction of high throughput technologies such as ultra-high-performance liquid chromatography coupled to high resolution mass spectrometers [24, 41], the high mass accuracy proteomics data can be reliably generated from samples and further processed and modeled using various statistical approaches. However, the heterogeneity in the LC-MS data due to variations in the biological samples or technical approaches can be problematic for accurate biological modeling [56]. Here, we are dealing with expression proteomics including the analysis of features (peptides and/or proteins) at large scale. Differential expression (DE) analysis of features is carried out to detect significant features in two or more conditions, such as healthy versus different disease conditions. Despite the availability of tools for analyzing proteomics data [57, 77-79], there are various statistical challenges in analyzing proteomics data, such as data heterogeneity

and missing values (MVs) [52, 55]. The biological variability among the samples and technical approaches of data generation lead to heterogeneity. Biological variability arises from genetic and environmental factors. The technical approaches such as sample extraction, storage, different batches of buffer, repeating mass spectrometer runs, etc., lead to changes in the expression data. Therefore, biological and technical variability along with other covariates such as race, gender, age, height, etc., should also be taken in account in the analysis. Furthermore, there is problem of MVs in the proteomics data that can occur due to biological and/or technical issues. There are three broad categories of MVs, namely, MCAR, MAR and MNAR [53].

Here, we introduce a user-friendly shiny tool to analyze and compare proteomics expression data scalable from small cell-culture based studies to large clinical proteomic studies using various statistical approaches. We have enabled the use of various input parameters to perform DE analysis by various approaches. Our tool will be helpful to detect differentially expressed features while considering the variability due to biological and technical replicates as well as missing observations. We have provided options to adjust the effect due to additional covariates such as age, race, etc. We have implemented the methods in R [1]. The tool has been made using “shiny” package [3]. The interactive plots were implemented using “plotly” [80].

## **Methods**

We have provided various options at each step to perform the data analysis. The two main pipeline inputs required for our platform are label-free proteomics

expression data and the additional pre-clinical or clinical information, such as patient demographic covariates. The various steps in the workflow of DE analysis of proteomics data are given below.

1. Upload the proteomics expression data
2. Select the type of feature for the analysis: The analysis can be done either at the protein or peptide level. So, the user then selects either “Protein” or “Peptide” for the analysis.
3. Select the aggregation method (Mean/Median/Sum/Max): Data aggregation is required if the user has provided the peptide data and wants analysis at the protein level. It is also applicable to other situations, such as when the features are redundant.
4. Upload the additional information of the data: The additional information of the data such as samples, groups, biological samples, etc., is required for the analysis. In complex experiments, there can be other independent variables present such as run, biological replicate, gender, age, etc. We have provided the user a method to include these variables in such situation. After the file is uploaded, the user must specify the nature of variables.
5. Select the categorical fixed effect: The fixed effects are the effects that remain constant across individuals. Here, the user has to select the categorical fixed effects such as groups, genotype, race, gender, etc.
6. Select the numeric (continuous) fixed effect: The user has to select the numeric (continuous) fixed effects such as age, height, etc.



7. Select the random effects: The random effects are the effects that are random and unpredictable. The random effects cannot be controlled by the experimenter, e.g., MS run, biological replicates, technical replicates, etc.
8. Select the categorical fixed effect of interest: The user has to specify the variable of interest which is considered to have a fixed effect.
9. Select a comparison of interest: After selecting the categorical fixed effect of interest, all the available pairwise comparisons will appear in the drop-down menu. The user can select one comparison at a time.
10. Choose the method of analysis:
  - (i) Excluding MVs: We retain the features having complete observations for all the samples. The features with MV in any of the samples are discarded from the analysis. However, this approach is generally not preferred as it leads to exclusion and loss of various features. Therefore, we have provided the option of imputing MVs.
  - (ii) Imputing MVs: We provide a hybrid imputation method of the R package “imputeLCMD” [62] that assumes the MVs are both MAR and MNAR. We have given the users two different options for data imputation under the assumption of MAR or MCAR, (a) Singular value decomposition [64] and (b) K-nearest neighbor [65, 66]. The MNAR assumption uses a quantile regression method for the imputation of left-censored missing data in quantitative proteomics. The MVs are imputed for each group separately after normalizing the data [61].
11. Choose the transformation and/or normalization methods: Data transformation and/or normalization is required to achieve consistency across the

samples. The normalization is done group-wise, based on the variable of interest.

We have provided the following options:

(i) Logarithmic transformation: The raw data are transformed to log base 2.

(ii) Quantile normalization: The raw data are transformed to log base 2 followed by QN using the function “normalize.quantiles” [67] of the R package “preprocessCore” [68].

(iii) Variance stabilizing normalization: The raw data are normalized by using the “justvsn” function of the R package “vsn” [69], separately for each group.

(iv) None: The users can use their own normalized data in place of raw intensity data.

12. Method of DE analysis: We have provided various options to detect differentially expressed features. It is assumed that the data follow normal distribution after applying the data transformation and/or normalization method. The user must select the appropriate test, depending on the experimental design, which are discussed below.

(i) LIMMA/Moderated t-test: We have used “limma” R package [81]. This is the most robust statistical analysis method. We fit the linear model using “lmFit” function for each feature. The moderated t-statistics and the coefficient estimates are computed by using empirical Bayes (eBayes) method [82, 83]. The feature-wise residual variances are squeezed toward a common value using eBayes method. However, this method can handle only a single random effect.

(ii) Linear fixed or mixed model approach: A general linear approach with fixed and random effects is a much more flexible and powerful technique that can be

applied to more complex designs. If random effects have been specified by the user, then we fit the linear mixed-effects models using R package “lme4” [84]. We have used R package “afex” for estimating the mixed models and calculating the p-values of fixed effects [85]. This option can also handle more than one variable having random effects. In the absence of any random effect, we fit a linear model using the “lm” function [1]. The contrasts of the estimated marginal means for linear and mixed models are computed using R package “emmeans” [86].

(iii) Pairwise t-tests: On clicking the radio button “T-test”, a dropdown menu showing three different types of t-test will appear. The three types of t-test are given below:

(a) Two sample t-test assuming equal variances: This option performs t-test with the assumption that the two populations have equal variances with equal and unequal samples sizes.

(b) Two sample Welch’s t-test assuming unequal variances: The Welch’s t-test is used when the population variances are assumed to be unequal. The sample sizes may be equal or unequal.

(c) Paired t-test: This option performs the paired t-test. This test is used to compare two population means from the same population at two different times (repeated measures) or to compare two population means from different populations in which the observations have been matched or “paired”.

If the user chooses options, t-test or Welch’s t-test, the test will consider only the fixed effect of interest. The analysis will not control the effects of other variables.

If the user wants to control the effects due to other variables, then the user must choose the options (i) LIMMA/moderated t-test or (ii) linear fixed/mixed model approach. If there is only one fixed effect and no other covariates, then the first option gives the result based on moderated t-test, whereas the second option gives the result equivalent to Welch's t-test.

13. Select the significance level: By default, it is 0.05. However, the user can specify any cut-off between 0 and 1.

14. Desired  $\log_2$  fold change (FC) cut-off: We have assumed the data is approximately on log 2 scale for easier interpretation of the results. The user can specify a desired  $\log_2$  FC cut-off. By default, the value of  $\log_2$  FC is 1, which means a doubling in intensity (abundance values).

15. Method of adjustment: We have provided several methods of adjusting p-values for testing multiple features available in R [1]. These are Benjamini-Hochberg (BH), Bonferroni, Holm, Hochberg, Hommel, and Benjamini-Yekutieli (BY). The BH method is the default adjustment method.

After specifying all the input parameters, the user has to hit the "Submit" button to get the results. The results are displayed under each tab. The tab "Inputs selected" contains the various input parameters provided by the user. We provide various exploratory plots such as box plot, density plot, correlation heatmap and multidimensional scaling (MDS) plot. The summary of differentially expressed features can be viewed under "Summary" tab. The user can view the result of DE analysis under the tab "Table of differentially expressed features". We provide interactive volcano plots for both with and without adjustment. All the

results can be download by clicking on the download links provided under each tab. The summary and table of differentially expressed features are downloaded in “xlsx” format. By clicking on the download link under “Table of differentially expressed features”, the user will obtain the DE analysis result, complete results having the overall F-statistic, p-value and other values for different contrast or comparison, and the preprocessed data. All the plots are downloaded in “png” format.

### **Demonstration and results**

To demonstrate our web-based application, we generated a test proteomics expression data set abstracted from locally available clinical proteomics data sets consisting of 200 proteins corresponding to 1000 peptides. The additional data information file contains the information with headings, “Samples”, “Group”, “Race”, “Bio”, “Run” and “Age”. The “Group” is the fixed effect of interest. There are three groups, namely, control, case1 and case2, each having three biological replicates with two MS runs. Thus, there are six samples in each group leading to total 18 samples. After specifying the input parameters and submitting the job, we obtained various results such as results summary, result showing differentially expressed features, graphical plots (box plots, density plots, correlation heatmap, MDS plot, volcano plots). The steps involved in the analysis with screenshots are as follows.

**SATP: Statistical Analysis Tool for Proteomics**

**Choose file to upload expression data**

Browse... No file selected

**Select feature type**

\_\_\_\_\_

**Aggregation method**

Mean

**Choose file to upload additional information**

Browse... No file selected

**Select categorical fixed effect(s)**

\_\_\_\_\_

**Select continuous fixed effect(s)**

\_\_\_\_\_

**Select random effect(s)**

\_\_\_\_\_

**Select categorical fixed effect of interest**

\_\_\_\_\_

**Select comparison of interest**

\_\_\_\_\_

**Choose analysis method**

Excluding missing values

Imputing missing values

**Select imputation method**

SVD

**Choose transformation/normalization method**

Log base 2

Quantile Normalization

Variance Stabilizing Normalization

None

**Choose statistical testing method**

Test based on LIMMA/ Moderated t-test

Test based on linear fixed or mixed model

T-test

**Specify level of significance**

0.05

**Specify log fold change**

1

**Select adjustment method**

BH

Submit

Inputs selected | Box plot | Density plot | Correlation heatmap | MDS plot | Summary | Table of DE features

Volcano plot (Not adjusted) | Volcano plot (Adjusted)

---

Download\_Inputs\_Selected

**Figure 5.1.** Webpage of the tool “SATP”

## Input specifications

We have demonstrated the webtool using a test proteomics dataset. The inputs to be provided by the user are as follows:

1. Upload the proteomics expression data: The first step is to upload the proteomics expression data either in csv, txt, tsv, xls or xlsx format. The first two columns are reserved for proteins and peptides. If the data are available at the protein level only and there are no peptide data, then user must leave the second column blank. The expression data must start from the third column and onwards. The first row must contain the labels such as “Protein”, “Peptide” followed by the sample names (starting from third column). A portion of input expression data is shown below.

Protein	Peptide	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
Protein_1	NTHVFSF	NA	143000	230000	550000	1662000	426000	195000	669000	NA	96000	334000	NA	1055000	248000	1519000	963800	505200	NA
Protein_2	SHYPCLHRML	NA	829000	1685000	1386000	1627000	2468000	2322000	3894000	NA	NA	NA	NA	1464000	715000	1197000	NA	1437000	1274000
Protein_2	DIHMQYSPQVDWHEIRWY	1813000	1783000	1371000	1201000	NA	NA	NA	4454000	4206000	2037000	3552000	NA	NA	NA	1127000	NA	NA	NA
Protein_2	LIKYDLARASNE	280000	527000	550000	1052000	638000	NA	1865000	3826000	2997000	1104000	3386000	486000	346000	NA	353000	NA	NA	NA
Protein_3	HVGMCKVVEYHRKWL	NA	NA	1356000	1148000	714000	1584000	NA	3754000	NA	2489000	4213000	1658000	911000	1533000	929000	1364000	963000	NA
Protein_3	YFPFCWKWRMY	1073000	NA	NA	NA	2850000	NA	1255000	NA	788000	NA	NA	3367000	NA	1273000	1993000	1286000	NA	4031000
Protein_3	WNSDYKTAYGMVAISMOK	NA	532200	46000	NA	1840000	1841000	1223000	NA	194000	749400	252000	3272000	1213000	1953000	NA	332000	1530000	4352000
Protein_3	DQQLPKQN	1222000	NA	836000	1696000	2595000	2897000	NA	929000	1451000	1414400	1053000	3459000	1728000	2303000	3018000	999000	2467000	5229000
Protein_3	RADFFIIGPYNCFYR	858000	NA	887000	1904000	NA	2578000	1766000	1835000	NA	NA	1260000	3573000	2001000	NA	NA	1601000	2321000	NA
Protein_3	TWWQSSNKSRVLDLPWL	1306000	1501200	1131000	1359000	NA	2855000	1330000	1306000	946000	1003400	1075000	3703000	2072000	2699000	3238000	1774000	2387000	5377000
Protein_3	VVWTEICQHWFLMLGPH	1346000	1269200	1212000	1946000	2809000	3269000	1378000	1533000	1773000	1116400	1468000	4614000	2457000	2022000	2920000	2145000	NA	5070000
Protein_3	VRREWHEKQNCPHNE	NA	NA	NA	747000	NA	NA	NA	851000	384000	NA	846000	NA	NA	1701000	NA	NA	NA	4650000
Protein_3	WHVCNPPQGFVFIQ	NA	NA	873000	1664000	NA	NA	NA	1606000	1838400	1078000	4122000	NA	2154000	2997000	NA	2556000	4518000	
Protein_3	ERVLNY	314000	298200	NA	NA	2736000	2095000	1423000	NA	NA	NA	NA	NA	NA	2513000	447000	NA	4765000	
Protein_3	WIMQSHMEFWHQFKPA	1057000	NA	NA	NA	2485000	NA	1715000	1687000	1192000	NA	NA	NA	NA	3207000	1158000	2620000	5391000	
Protein_4	LDPAHLG	2646000	224200	1267000	NA	NA	997000	121068000	303205000	143301000	252317000	NA	41000	NA	NA	839000	NA	578000	52000
Protein_4	ILEVTNC	NA	NA	1974000	NA	1294000	1208000	122378000	NA	144550000	NA	NA	1417000	NA	1009000	1896000	1188000	1507000	NA
Protein_4	HATATQS	3403000	NA	NA	NA	399000	1135000	122008000	NA	143648000	252576000	NA	NA	NA	340000	NA	1348000	NA	772000
Protein_4	DWMPHRC	3661000	1176200	1956000	1428000	809000	1327000	122149000	303744000	144581000	253577000	284023000	817000	746000	1057000	1284000	1309000	970000	1448000
Protein_4	NGWGNHEECLPISK	3847000	1012200	1933000	1812000	954000	1281000	NA	304229000	NA	253031000	NA	1171000	1085000	1687000	1648000	1056000	NA	1222000

Figure 5.2. A portion of proteomics expression data

The user has to click on the “Browse...” button for selecting the expression data file as given below in Figure 5.3.

**Choose file to upload expression data**

Browse... ProteomicsData\_SATP.xlsx

Upload complete

Select feature type

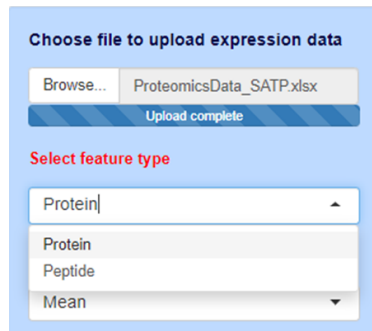
Protein

Aggregation method

Mean

Figure 5.3. Upload the proteomics expression data

2. Select the type of feature for the analysis: The feature type available will automatically be detected after uploading the expression data file. There are two options available: “Protein” or “Peptide”. We selected the analysis to be done at “Protein” level as given below.



Choose file to upload expression data

Browse... ProteomicsData\_SATP.xlsx

Upload complete

Select feature type

Protein

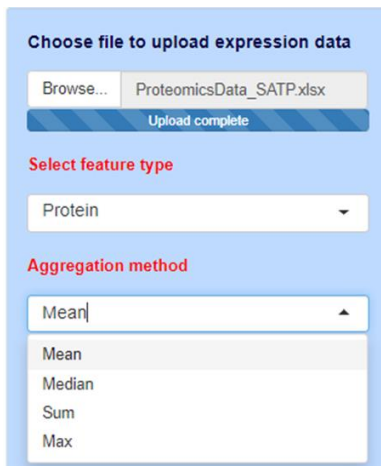
Protein

Peptide

Mean

**Figure 5.4.** Choose the type of feature– “Protein” or “Peptide”

3. Select the aggregation method: There are four options available for data aggregation (Mean/Median/Sum/Maximum). We selected “Mean” for aggregating the peptide data at protein level as given below.



Choose file to upload expression data

Browse... ProteomicsData\_SATP.xlsx

Upload complete

Select feature type

Protein

Aggregation method

Mean

Mean

Median

Sum

Max

**Figure 5.5.** Choose the aggregation method

4. Upload the additional information: Now the user has to upload the additional information about the data either in csv, txt, tsv, xls or xlsx format. The first row



must contain the labels of various information. The first column contains the sample names. The screen shot of the additional data is shown below.

	A	B	C	D	E	F	G
1	Samples	Group	Gender	Race	Bio	Run	Age
2	S1	Control	Male	White	1	1	36
3	S2	Control	Female	White	2	1	41
4	S3	Control	Male	Black	3	1	39
5	S4	Control	Male	White	1	2	36
6	S5	Control	Female	White	2	2	41
7	S6	Control	Male	Black	3	2	39
8	S7	Case1	Female	White	4	1	46
9	S8	Case1	Male	White	5	1	41
10	S9	Case1	Male	White	6	1	42
11	S10	Case1	Female	White	4	2	46
12	S11	Case1	Male	White	5	2	41
13	S12	Case1	Male	White	6	2	42
14	S13	Case2	Male	White	7	1	36
15	S14	Case2	Male	Black	8	1	45
16	S15	Case2	Male	Black	9	1	34
17	S16	Case2	Male	White	7	2	36
18	S17	Case2	Male	Black	8	2	45
19	S18	Case2	Male	Black	9	2	34

**Figure 5.6.** Additional information of data

In the given example, labels in the first row are: “Samples”, “Group”, “Gender”, “Race”, “Bio”, “Run”, “Age”. There are three groups (column with label “Groups”), namely, “Control”, “Case1” and “Case2”, each having three biological replicates (column with label “Bio”) with two MS runs (column with label “Run”). There are six samples in each group leading to total 18 samples (column with label “Samples”). There are three additional covariates (gender, race and age) in the data. The covariates gender (column with label “Gender”) and race (column with label “Race”) are categorical fixed effects, each having two levels. Gender has two levels: “Male” and “Female”. Race has two levels: “White” and “Black”. The covariate age (column with label “Age”) is continuous/numeric fixed effect.

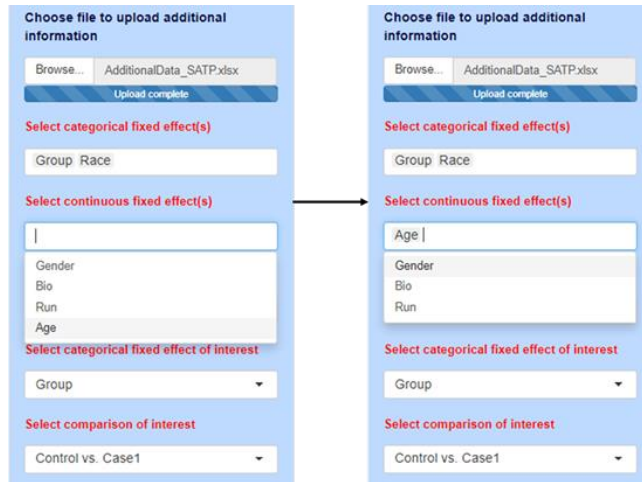
The file can be uploaded by first by clicking on the “Browse...” button and then selecting the file as given in Figure 5.7.

**Figure 5.7.** Upload the file with additional information of data

5. Select the categorical fixed effect: The user has to select the categorical fixed effects one by one which will automatically pop out after uploading the file containing additional information. We have selected “Group” and “Race” as the categorical fixed effects as given below in Figure 5.8.

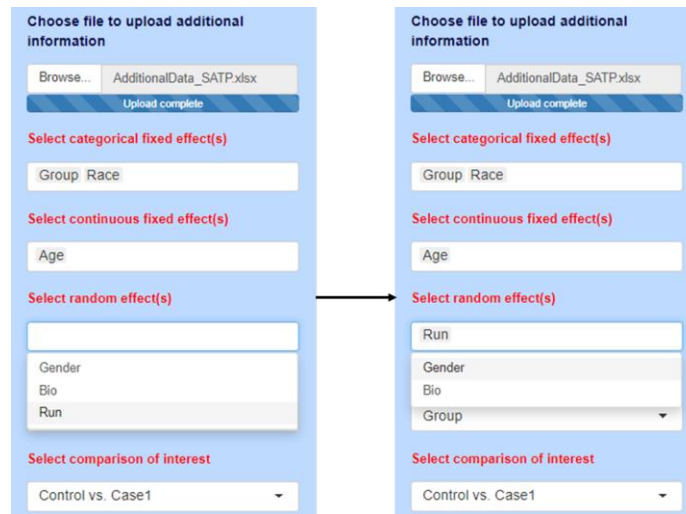
**Figure 5.8.** Selection of categorical fixed effects

6. Select the continuous fixed effect: After selecting the categorical fixed effects, the user can now select the continuous variable from the remaining variables, if available. We have selected “Age” in the given example (Figure 5.9).



**Figure 5.9.** Selection of continuous variables

7. Select the random effects: After selecting the fixed effects, the user has to select the random effects, if available. We have selected “Run” as random effect as given below in Figure 5.10.



**Figure 5.10.** Selection of random effects

8. Select the categorical fixed effect of interest: The user has to specify the variable of interest which is considered to be having a fixed effect. We have selected “Group” as the variable under study (Figure 5.11).



**Figure 5.11.** Selection of categorical fixed effect of interest

9. Select a comparison of interest: We selected “Control vs. Case1” from the list of all available pairwise comparisons in the drop-down menu (Figure 5.12). “Control vs. Case1” means “Case1” is compared to “Control”.



**Figure 5.12.** Selection comparison of interest

10. Choose method of analysis: We have provided two options for the analysis: (i) Excluding missing values, and (ii) Imputing missing values (default option). There are two methods of data imputation available: (a) SVD, and (b) KNN. We selected the radio button “Imputing missing values” and “SVD” method for data imputation. The screenshots are given below in Figure 5.13.



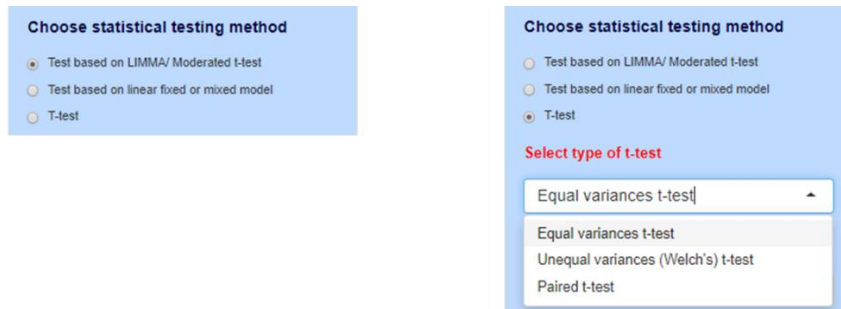
**Figure 5.13.** Select analysis method

11. Choose the transformation and/or normalization methods : There are four options available for data transformation and/or normalization (log2/QN/VSN/None). We selected “Quantile Normalization” for data normalization as given below.



**Figure 5.14.** Selection of normalization method

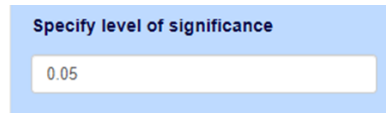
12. Method of DE analysis: We provide: (i) LIMMA/ Moderated t-test, (ii) linear fixed or mixed model approach, and (iii) various forms of t-test: The LIMMA method provides the most reliable and robust statistical test. Therefore, we have made this option as the default method. On clicking the radio button “T-test”, a dropdown menu showing three different types of t-test will appear as given below in Figure 5.15.



**Figure 5.15.** Selection of statistical testing method

We have selected the default method (LIMMA/Moderated t-test) for the demonstration.

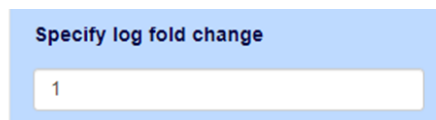
13. Select the significance level: The user has to specify the level of significance. We have selected the default value 0.05 as the significance level (Figure 5.16).



A screenshot of a web form titled "Specify level of significance". It features a light blue header with the title in bold. Below the header is a white text input field with a thin border, containing the value "0.05".

**Figure 5.16.** Specify the level of significance

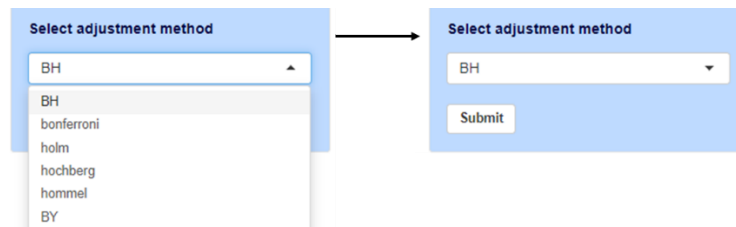
14. Desired log<sub>2</sub> fold change (FC) cut-off: The user can specify a desired log<sub>2</sub> FC cut-off. We have specified the default value of log<sub>2</sub> FC as 1 (Figure 5.17).



A screenshot of a web form titled "Specify log fold change". It features a light blue header with the title in bold. Below the header is a white text input field with a thin border, containing the value "1".

**Figure 5.17.** Specify the desired log fold change

15. Method of adjustment: The user has to select the method of adjusting the p-values for multiple testing of features. We have provided six adjustment methods. We selected “BH” adjustment method (Figure 5.18).



Two screenshots of a web form titled "Select adjustment method". The left screenshot shows a dropdown menu with "BH" selected and a list of other methods: "BH", "bonferroni", "holm", "hochberg", "hommel", and "BY". An arrow points to the right screenshot, which shows the same dropdown menu with "BH" selected and a "Submit" button below it.

**Figure 5.18.** Specify the adjustment method

## Output specifications

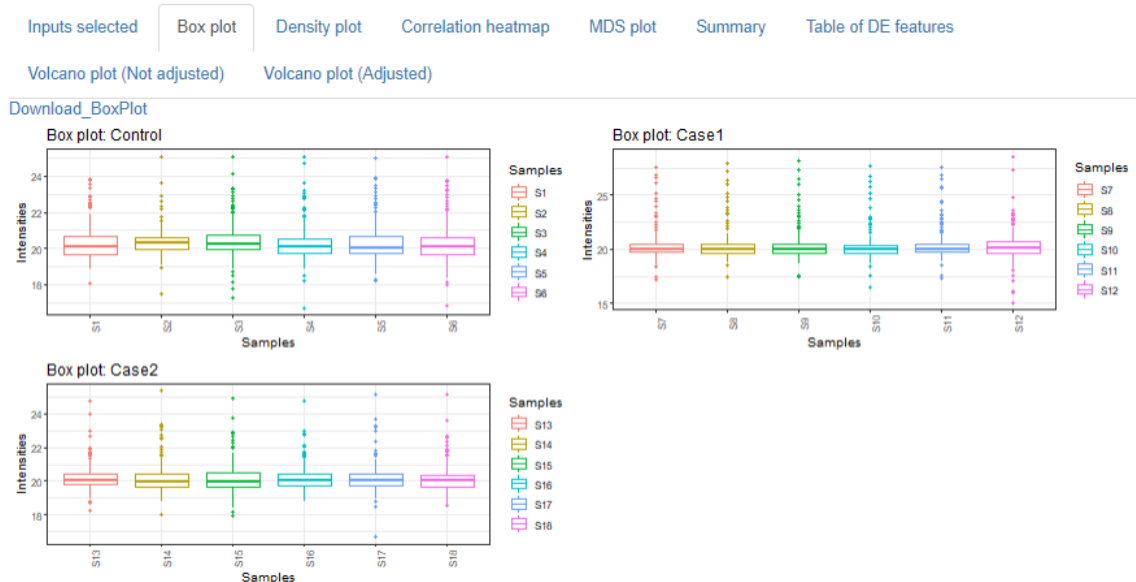
After specifying all the inputs, the user has to hit the “Submit” button and wait for the results. The results are displayed by clicking on the respective tabs. The screenshots for all the results are as follows.

1. Inputs selected: The various inputs defined by the user for the analysis can be viewed under the tab “Inputs selected” as given in Figure 5.19.

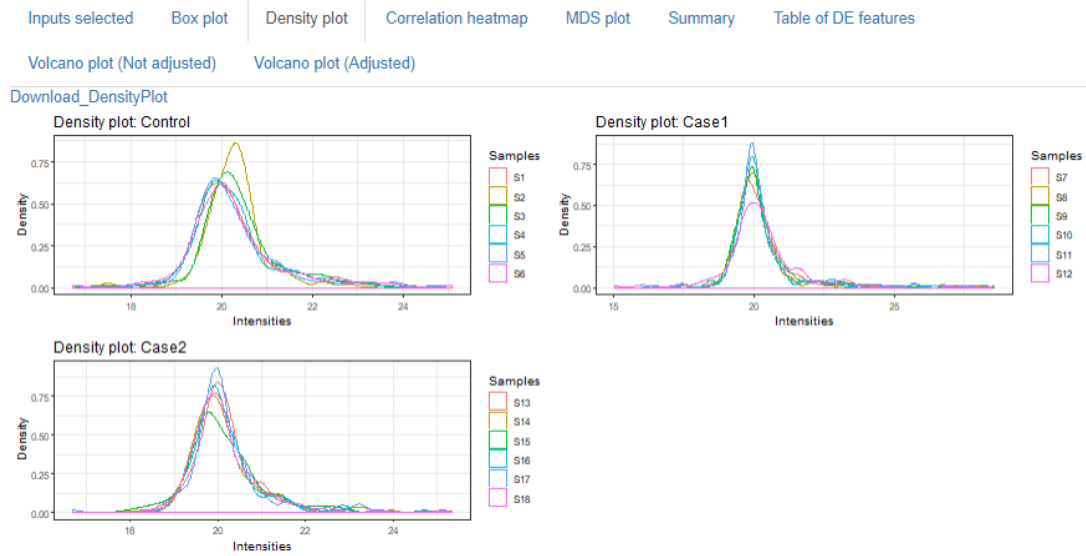
Inputs selected	Box plot	Density plot	Correlation heatmap	MDS plot	Summary	Table of DE features
Volcano plot (Not adjusted)		Volcano plot (Adjusted)				
<a href="#">Download_Inputs_Selected</a>						
Input parameters	Inputs selected					
Feature type	Protein					
Aggregation method	Mean					
Categorical fixed effect(s)	Group, Race					
Continuous fixed effect(s)	Age					
Random effect(s)	Run					
Categorical fixed effect of interest	Group					
Comparison of interest	Control vs. Case1					
Analysis method	Imputing missing values using SVD method					
Normalization method	Quantile Normalization					
Statistical testing method	Test based on LIMMA/ Moderated t-test					
Desired logFC	1					
Level of significance	0.05					
Adjustment method	BH					

**Figure 5.19.** Inputs selected

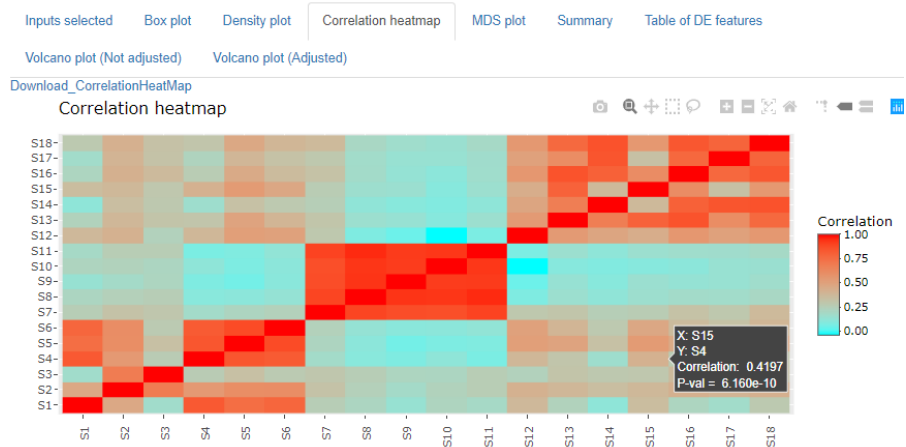
2. Visual plots of the preprocessed data: Various exploratory plots of the preprocessed data such as box plots, density plots, correlation heatmap and MDS plot can be viewed under their respective tabs as shown in Figures 5.20-5.23.



**Figure 5.20.** Box plots of preprocessed expression data for different groups



**Figure 5.21.** Density plots of preprocessed expression data for different groups



**Figure 5.22.** Interactive correlation heatmap of preprocessed expression data



**Figure 5.23.** Interactive MDS plot



3. Differential expression analysis results: Summary of results showing the total number of features (proteins or peptides) analyzed, number of differentially expressed features, number of differentially expressed features between desired log FC cutoffs, number of upregulated and downregulated features will be obtained under “Summary” tab. The summary will be for both adjusted and not adjusted. An example is shown below in Figure 5.24.

	Not.adjusted	Adjusted
# of proteins analyzed	200	200
# of DE proteins	67	39
# of DE proteins (-1 < log <sub>2</sub> FC < 1)	31	10
# of up-regulated proteins (log <sub>2</sub> FC ≥ 1)	16	14
# of down-regulated proteins (log <sub>2</sub> FC ≤ -1)	20	15

**Figure 5.24.** Summary of result

The result of DE analysis of the features can be viewed under the tab “Table of DE features” (Figure 5.25). It will show the result for each feature analyzed. We have analyzed the data at “Protein” level using the SVD imputation method. Therefore, the table shows the names of proteins in first column and the number of peptides belonging to a protein in second column. If the data are analyzed at the “Peptide” level, then the first column contains the protein names and the second column contains the peptide sequence. The table also shows the percent of MVs in each group, e.g., “MV (%) Control” and “MV (%) Case1”. If the user chooses the method excluding MVs, then the percent of MVs will not appear. The table also shows the estimate (equivalent to log<sub>2</sub> FC), t-value, degree of freedom (df), p-values without adjustment and adjusted p-values for each feature. The user can download the results by clicking on the download link

button. The table of differentially expressed features, complete results and preprocessed data will be downloaded in zip format. If more than two groups are present, then the complete results will have overall F-value, F statistic (without and with adjustment), DE analysis results based on all pairwise contrast or comparison.

Inputs selected   Box plot   Density plot   Correlation heatmap   MDS plot   Summary   Table of DE features

Volcano plot (Not adjusted)   Volcano plot (Adjusted)

Download\_Result\_DE\_Analysis

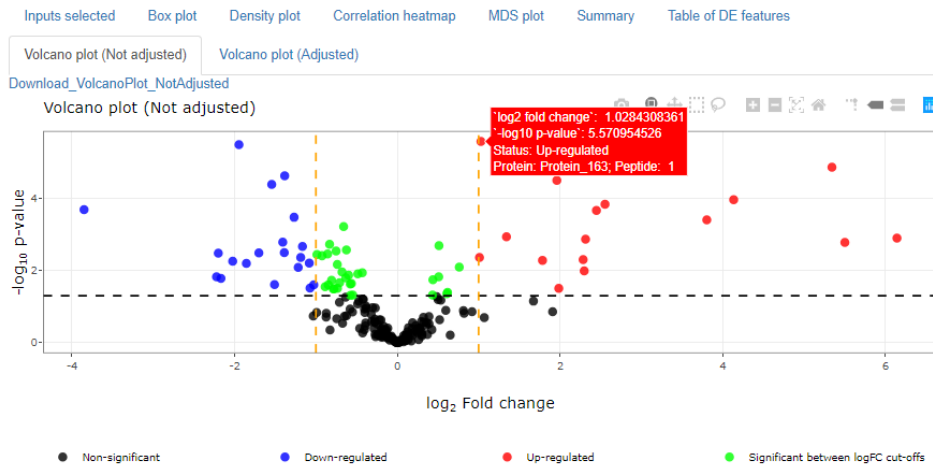
Show 10 entries   Search:

Protein	# Peptides	MV (%) Control	MV (%) Case1	logFC	t	df	p-value	adj.p-value
Protein_1	1	16.67	33.33	-0.99	-1.53	13	0.1468	0.3262
Protein_10	10	33.33	35.00	1.07	1.34	13	0.2008	0.4016
Protein_100	10	35.00	40.00	0.30	0.47	13	0.6455	0.8018
Protein_101	3	38.89	22.22	0.04	0.20	13	0.8457	0.9293
Protein_102	4	41.67	12.50	-2.22	-2.74	13	0.0151	0.0655
Protein_103	1	16.67	33.33	0.13	0.50	13	0.6273	0.7967
Protein_104	1	16.67	16.67	0.04	0.11	13	0.9153	0.9634
Protein_105	10	33.33	43.33	-0.40	-1.67	13	0.1156	0.2818
Protein_106	4	25.00	33.33	2.31	3.90	13	0.0014	0.0170
Protein_107	5	20.00	23.33	-0.02	-0.12	13	0.9027	0.9553

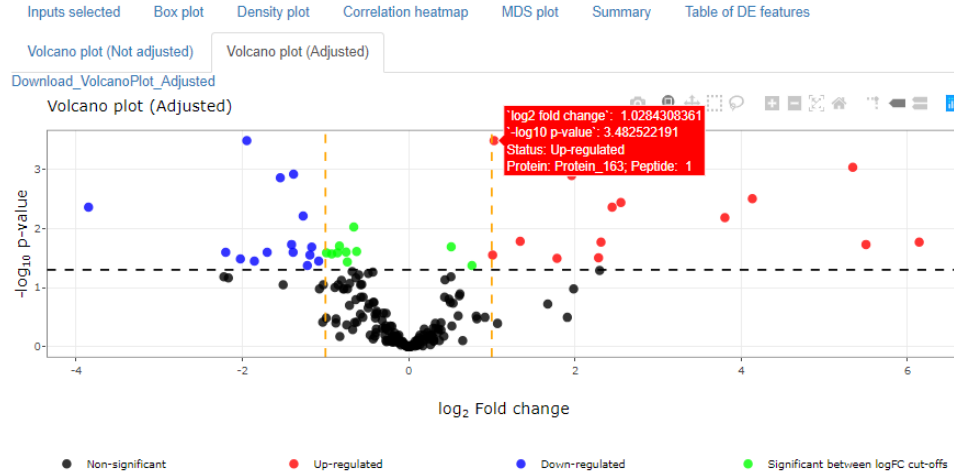
Showing 1 to 10 of 200 entries   Previous   1   2   3   4   5   ...   20   Next

**Figure 5.25.** Result of differential expression analysis

The interactive volcano plots without adjustment and with adjustment can be viewed respectively under the tabs “Volcano plot (Not adjusted)” and “Volcano plot (Adjusted)” (Figures 5.26 and 5.27).



**Figure 5.26.** Volcano plot without adjustment



**Figure 5.27.** Volcano plot with adjustment

All the results can be downloaded by clicking on the download link button provided under each tab. The table results will be download in xlsx format. The visual plots will be download in png format. Some results will be download in zip format. The user has to unzip the files and the individual files in xlsx or png format can be viewed separately.

Here, we analyzed the data at protein level. We used the mean of expression values of peptides corresponding to a protein. We compared the groups “Control vs. Case1” for differential expression analysis. We adjusted the effect due to race (categorical) and age (continuous). The run effect is considered as the variable having random effect. We normalize the data using QN and imputed the data using “SVD” method. We used the LIMMA method with desired log<sub>2</sub> FC 1 at the significance level 0.05. We used the “BH” method for adjusting multiple testing of proteins. The distribution of the expression data of different samples in each group can be examined by exploratory plots such as box plots (Figure 5.20) and density plots (Figure 5.21). The interactive correlation heatmap (Figure 5.22) shows the correlation coefficients and p-values of

correlation for all possible pair of samples. The MDS plot allows the user to find the relationship among the samples based on the group. In Figure 5.23, the samples are clustered well together in each group except for one sample in “Case1” group. The summary of DE analysis is given in Figure 5.24. The number of differentially expressed proteins without adjustment is 67 out of total 200 proteins. Without adjustment, there are 16 upregulated and 20 downregulated proteins at  $\log_2$  FC cut-off of  $\pm 1$ ; and 31 significant proteins between the  $\log_2$  FC cut-offs. However, the p-values need to be adjusted for testing the multiple proteins. Therefore, with adjustment, we found only 39 proteins to be significant (14 upregulated, 15 downregulated and 10 significant between the  $\log_2$  FC cut-offs). The result of DE analysis for each protein can be viewed and a portion of the result is given in Figure 5.25. The volcano plot is generally used to display the result of DE analysis. The interactive volcano plots without and with adjustment are given in Figure 5.26 and 5.27 respectively. The most upregulated proteins are towards the right (red color) and the most downregulated proteins are towards the left (blue color) with the most statistically significant proteins are towards the top (above the dotted horizontal line). The non-significant proteins are towards the down (black color). The plot also displays the significant proteins between the  $\log$  FC cut-offs (green color). All the results can be viewed and downloaded.

## **Discussion**

Our tool is a valuable source for analyzing proteomics expression data even in the presence of MVs and accommodating complex experimental design. We

have fully tested our tool for proper function. The user can perform the analysis interactively and can download the results for each comparison. In case of more than two groups, the user can download the complete results with DE analysis results for all possible pairwise comparisons. We compared our tool “SATP” with the existing tools, RepExplore [57] and MSqRob [79] for the DE analysis of proteomics expression data. A brief comparison among the tools is given below:

**Table 5.1.** Comparison among the tools: SATP, RepExplore and MSqRob

	SATP	RepExplore	MSqRob
Ability to handle MVs	Yes	No	No
Ability to compare more than two groups	Yes	No	Yes
Ability to adjust additional covariates	Yes	No	No

As compared to the existing tools, our tool has several advantages. The first advantage is that the tool can analyze data having missing observations while considering the heterogeneity due to biological and technical replicates. We also provide the percentage of MVs for each feature for various groups under comparison in the results obtained using imputation method. This will be helpful for deciding whether the feature is significant or not. The second advantage is that we have provided robust statistical methods such as LIMMA and linear fixed/mixed models that can accommodate complex experimental design. The user can also control the effects of additional covariates such as gender, age, height, etc. Our tool can analyze the data for two or more than two groups. The third advantage is the user can analyze the data both at the protein and peptide levels.

## **Conclusion**

Our tool will be a useful resource for the researches working in the field of proteomics and bioinformatics. We have provided different ways to analyze proteomics abundance data. Furthermore, this can be used to analyze data from similar experiments with expression values (e.g., microarray and metabolomics data).

## CHAPTER 6

### SAMPLE SIZE ESTIMATION FOR HETEROGENEOUS PROTEOMICS EXPERIMENTS USING STATISTICAL AND COMPUTATIONAL APPROACHES

#### **Introduction**

Proteomics studies are carried out on large scale and designed to address both qualitative and quantitative aspects of the proteome [11, 12]. The proteomics experiment can lead to identification of novel biomarkers which are measurable indicator of some biological state or condition [87]. The biomarker discovery will lead to better understanding of the biological or physiological process such as mechanism of disease. Despite the major advances in proteomics and bioinformatics approaches, still there are limitations and challenges in the experimental design.

Design and sample size estimation are important for carrying out proteomics experiments. Various studies have been done with respect to the design, power analysis and sample size calculation for proteomics experiments [51, 88-91]. The sample size required in a study depends on various constraints such as data availability, budget, support facilities, time requirement, etc. Sample size can be estimated by either using simulation methods or using pilot data or using similar data sets. However, the proteomics data obtained from proteomics experiments have a lot of missing values (MVs) and are highly heterogeneous. In previous chapters, we have suggested the use of the imputation methods for

better analysis of proteomics experiments. In this chapter, we have provided various statistical approaches for sample size calculation.

In Chapter 5, we have developed a tool for differential expression analysis of proteomics experiment. In this chapter, we developed sample size calculation methods to test the significance of features for quantitative proteomics expression data. Sample size calculation for testing the significance of features between two groups is based on Welch's t-test [92, 93]. We have implemented all the methods in R [1] and we have developed user-friendly shiny apps [3] for estimating sample size for proteomics experiment under allocation and cost constraints.

In Chapters 3 and 4, we studied and implemented various approaches of standardizing proteomics workflow for LC-MS data. In the last section of this chapter, we studied the impact of technical variability on the study design for proteomics experiment. The sample size calculation is based on the coefficient of variation (CV) [94].

### **Sample size calculation for detecting differentially expressed features between two classes**

The sample size and cost estimation are important for carrying out the experiments successfully. Our method of sample size calculation is based on Welch's t-test for comparing means between two groups (or classes) [92]. We have used the general methods of optimal sample sizes for Welch's test given by Jan and Shieh [93]. The methods were modified and extended to estimate



sample size for proteomics experiment involving multiple features. These methods are discussed below:

**A. Sample size calculation for comparing means between two groups for a single feature**

The study aims at class comparison, that is, detecting features which significantly differ in abundance between two groups. We construct the hypothesis setting for testing group effect, i.e., whether a feature is differentially expressed between two groups, as given below:

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_1: \mu_1 \neq \mu_2$$

where,  $\mu_1$  and  $\mu_2$  are the population means corresponding to group 1 and group 2 respectively.

Alternatively, the above setting can be written as

$$H_0: \mu_d = 0 \text{ vs. } H_1: \mu_d \neq 0$$

where,  $\mu_d = \mu_1 - \mu_2 = 0$ .

The outcomes of testing the null hypothesis belong to one of the four scenarios as given below in Table 6.1.

**Table 6.1.** The outcomes of testing null hypothesis vs. alternative hypothesis

		Null hypothesis (equal abundance)	
		True equal abundance	False equal abundance
Decision about null hypothesis	Fail to reject (abundances are equal)	Correct decision ( $1 - \alpha$ ) True Negative	Type II error ( $\beta$ ) False Negative
	Reject (abundances are unequal)	Type I error ( $\alpha$ ) False Positive	Correct decision ( $1 - \beta$ ) True Positive

The significance level of test  $\alpha$  is the probability of making type I error. The probability of type II error is denoted as  $\beta$ . The power of the test,  $(1 - \beta)$ , is defined as the probability of correctly rejecting the false null hypothesis. We need to fix the significance level and power of the test at desired levels in advance. Furthermore, we must specify the desired fold change (FC) or difference between population means to be detected.

Two-sample t-test is derived under the assumptions that the populations are normally distributed and have equal variance. The Welch's t-test is an adaptation of Student's t-test and is more robust when the populations have unequal variance, and/or the sample sizes are unequal. Let us consider independent random samples from two normal populations,  $X_{ij} \sim N(\mu_i, \sigma_i^2)$ , ( $i = 1, 2; j = 1, 2, \dots, N_i$ ) where  $\mu_1, \mu_2, \sigma_1^2$  and  $\sigma_2^2$  are unknown parameters. The Welch's test statistic is defined as

$$t_W = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} \quad (6.1)$$

where  $\bar{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i}$  and  $S_i^2 = \frac{\sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2}{(N_i - 1)}$ .

Under null hypothesis  $H_0$ , the approximate distribution of  $t_W$  given by Welch [92] is  $t_W \sim t(\hat{n})$ , i.e.,  $t$  with  $\hat{n}$  degrees of freedom given by

$$\hat{n} = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{S_1^4}{N_1^2(N_1-1)} + \frac{S_2^4}{N_2^2(N_2-1)}} \quad (6.2)$$

The null hypothesis is rejected if  $|t_W| > t_{\hat{n}, \alpha/2}$ , where  $t_{\hat{n}, \alpha/2}$  is the upper  $100(\alpha/2)$ th percentile of the  $t$ -distribution  $t(\hat{n})$ . The same concept was also suggested

by Smith [95] and Satterthwaite [96]. Therefore, the test is also sometimes referred to as the Smith-Welch-Satterthwaite test. The test is an approximate solution of Behrens-Fisher problem. The exact distribution of Welch's t-test is complicated, and it can be expressed in different forms. We use the following notations for the alternate expression of Welch's t-test [93].

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma} \sim N(\delta, 1)$$

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

$$\sigma^2 = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

$$W = \frac{(N_1 - 1)S_1^2}{\sigma_1^2} + \frac{(N_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(N_1 + N_2 - 2)$$

$$B = \frac{\frac{(N_1 - 1)S_1^2}{\sigma_1^2}}{W} \sim \text{Beta}\left(\frac{(N_1 - 1)}{2}, \frac{(N_2 - 1)}{2}\right)$$

The alternate expression of Welch's test statistic is given by

$$t_W = \frac{T}{\sqrt{H}} \quad (6.3)$$

where  $T = \frac{Z}{\sqrt{\frac{W}{N_1 + N_2 - 2}}} \sim t(N_1 + N_2 - 2, \delta)$ , which is the non-central t-distribution with

degrees of freedom  $N_1 + N_2 - 2$  and non-centrality parameter  $\delta$ ;  $H = \frac{\sigma_1^2 B}{N_1 p} +$

$\frac{\sigma_2^2 (1-B)}{N_2 (1-p)}$  with  $p = \frac{N_1 - 1}{N_1 + N_2 - 2}$ . The random variables,  $Z$ ,  $W$  and  $B$  are mutually

independent. Here, the variables  $T$  and  $B$  are also independent. The alternate expression of degrees of freedom can be written as

$$\hat{n} = \frac{1}{\frac{B_1^2}{(N_1 - 1)} + \frac{B_2^2}{(N_2 - 1)}} \quad (6.4)$$

where,  $B_1 = \frac{\frac{\sigma_1^2 B}{N_1 p}}{\frac{\sigma_1^2 B}{N_1 p} + \frac{\sigma_2^2 (1-B)}{N_2 (1-p)}}$  and  $B_2 = 1 - B_1$ . The power function of  $t_W$  is given by

$$\pi(\mu_d, \sigma_1^2, \sigma_2^2, N_1, N_2) = P\{|t_W| > t_{\hat{n}, \alpha/2}\} = P\{|T| > t_{\hat{n}, \alpha/2} \sqrt{H}\} \quad (6.5)$$

The exact power can be calculated by using Simpson's rule.

### A1. Allocation of samples between two groups

Let the sample size ratio  $\left(\frac{N_2}{N_1} = r \geq 1\right)$  between two groups be fixed in advance.

Then the power function becomes a strictly monotone function of  $N_1$  with other parameters held constant. A simple incremental search can be used to find out the minimum sample size  $N_1$  required to achieve the given power at a significance level  $\alpha$ . The large sample normal approximation can be used as the starting values for the iteration. According to Jan and Shieh [93], the starting sample size  $N_{1Z}$  would be the smallest integer satisfying the inequality

$$N_{1Z} = (\sigma_1^2 + \sigma_2^2/r)(z_{\alpha/2} + z_\beta)^2 / \mu_d^2 \quad (6.6)$$

However, for large values of  $\mu_d$ , the program sometimes return error as the starting value of  $N_{1Z}$  is less than 1. Therefore, we are using  $\max(N_{1Z} - 1, 1)$  as the starting value. For example, if we use the original program with  $\mu_d = 4$ , it will not return any result. With the input parameters given below, we found the following sample sizes with the actual power achieved. The original program was unable to calculate the sample sizes with parameters  $\mu_d = 4, \alpha = 0.05, 1 - \beta = 0.90$  for different values of  $r$  and  $\sigma_1 : \sigma_2$ . The sample sizes and the exact power obtained with these parameters are shown in Table 6.2.

**Table 6.2.** Sample sizes ( $N_1, N_2$ ) and power computed when  $r$  is fixed with parameters  $\mu_d = 4, \alpha = 0.05, 1 - \beta = 0.90$  for different combinations of  $r$  and  $\sigma_1: \sigma_2$ .

		$\sigma_1: \sigma_2$														
		1/3:1			1/2:1			1:1			2:1			3:1		
$r$		$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power
1	3	3	0.9245	3	3	0.9330	4	4	0.9926	6	6	0.9590	9	9	0.9235	
2	2	4	0.9876	2	4	0.9313	3	6	0.9479	5	10	0.9077	9	18	0.9307	
3	2	6	0.9907	2	6	0.9307	3	9	0.9443	5	15	0.909	9	27	0.9328	

## A2. Allocation of samples between two groups under a budget constraint

The various costs incurred for conducting a proteomics experiment such as quantitative LC-MS/MS are as follows:

(i) Sample procurement cost: It depends on the number biological and/ or technical replicates for each condition. For a case-control study, with  $N_1$  and  $N_2$  replicates respectively, in control and case groups, the cost will be  $c_1 = a_{11}N_1 + a_{12}N_2$ , where  $a_{11}$  and  $a_{12}$  are respectively the sample procurement cost per sample in control and case.

(ii) Sample preparation cost: It involves the methods such as digestion (e.g. trypsin), alkylation,  $\mu$ -Solid Phase Extraction and sample cleaning. The cost will be  $c_2 = a_2(N_1 + N_2)$ , where  $a_2$  is the sample preparation cost per sample.

(iii) LC-MS/ MS analysis: The cost of LC-MS/MS analysis will depend on the type of sample (e.g., simple or complex) and duration. The cost for LC-MS/MS analysis will be  $c_3 = a_3(N_1 + N_2)$ , where  $a_3$  is per sample LC-MS/MS analysis cost.

(iv) Database search and protein identification:  $c_4 = a_4(N_1 + N_2)$ , where  $a_4$  is the cost for database search and identification.

(v) Analysis:  $c_5 = a_5(N_1 + N_2)$ , where  $a_5$  is the average cost of analysis per sample.

The total cost for conducting the experiment is given by

$$C = c_1 + c_2 + c_3 + c_4 + c_5$$

$$= (a_{11} + a_2 + a_3 + a_4 + a_5)N_1 + (a_{12} + a_2 + a_3 + a_4 + a_5)N_2$$

Therefore, the total cost can be written as  $C = C_1N_1 + C_2N_2$ , where  $C_1$  and  $C_2$  are the average cost per sample in control and case, respectively. The list of prices for each step are available at various online sources. A hypothetical example of cost calculation per sample in a quantitative proteomics experiment is given below in Table 6.3.

**Table 6.3.** A hypothetical example of various costs involved in quantitative proteomics experiment

<b>Services</b>	<b>Price per sample (in USD)</b>
Sample procurement cost	50
Sample preparation: digestion, extraction and cleanup	60
LC-MS/MS	100
Data base search and protein identification	50
Analysis	40
<b>Total cost per sample</b>	<b>300</b>

Let the total cost  $C$  is fixed in advance as given below:

$$C = C_1N_1 + C_2N_2 \tag{6.7}$$

The optimal sample size ratio is proportional to the ratio of standard deviations divided by the square root of ratio of costs [97]. Therefore, the optimal allocation is obtained when the ratio of sample sizes assumes the equality

$$\frac{N_2}{N_1} = \frac{\sigma_2 C_1^{1/2}}{\sigma_1 C_2^{1/2}} = \theta \quad (6.8)$$

#### A2.1. Sample allocation with maximum power under a fixed cost

When the total cost is fixed, then the maximum power is obtained using the sample size combination given below:

$$N_{1Z} = \frac{c(\sigma_1 C_2^{1/2})}{c_1(\sigma_1 C_2^{1/2}) + c_2(\sigma_2 C_1^{1/2})} \quad (6.9)$$

$$N_{2Z} = \frac{c(\sigma_2 C_1^{1/2})}{c_1(\sigma_1 C_2^{1/2}) + c_2(\sigma_2 C_1^{1/2})} \quad (6.10)$$

We calculate the power for various combinations of  $N_1$  and  $N_2$  and find the optimal allocation, that is, the combination giving the maximum power. We vary the value of  $N_1$  from  $N_{1Min}$  to  $N_{1Max}$ , where  $N_{1Min} = \text{floor}(N_{1Z}) - 1$  and  $N_{1Max} = \text{ceiling}\left[\frac{C - C_2\{\text{floor}(N_{2Z}) - 1\}}{c_1}\right]$ . The function  $\text{floor}(x)$  in R rounds to the nearest integer that is smaller than  $x$ . The function  $\text{ceiling}(x)$  in R rounds to the nearest integer that is larger than  $x$ . In their work, the value of  $N_{1Max}$  was rounded using floor function. However, we found the maximum power is achieved on using ceiling function. Also, we found that the number of samples required is less when the ceiling function is used instead of floor function. For example, we estimate sample sizes for a fixed cost  $C$  as given below in Table 6.4. We found that our method has more power for a given fixed cost with less number of samples as compared to the original method.

**Table 6.4.** Sample sizes ( $N_1$ ,  $N_2$ ), total number of samples ( $N_1 + N_2$ ) and the power obtained for fixed cost  $C$  with parameters  $\mu_d = 4, \alpha = 0.05$ ,  $C_1: C_2 = 1: 1/3$  and different values of  $\sigma_1: \sigma_2$  using our method and the original method.

	$C_1: C_2$	$\sigma_1: \sigma_2$									
		1/3:1					1/2:1				
		<b>C</b>	$N_1$	$N_2$	$N_1 + N_2$	<b>Power</b>	<b>C</b>	$N_1$	$N_2$	$N_1 + N_2$	<b>Power</b>
Our method	1:1/3	25	10	45	55	0.999646	30	15	45	60	0.998665
Original method	1:1/3	25	9	48	57	0.999626	30	14	48	62	0.998663

### A2.2. Sample allocation with minimum cost for a fixed power

When the power is fixed, then the minimum total cost can be obtained using the sample size combination given below:

$$N_{1Z} = \frac{(\theta\sigma_1^2 + \sigma_2^2)(z_{\alpha/2} + z_{\beta})^2}{\theta\mu_d^2} \quad (6.11)$$

$$N_{2Z} = \frac{(\theta\sigma_1^2 + \sigma_2^2)(z_{\alpha/2} + z_{\beta})^2}{\mu_d^2} \quad (6.12)$$

The optimal allocation is found by screening the different sample size combinations and finding the combination that gives the minimum cost at the desired power. We vary the value of  $N_1$  from  $N_{1Min}$  to  $N_{1Max}$ , where  $N_{1Min} =$

$$\max\{\text{floor}(N_{1Z}), 2\} \text{ and } N_{1Max} = \max\left\{\text{ceiling}\left[\frac{\frac{\sigma_1^2}{\mu_d^2}}{\frac{(z_{\alpha/2} + z_{\beta})^2}{(\text{floor}(N_{2Z}) - 1)}}\right], 2\right\}. \text{ We use}$$

different form of  $N_{1Min}$  and  $N_{1Max}$  that differs from the original work. The minimum and maximum value of  $N_1$  must be at least 2.



## B. Sample size calculation for comparing means between two groups for multiple features

In proteomics experiments, the experimenter is interested in comparing the group means for many features. Several multivariate generalizations of type I error and power of the test exist along with several statistical techniques of their control. Suppose we simultaneously test  $m$  null hypotheses (or compare the abundance of  $m$  features)  $(H_1, H_2, \dots, H_m)$ . We reject the null hypothesis if the test is declared significant. We do not reject the null hypothesis if the test is non-significant. Let  $m_0$  features do not differ significantly between the two populations (number of true null hypothesis). The various possible outcomes for testing multiple null hypotheses are shown below in Table 6.5.

**Table 6.5.** The possible outcomes of testing multiple null hypotheses

		True state		
		Null/ Non-significant	Alternative/ Significant	Total
Decision about null hypothesis	Failed to reject null/ Declared non-significant	$U$ (TN)	$T$ (FN)	$m - R$
	Rejected null/ Declared significant	$V$ (FP)	$S$ (TP)	$R$
Total		$m_0$	$m_1 = m - m_0$	$m$

We define the following terms based on Table 6.5:

$m$  is the number of features/hypotheses tested.

$m_0$  is the number of true null hypothesis (unknown parameter).

$V$  is the number of false positives (type I error)/ false discoveries.

$U$  is the number of true negatives.

$T$  is the number of false negatives (type II error)

$S$  is the number of true positives/ true discoveries.

$R = V+S$ , the number of rejected null hypotheses/ discoveries.

$S$ ,  $T$ ,  $U$  and  $V$  are unobservable random variables.

$R$  is observable random variable.

False discovery rate (FDR) is considered as one of the most powerful multivariate generalization of type I error. FDR-controlling procedures are designed to control the expected proportion of false discoveries (incorrectly rejected null hypotheses). Mathematically, FDR is defined as

$$FDR = E \left[ \frac{V}{\max(R,1)} \right] \quad (6.13)$$

where  $E[.]$  denotes the expected value. The Benjamini-Hochberg (BH) procedure [98] can be used to calculate the number of replicates required for future experiments with multiple features while controlling the FDR. In BH procedure, we first arrange the p-values of the  $m$  comparisons from largest (least significant) to smallest (most significant) values. Then, we compare  $p_{(j)}$  with  $(j/m)*q$ . We reject the null hypotheses if  $p_{(j)} \leq (j/m)*q$ . Let  $R_{ave}$  be the average number of rejections and  $(1 - \beta_{ave})$  be the average power. Then it follows that

$$m\alpha_{ave} \leq R_{ave}q \cong [m_0\alpha_{ave} + m_1(1 - \beta_{ave})]q \quad (6.14)$$

Thus, the BH procedure controls the average type I error over all the features at

$$\alpha_{ave} \leq \frac{(1-\beta_{ave})q}{1+(1-q)\frac{m_0}{m_1}} \quad (6.15)$$

The procedure provides less stringent control of type I errors compared to familywise error rate controlling procedures such as the Bonferroni correction.

The sample size calculation methods for comparing group means using Welch's t-test were extended for multiple features. The users have to specify extra input parameters, namely, FDR, average power ( $1 - \beta_{ave}$ ), number of features ( $m$ ) and expected number of differentially expressed features ( $m_1$ ). Based on these extra input parameters, we compute the average type I error ( $\alpha_{ave}$ ) for overall features. Then we replace the significance level  $\alpha$  and power ( $1 - \beta$ ) by  $\alpha_{ave}$  and  $(1 - \beta_{ave})$ , respectively in the formulae given in Section A. The three options available for estimating sample size without using pilot data by specifying only the input parameters are given below:

**B1.1.** Sample allocation with no cost constraint: Please see Section A1.1.

**B2.1.** Sample allocation with maximum power for a fixed cost: Please see Section A2.1.

**B2.2.** Sample allocation with minimum cost for a fixed power: Please see Section A2.2.

We have developed a shiny application for computing sample size under various constraints as discussed in *Sections A and B*.

### **Sample size calculation using pilot data**

We studied various ways to calculate sample size for detecting differentially expressed features between two groups based on pilot data for conducting future experiments. We used the data corresponding to first two groups as given in previous chapter and estimated the sample sizes for detecting differentially expressed proteins between two groups. We normalized the data first by taking logarithmic base 2 followed by quantile normalization (QN). We used the singular

value decomposition (SVD) method to impute the data. There are total 200 proteins ( $m = 200$ ). We assumed the expected proportion of differentially expressed proteins to be 0.10. We computed  $\alpha_{ave}$  ( $= 0.0042$ ) by assuming the average power to be 0.80 and FDR ( $q$ ) to be 0.05 . We estimated the standard deviations (SDs),  $\hat{\sigma}_1$  and  $\hat{\sigma}_2$  in two groups for each feature from the data. Then, we used the mean, median, 90<sup>th</sup> percentile value and maximum value of estimated SDs for sample size calculation for each of the three options. Please see Table 6.6 for summary of estimates of SDs for two groups.

**Table 6.6.** Summary of estimated standard deviations for two groups

	$\hat{\sigma}_1$	$\hat{\sigma}_2$
<b>Mean</b>	0.52	0.59
<b>Median</b>	0.41	0.36
<b>90<sup>th</sup> percentile</b>	1.09	1.30
<b>Maximum</b>	2.27	3.64

The sample size calculation for each of the three options under various scenarios are given below.

(i) Sample allocation with no cost constraint: We computed the sample size for experiment with equal samples ( $r = \frac{N_2}{N_1} = 1$ ) and unequal samples ( $r = \frac{N_2}{N_1} = 2$ ).

The results are given in Table 6.7. For equal sample sizes, we found that it requires minimum number of samples with maximum power using median values of SDs. For unequal sample sizes, we found that it requires minimum number of samples using median values of SDs whereas the maximum power is obtained using mean values of SDs.

**Table 6.7.** Sample sizes ( $N_1$ ,  $N_2$ ), total number of samples ( $N_1 + N_2$ ) and the exact power obtained with  $\log_2$  FC 1 for fixed sample size ratio (1 and 2).

		$N_1$	$N_2$	$N_1 + N_2$	Exact power
$r = \frac{N_2}{N_1} = 1$	Mean	11	11	22	0.8160
	Median	7	7	14	0.8463
	90 <sup>th</sup> percentile	42	42	84	0.8061
	Maximum	256	256	512	0.8020
$r = \frac{N_2}{N_1} = 2$	Mean	9	18	27	0.8622
	Median	6	12	18	0.8187
	90 <sup>th</sup> percentile	30	60	90	0.8072
	Maximum	163	326	489	0.8002

(ii) Sample allocation with maximum power for a fixed cost: We assumed the cost (in USD) per sample in group 1 and 2 are 300 and 325, respectively. Then, we estimated the sample sizes giving maximum power for conducting experiments with the total budget of 5000, 10000 and 15000, respectively. The results are shown below in Table 6.8. We found that maximum power with minimum number of samples are obtained using median values of SDs for all the three fixed costs.

**Table 6.8.** Sample sizes ( $N_1$ ,  $N_2$ ), total number of samples ( $N_1 + N_2$ ) and the maximum power obtained with  $\log_2$  FC 1 for fixed cost (5000, 10000 and 15000)

		$N_1$	$N_2$	$N_1 + N_2$	Exact power
$C = 5000$	Mean	8	8	16	0.562
	Median	8	8	16	0.9258
	90 <sup>th</sup> percentile	8	8	16	0.0697
	Maximum	8	8	16	0.0106

<b>C = 10000</b>	<b>Mean</b>	16	16	32	0.9698
	<b>Median</b>	17	15	32	1
	<b>90<sup>th</sup> percentile</b>	16	16	32	0.2492
	<b>Maximum</b>	12	19	31	0.0229
<b>C = 15000</b>	<b>Mean</b>	24	24	48	0.9991
	<b>Median</b>	25	23	48	1
	<b>90<sup>th</sup> percentile</b>	24	24	48	0.4571
	<b>Maximum</b>	18	29	47	0.0391

(iii) Sample allocation with minimum cost for a fixed power: We assumed the cost (in USD) per sample in group 1 and 2 are 300 and 325, respectively. We calculated the minimum cost for conducting the experiment to achieve a desirable power of 0.80. The results are shown below in Table 6.9

**Table 6.9.** Sample sizes ( $N_1$ ,  $N_2$ ), total number of samples ( $N_1 + N_2$ ) and exact power obtained for a minimum cost of experiment with  $\log_2$  FC 1

	$N_1$	$N_2$	$N_1 + N_2$	Power	Minimum cost
<b>Mean</b>	11	11	22	0.8160	6875
<b>Median</b>	8	6	14	0.8376	4350
<b>90<sup>th</sup> percentile</b>	39	44	83	0.8040	26000
<b>Maximum</b>	191	293	484	0.8006	152525

### **Web app for calculating sample size using pilot data**

We have developed a tool/app for calculating sample size for detecting differentially expressed features between two groups based on pilot data for conducting future experiments. We have provided various options of estimating

sample size using pilot data. A screen shot of the app “SSCP: Sample Size Calculator for Proteomics Experiment” is given below:

#### SSCP: Sample Size Calculator for Proteomics Experiment

Choose file to upload expression data  
Browse... No file selected

Select feature type  
▼

Choose file to upload additional information  
Browse... No file selected

Select name of class under comparison  
▼

Expected proportion of significant features  
0.1

Specify false discovery rate  
0.05

Specify average power  
0.8

Specify  $\log_2$  fold change  
1

Choose sample size calculation method  
 Sample allocation with no cost constraint  
 Sample allocation with maximum power for a fixed cost  
 Sample allocation with minimum cost for a fixed power

Submit

**Figure 6.1.** Screenshot of the tool “SSCP”

Various inputs to be provided by the user are as follows:

- (1) Choose file to upload pilot data: A pilot dataset in matrix form with  $N$  ( $N=N_1+N_2$ ) subjects in columns and  $m$  features in rows. In proteomics experiment, the proteomics expression data may have MVs. We normalize the data first by taking logarithmic base 2 followed by quantile normalization. We use SVD method to impute the data in case of data with missing values.
- (2) Select feature type (“Protein” or “Peptide”): The calculation will be based on detection of differentially expressed features either at protein or peptide level. We

summarize/aggregate the expression data by taking mean of common or redundant features.

(3) Choose file to upload additional information: The file contains the sample names and the group/class names under comparison.

(4) The expected proportion of significant features ( $\pi_1$ ): The user has to specify the expected proportion of differentially expressed features. The default value is 0.10.

(5) The desired FDR ( $q$ ): We use the procedure given in Equations 6.13-6.15 for controlling the FDR. The default value of FDR is 0.05.

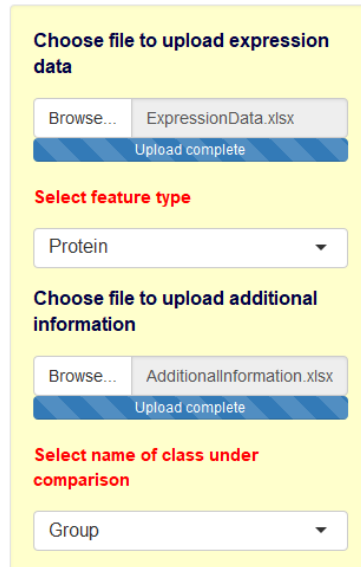
(6) The desired average power ( $1 - \beta_{ave}$ ): The average power is specified to calculate the  $\alpha_{ave}$  using Equation 6.15.

(7) Desirable  $\log_2$  FC: We assumed that the normalized data (log2 transformation followed by QN) is normally distributed.

(8) Choose the method of sample size calculation: We have provided three options for calculating sample size: (i) Sample allocation with no cost constraint, (ii) Sample allocation with maximum power for a fixed cost and (iii) Sample allocation with minimum cost for a fixed power. The user has to define other input parameters for selected method of sample size calculation. The user has to specify the sample size ratio after selecting the first method. If the user selects second method, then he has to specify cost per sample in group 1 and 2 as well as the total cost of the experiment. On selecting the third method, the user has to specify the cost per sample in group 1 and 2.



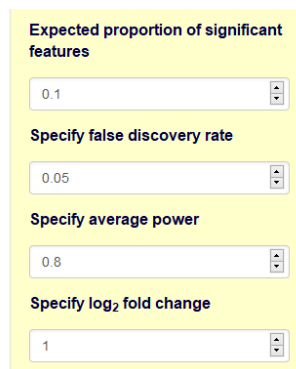
Based on the various input parameters, the user will get two tables (inputs selected and outputs obtained). Example of inputs provided by the user from steps 1 to 3 is given below.



The screenshot shows a yellow background with four sections. The first section is titled "Choose file to upload expression data" and contains a file upload button labeled "Browse..." with "ExpressionData.xlsx" next to it, and a blue "Upload complete" button below. The second section is titled "Select feature type" and contains a dropdown menu with "Protein" selected. The third section is titled "Choose file to upload additional information" and contains a file upload button labeled "Browse..." with "AdditionalInformation.xlsx" next to it, and a blue "Upload complete" button below. The fourth section is titled "Select name of class under comparison" and contains a dropdown menu with "Group" selected.

**Figure 6.2.** Uploading the two input files and selecting feature type and class name under comparison

Now the user has to specify other input parameters (steps 4-7) as given below.



The screenshot shows a yellow background with four sections, each with a numerical input field and a dropdown arrow. The first section is titled "Expected proportion of significant features" with the value "0.1". The second section is titled "Specify false discovery rate" with the value "0.05". The third section is titled "Specify average power" with the value "0.8". The fourth section is titled "Specify log<sub>2</sub> fold change" with the value "1".

**Figure 6.3.** Specifying expected proportion of significant features, false discovery rate, average power and log<sub>2</sub> fold change

Now the user has to choose any one method of sample size calculation (Step 8).

The screenshot of example under each constraint are given in Figure 6.4.

**Figure 6.4.** The input parameters under each method of sample size calculation

An example of inputs and outputs after submitting the job for sample allocation with maximum power for a fixed cost of 5000 assuming the cost per sample in group 1 and 2 to be 300 and 325, respectively is given below:

**SSCP: Sample Size Calculator for Proteomics Experiment**

Inputs	
Feature type	Protein
Expected proportion of significant features	0.1
False discovery rate	0.05
Average power	0.8
Desired log <sub>2</sub> fold change	1
Cost per sample in Control	300
Cost per sample in Case1	325
Total cost of the experiment	5000
Sample size calculation method	Sample allocation with maximum power for a fixed cost
<a href="#">Download_Inputs_Selected</a>	
Output	
Number of proteins	200
Expected number of significant proteins	20
Estimated SD of Control	0.41
Estimated SD of Case1	0.36
Number of samples in Control (N <sub>1</sub> )	8
Number of samples in Case1 (N <sub>2</sub> )	8
Average alpha	0.0042
Maximum power	0.9258
<a href="#">Download_Result</a>	

**Figure 6.5.** Example of inputs and output obtained for sample allocation with maximum power for a fixed cost

The results obtained are same as given in Table 6.6 (estimates of standard deviations) and Table 6.8 (sample sizes for fixed cost of 5000). The user can use the app for sample size calculation under various constraints using pilot data. However, in some situations, it may take more time to compute the sample sizes. In such situations, the user can run the R programs on high performance computing facility to get the result.

### **Impact of technical components on the study design and sample size estimation of LC-MS proteomics workflow**

The purpose of the approaches given in Chapters 3 and 4 was to standardize proteomics workflow and to study variability in proteomics expression data. This will help in designing and conducting proteomics experiments. As discussed in Chapter 3, we have two tissue storage methods (FFPE and FR) and three tissue extraction methods (MAX, TX.MAX and SDS.MAX). We studied the technical variability associated with proteomics expression data. In this section, we studied the impact of technical variability on the study design using real dataset given in Chapter 3. We estimated the sample size based on CV and % effect sizes [94].

#### **Sample size formulation**

Let us consider two normal populations with means  $\mu_1, \mu_2$  and variances  $\sigma_1^2, \sigma_2^2$ . When the variances are known, then the sample size per group with significance level  $\alpha$  and power  $(1 - \beta)$  is given by

$$N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 v^2 (\mu_r^2 + 1)}{(\mu_r - 1)^2}$$

where,  $\mu_r = \frac{\mu_1}{\mu_2}, \frac{\sigma_1}{\sigma_2} = \frac{\mu_2}{\sigma_2} = v$ ,  $z_{1-p}$  is the 100(1-p)th percentile of standard normal distribution. The above formula can be adjusted by FDR-controlling procedures for testing the significance of multiple features.

### **Sample size estimation**

We studied the effect of tissue storage methods (FFPE and FR) and tissue extraction methods (MAX, TX.MAX and SDS.MAX) on sample size estimation. For example, suppose the experimenter has used the FR method for tissue storage and he wants to estimate sample size for two-class comparison. Then, the formula given in previous section can be used to estimate sample size based on CV and the percentage change in means.

We used variance stabilizing normalization (VSN) method for data normalization and SVD method for data imputation. We estimated the sample size using the median and maximum value of CV for two TSMs and three TEMs as given in Table 3.3. of Chapter 3. We have provided the sample sizes computed for all the technical approaches at different percent change between means ( $\mu_r$  - fold difference in means) in Table 6.10. The sample sizes corresponding to without adjustment and FDR-adjusted are separated by “/”.

**Table 6.10.** Computed sample sizes for different technical approaches

	Using median value of CV					Using maximum value of CV				
	TSM		TEM			TSM		TEM		
	FR	FFPE	MAX	TX.MAX	SDS.MAX	FR	FFPE	MAX	TX.MAX	SDS.MAX
<b>1.05</b>	66/ 78	6/ 8	13/ 15	13/ 15	72/ 85	445/ 524	246/ 290	339/ 400	306/ 361	461/ 543
<b>1.1</b>	18/ 21	2/ 2	4/ 4	4/ 4	19/ 23	117/ 138	65/ 77	90/ 105	81/ 95	122/ 143
<b>1.15</b>	9/ 10	1/ 1	2/ 2	2/ 2	9/ 11	55/ 65	31/ 36	42/ 49	38/ 45	57/ 67
<b>1.2</b>	5/ 6	1/ 1	1/ 2	1/ 2	6/ 7	33/ 38	18/ 21	25/ 29	23/ 27	34/ 40
<b>1.25</b>	4/ 4	1/ 1	1/ 1	1/ 1	4/ 5	22/ 26	12/ 15	17/ 20	15/ 18	23/ 27
<b>1.3</b>	3/ 3	1/ 1	1/ 1	1/ 1	3/ 3	16/ 19	9/ 11	13/ 15	11/ 13	17/ 20
<b>1.35</b>	2/ 3	1/ 1	1/ 1	1/ 1	2/ 3	13/ 15	7/ 8	10/ 11	9/ 10	13/ 15
<b>1.4</b>	2/ 2	1/ 1	1/ 1	1/ 1	2/ 2	10/ 12	6/ 7	8/ 9	7/ 8	11/ 12
<b>1.45</b>	2/ 2	1/ 1	1/ 1	1/ 1	2/ 2	9/ 10	5/ 6	7/ 8	6/ 7	9/ 10
<b>1.5</b>	2/ 2	1/ 1	1/ 1	1/ 1	2/ 2	7/ 9	4/ 5	6/ 7	5/ 6	8/ 9
<b>1.75</b>	1/ 1	1/ 1	1/ 1	1/ 1	1/ 1	4/ 5	3/ 3	3/ 4	3/ 4	4/ 5
<b>2</b>	1/ 1	1/ 1	1/ 1	1/ 1	1/ 1	3/ 4	2/ 2	3/ 3	2/ 3	3/ 4
<b>2.5</b>	1/ 1	1/ 1	1/ 1	1/ 1	1/ 1	2/ 3	1/ 2	2/ 2	2/ 2	2/ 3
<b>3</b>	1/ 1	1/ 1	1/ 1	1/ 1	1/ 1	2/ 2	1/ 1	2/ 2	1/ 2	2/ 2

The FR TSM requires more sample as compared to FFPE TSM. In TEM, the SDS.MAX method requires more number of samples as compared to those of MAX and TX.MAX. The sample size estimated is more when adjusting for multiple proteins. The sample size with value 1 should be ignored. Here, we have considered only the technical variability for sample size estimation. We have not considered the biological variability. The sample size estimated will be much more on inclusion of biological variability.

## CHAPTER 7

### SAMPLE SIZE CALCULATION FOR RNA-SEQ EXPERIMENTS CONSIDERING HETEROGENEITY

#### **Introduction**

RNA-Seq has become the standard for measuring gene expression levels in biological experiments. It differs from the microarray technology in various aspects such as nature of data, normalization methods, differential expression analysis methods, sensitivity, accuracy, etc. [99-101]. The RNA-Seq method is developing rapidly and the cost of sequencing is declining. So, in the coming future, more samples will be sequenced, and more experiments will be performed. But still the cost per sample is the limiting factor in most of the laboratories. There are two important points to be considered while designing RNA-Seq experiments which are namely, the sequencing depth and the number of replicates (biological and technical) required to observe significant changes in expression. The other points should also be considered such as length of transcripts, GC content and sequencing bias (influencing counts of transcripts within a sample). The cost can be reduced by optimizing the designing process of these experiments.

Various tools and software have been developed to address the problem of sample size estimation and power analysis. Some of the examples are RNASeqPowerCalculator [102], RNASeqPower [103], Scotty [104], PROPER

[105], etc. The RNA-seq experiments are complex in nature, and still there is requirement of advanced method to calculate sample size for differential expression analysis using RNA-Seq data. In spite of various developments, this field still lacks a general approach to estimate optimal sample size and power for complex RNA-Seq experiments under the assumptions of various distributions. There are various issues with the read counts for sample size and power calculation such as over dispersion parameter estimation, excess zeros, complexity of model, etc. The results obtained using the various methods for differential expression analysis of RNA-Seq data from single organism or from various sources of RNA-Seq data, do not lead to a common conclusion and sometimes the results are not meaningful [106]. The misleading results are caused due to heterogeneity issues at each step of RNA-Seq experiments. Therefore, it is imperative to devise a statistical procedure for optimizing the sample size calculation with reasonable statistical power and cost required for conducting the experiments in the presence of heterogeneity. Therefore, we have developed the statistical approaches for designing and sample size calculation required to carry out the RNA-Seq experiments in the presence of heterogeneity under the assumptions of various models.

### **Modeling the count data in RNA-Seq experiments**

The RNA-Seq data is comprised mainly of the mapped read counts. The counting of feature can be done at various levels such as gene level, transcript level, exon level, etc. Instead of raw counts, normalized read counts such as RPKM (reads aligned per kilobase of exon per million reads mapped) [107],

FPKM (fragments per kilobase of exon per million fragments mapped), TPM (transcripts per kilobase million), etc. are also used. However, from the statistical point of view, actual counts are used as input for differential expression analysis in many cases [108, 109]. As the raw count data are discrete in nature, therefore, cannot be necessarily approximated well by normal (Gaussian) distributions, therefore, the use of standard linear models like t-tests, ANOVA, regression should not be preferred as the modeling framework. There are two popular distributions for modelling the read counts which are given below:

(i) Poisson distribution [99, 110-113]: Let  $Y \sim \text{Poisson}(\lambda)$  be a random variable representing the read counts for a gene in a sample, then its probability mass function is given by

$$p(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, 2, \dots \quad (7.1)$$

where  $\lambda$  is the rate parameter, i.e., expected number of reads per sample. The mean and variance are equal to  $\lambda$ . When  $\lambda$  is small, the data is over-dispersed, i.e., there is more variation than expected under  $\text{Poisson}(\lambda)$ . Similarly, when  $\lambda$  is large, there is less variation than expected under  $\text{Poisson}(\lambda)$ . Therefore, in most of the cases, the RNA-Seq data is not modeled well by Poisson distribution as the relationships between means and variances tend to be far more complicated among (and within) biological replicates. The Poisson distribution accounts only for the technical replicates. It is not well suited to account for the biological replicates due to the problem of over dispersion caused by biological variations. Various other forms of Poisson distribution such as Quasi-Poisson have been developed to account for it with count data.



(ii) Negative Binomial (NB) distribution [108, 109, 114-116]: The NB distribution is assumed to be the natural distribution for modelling the read counts. If a random variable  $Y$  has NB distribution with mean parameter  $\mu$  and dispersion parameter  $\phi$ , then its probability mass function is given by

$$p(Y = y) = \frac{\Gamma(y + \frac{1}{\phi})}{\Gamma(\frac{1}{\phi})\Gamma(y+1)} \frac{(\mu\phi)^y}{(1+\mu\phi)^{y+\frac{1}{\phi}}}; y = 0, 1, 2, \dots \quad (7.2)$$

with expected number of counts =  $E(Y) = \mu$  and  $Var(Y) = \mu + \phi\mu^2$ . Here, the dispersion parameter,  $\phi$  is a measure of extra variance of  $Y$  that the Poisson distribution does not account.

The regulation of gene expression can be studied across different conditions such as levels of different factors, genotypes, environmental conditions, developmental stages, etc. The major goal of RNA-Seq experiments is to determine which features (e.g., genes, transcripts) show significant changes in abundance across different condition or treatment. For differential expression analysis,  $Y$  can be considered as the number of reads mapped to a reference genome (read counts) that is generally modelled by assuming Poisson distribution or NB distribution. Let us consider an RNA-Seq experiment including a total of  $N$  samples. The samples are sequenced and resulting reads are aligned with a reference genome. The numbers of reads mapped to each of the reference gene are calculated.

Let us consider a study in which there are  $I$  conditions/groups denoted by  $G_i (i = 1, 2, \dots, I)$  and there are  $N_i$  samples denoted by  $S_{i,j} (j = 1, 2, \dots, N_i)$  corresponding to group  $G_i$ . Now suppose, there are  $K$  genes/features denoted by

$F_k (k = 1, 2, \dots, K)$  (Please see Table 2.1 in Chapter 2). Let  $y_{i,j,k}$  be the number of read counts corresponding to sample  $S_{i,j}$  ( $i = 1, 2, \dots, I; j = 1, 2, \dots, N_i$ ) of group  $G_i$  ( $i = 1, 2, \dots, I$ ) for gene/feature  $F_k (k = 1, 2, \dots, K)$ . Total number of samples in the study is  $N = \sum_{i=1}^I N_i$ . The library size/sequencing depth for the  $j^{\text{th}}$  sample of group  $G_i$  is denoted by  $L_{ij} = \sum_{k=1}^K y_{ijk}$  which varies for each sample. Normalization techniques are used to account for the within library and between library variability. The normalized count data can be modelled by Poisson distribution or NB distribution.

### **Estimation of parameters based on negative binomial distribution**

Let for any feature  $F_k$ , the observations  $Y_{ijk}$  are independently and identically distributed as

$$Y_{ijk} \sim NB(s_{ijk}\mu_{ik}, \phi_{ik})$$

where  $\mu_{ik}$  and  $\phi_{ik}$  are the true expression level and dispersion parameter, respectively for the feature  $F_k$  in group  $G_i$ , respectively;  $s_{ij}$  is the scaling/size factor to normalize the raw read counts corresponding to the  $j^{\text{th}}$  sample in  $i^{\text{th}}$  group. There are  $N_i$  observations in group  $G_i$  for each feature. The total number of reads in the  $i^{\text{th}}$  group for feature  $F_k$  is  $Y_{ik} = \sum_{j=1}^{N_i} Y_{ijk}$ . For simplicity, we suppressed the notation for feature  $F_k$  in the subscript of previous terms.

The estimation of parameters is an essential step for design and sample size calculation. The parameter estimation can be done by using various methods such as method of moments estimation (MME) [117], maximum likelihood estimation (MLE) [118-120], maximum quasi-likelihood estimation

(MQLE) [121]. The MME has certain limitations (when variance equals mean, dispersion parameter equals infinity; when variance is less than mean, dispersion parameter is negative; when variance-mean is small, dispersion parameter is very large). The MLE methods tend to underestimate the dispersion parameters. Besides these methods, there are various methods/models for estimation of parameters such as pseudo-likelihood [122, 123], quasi-likelihood [124], conditional maximum likelihood (CML) [125], conditional inference [126], quantile-adjusted CML [114], conditional weighted likelihood [109].

**Estimation of parameters without scaling factor:** Let  $Y_{ij}$  be a NB random variable with mean  $\mu_i$  and dispersion parameter  $\phi_i$ , i.e.,  $Y_{ij} \sim NB(\mu_i, \phi_i)$ , then its probability mass function is given by

$$p(Y_{ij} = y_{ij}) = \frac{\Gamma(y_{ij} + \frac{1}{\phi_i})}{\Gamma(\frac{1}{\phi_i})\Gamma(y_{ij} + 1)} \frac{(\mu_i \phi_i)^{y_{ij}}}{(1 + \mu_i \phi_i)^{y_{ij} + \frac{1}{\phi_i}}}; y_{ij} = 0, 1, 2, \dots \quad (7.3)$$

Then, the likelihood function is given by

$$L(\mu_i, \phi_i | y_{ij}; j = 1, 2, \dots, N_i) = \prod_{j=1}^{N_i} \frac{\Gamma(y_{ij} + \frac{1}{\phi_i})}{\Gamma(\frac{1}{\phi_i})\Gamma(y_{ij} + 1)} \frac{(\mu_i \phi_i)^{y_{ij}}}{(1 + \mu_i \phi_i)^{y_{ij} + \frac{1}{\phi_i}}} \quad (7.4)$$

The log-likelihood function is given by

$$\begin{aligned} l(\mu_i, \phi_i | y_{ij}; j = 1, 2, \dots, N_i) &= \sum_{j=1}^{N_i} \ln \Gamma\left(y_{ij} + \frac{1}{\phi_i}\right) - \sum_{j=1}^{N_i} \ln \Gamma\left(\frac{1}{\phi_i}\right) - \sum_{j=1}^{N_i} \ln \Gamma(y_{ij} + 1) \\ &+ \sum_{j=1}^{N_i} y_{ij} \ln(\mu_i \phi_i) - \sum_{j=1}^{N_i} \left(y_{ij} + \frac{1}{\phi_i}\right) \ln(1 + \mu_i \phi_i) \end{aligned} \quad (7.5)$$

Differentiating with respect to  $\mu_i$  and equating to zero, we get

$$\frac{\partial l}{\partial \mu_i} = \frac{\sum_{j=1}^{N_i} y_{ij}}{\mu_i} - \frac{\sum_{j=1}^{N_i} \left(y_{ij} + \frac{1}{\phi_i}\right) \phi_i}{(1 + \mu_i \phi_i)} = 0$$

$$\Rightarrow \hat{\mu}_i = \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} \quad (7.6)$$

Differentiating with respect to  $\phi_i$  and equating to zero, we get

$$\frac{\partial l}{\partial \phi_i} = \frac{\partial \sum_{j=1}^{N_i} \ln \Gamma\left(y_{ij} + \frac{1}{\phi_i}\right)}{\partial \phi_i} - \frac{\partial \sum_{j=1}^{N_i} \ln \Gamma\left(\frac{1}{\phi_i}\right)}{\partial \phi_i} + \frac{\sum_{j=1}^{N_i} y_{ij}}{\phi_i} - \frac{\sum_{j=1}^{N_i} \left(y_{ij} + \frac{1}{\phi_i}\right) \mu_i}{(1 + \mu_i \phi_i)}$$

$$+ \frac{N_i}{\phi_i^2} \ln(1 + \mu_i \phi_i)$$

Putting  $\hat{\mu}_i = \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i}$  in the above equation, we get

$$\frac{\partial l}{\partial \phi_i} = \frac{\partial \sum_{j=1}^{N_i} \ln \left( \frac{\Gamma\left(y_{ij} + \frac{1}{\phi_i}\right)}{\Gamma\left(\frac{1}{\phi_i}\right)} \right)}{\partial \phi_i} + \frac{N_i}{\phi_i^2} \ln \left( 1 + \phi_i \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} \right) = 0$$

Further simplification leads to

$$\frac{\partial l}{\partial \phi_i} = \frac{1}{\phi_i^2} \left\{ N_i \ln \left( 1 + \phi_i \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} \right) - \sum_{j=1}^{N_i} \sum_{m=0}^{y_{ij}-1} \frac{1}{\left(m + \frac{1}{\phi_i}\right)} \right\} = 0 \quad (7.7)$$

Since, the above equation is not in closed form, therefore, we have used Newton's method to estimate the dispersion parameter  $\phi$ . The second derivative of log-likelihood function with respect to  $\phi_i$  is given by

$$\frac{\partial^2 l}{\partial \phi_i^2} = -\frac{2N_i}{\phi_i^3} \ln \left( 1 + \phi_i \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} \right) + \frac{\sum_{j=1}^{N_i} y_{ij}}{\phi_i^2 \left( 1 + \phi_i \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i} \right)} \frac{2}{\phi_i^3} \sum_{j=1}^{N_i} \sum_{m=0}^{y_{ij}-1} \frac{1}{\left( m + \frac{1}{\phi_i} \right)} - \frac{1}{\phi_i^4} \sum_{j=1}^{N_i} \sum_{m=0}^{y_{ij}-1} \frac{1}{\left( m + \frac{1}{\phi_i} \right)^2} \quad (7.8)$$

**Estimation of parameters with scaling factor:** Let  $Y_{ij}$  be a NB random variable with mean  $\mu_i$  and dispersion parameter  $\phi_i$ , i.e.,  $Y_{ij} \sim NB(s_{ij}\mu_i, \phi_i)$ , then its probability mass function is given by

$$p(Y_{ij} = y_{ij}) = \frac{\Gamma(y_{ij} + \frac{1}{\phi_i})}{\Gamma(\frac{1}{\phi_i})\Gamma(y_{ij} + 1)} \frac{(s_{ij}\mu_i\phi_i)^{y_{ij}}}{(1 + s_{ij}\mu_i\phi_i)^{y_{ij} + \frac{1}{\phi_i}}}; y_{ij} = 0, 1, 2, \dots \quad (7.9)$$

where,  $s_{ij}$  is the scaling factor to normalize raw read counts in the  $j^{\text{th}}$  sample of group  $G_i$ . Then, the likelihood function  $L$  and the log-likelihood function  $l$  are given below:

$$L(\mu_i, \phi_i | y_{ij}; j = 1, 2, \dots, N_i) = \prod_{j=1}^{N_i} \frac{\Gamma(y_{ij} + \frac{1}{\phi_i})}{\Gamma(\frac{1}{\phi_i})\Gamma(y_{ij} + 1)} \frac{(s_{ij}\mu_i\phi_i)^{y_{ij}}}{(1 + s_{ij}\mu_i\phi_i)^{y_{ij} + \frac{1}{\phi_i}}} \quad (7.10)$$

$$l(\mu_i, \phi_i | y_{ij}; j = 1, 2, \dots, N_i) = \sum_{j=1}^{N_i} \ln \Gamma \left( y_{ij} + \frac{1}{\phi_i} \right) - \sum_{j=1}^{N_i} \Gamma \left( \frac{1}{\phi_i} \right) - \sum_{j=1}^{N_i} \ln \Gamma(y_{ij} + 1) + \sum_{j=1}^{N_i} y_{ij} \ln(s_{ij}\mu_i\phi_i) - \sum_{j=1}^{N_i} \left( y_{ij} + \frac{1}{\phi_i} \right) \ln(1 + s_{ij}\mu_i\phi_i) \quad (7.11)$$

Differentiating with respect to  $\mu_i$  and equating to zero, we get

$$\frac{\partial l}{\partial \mu_i} = \frac{\sum_{j=1}^{N_i} y_{ij}}{\mu_i} - \sum_{j=1}^{N_i} \frac{\left( y_{ij} + \frac{1}{\phi_i} \right) s_{ij}\phi_i}{(1 + s_{ij}\mu_i\phi_i)} = 0$$

$$\Rightarrow \hat{\mu}_i = \frac{\sum_{j=1}^{N_i} y_{ij}}{\sum_{j=1}^{N_i} s_{ij}} \quad (7.12)$$

Differentiating with respect to  $\phi_i$  and equating to zero, we get

$$\begin{aligned} \frac{\partial l}{\partial \phi_i} &= \frac{\partial \sum_{j=1}^{N_i} \ln \Gamma \left( y_{ij} + \frac{1}{\phi_i} \right)}{\partial \phi_i} - \frac{\partial \sum_{j=1}^{N_i} \ln \Gamma \left( \frac{1}{\phi_i} \right)}{\partial \phi_i} + \frac{\sum_{j=1}^{N_i} y_{ij}}{\phi_i} - \sum_{j=1}^{N_i} \left( \frac{y_{ij} + \frac{1}{\phi_i}}{1 + s_{ij} \mu_i \phi_i} \right) s_{ij} \mu_i \\ &\quad + \sum_{j=1}^{N_i} \frac{\ln(1 + s_{ij} \mu_i \phi_i)}{\phi_i^2} \\ &\Rightarrow \frac{\partial l}{\partial \phi_i} = \frac{\partial \sum_{j=1}^{N_i} \ln \left( \frac{\Gamma \left( y_{ij} + \frac{1}{\phi_i} \right)}{\Gamma \left( \frac{1}{\phi_i} \right)} \right)}{\partial \phi_i} + \sum_{j=1}^{N_i} \frac{\ln(1 + s_{ij} \mu_i \phi_i)}{\phi_i^2} = 0 \end{aligned}$$

Putting  $\hat{\mu}_i = \frac{\sum_{j=1}^{N_i} y_{ij}}{\sum_{j=1}^{N_i} s_{ij}}$  in the above equation, we obtain

$$\frac{\partial l}{\partial \phi_i} = \frac{1}{\phi_i^2} \left\{ \sum_{j=1}^{N_i} \ln \left( 1 + \frac{s_{ij} \sum_{j=1}^{N_i} y_{ij}}{\sum_{j=1}^{N_i} s_{ij}} \phi_i \right) - \sum_{j=1}^{N_i} \sum_{m=0}^{y_j-1} \frac{1}{\left( m + \frac{1}{\phi_i} \right)} \right\} = 0 \quad (7.13)$$

The second derivative with respect to  $\phi_i$  obtained is given by

$$\begin{aligned} \frac{\partial^2 l}{\partial \phi_i^2} &= -\frac{2}{\phi_i^3} \sum_{j=1}^{N_i} \ln \left( 1 + \frac{s_{ij} \sum_{j=1}^{N_i} y_{ij}}{\sum_{j=1}^{N_i} s_{ij}} \phi_i \right) + \frac{1}{\phi_i^2} \sum_{j=1}^{N_i} \left( \frac{\frac{s_{ij} \sum_{j=1}^{N_i} y_{ij}}{\sum_{j=1}^{N_i} s_{ij}}}{1 + \frac{s_{ij} \sum_{j=1}^{N_i} y_{ij}}{\sum_{j=1}^{N_i} s_{ij}} \phi_i} \right) + \\ &\quad \frac{2}{\phi_i^3} \sum_{j=1}^{N_i} \sum_{m=0}^{y_j-1} \frac{1}{\left( m + \frac{1}{\phi_i} \right)} - \frac{1}{\phi_i^4} \sum_{j=1}^{N_i} \sum_{m=0}^{y_j-1} \frac{1}{\left( m + \frac{1}{\phi_i} \right)^2} \end{aligned} \quad (7.14)$$

We used Newton's method to estimate the dispersion parameter  $\phi_i$ . If the scaling factor is 1 for all the samples in the group, i.e.,  $s_{ij} = 1$ , then the estimates of parameters,  $\hat{\mu}_i$  and  $\hat{\phi}_i$ , obtained are same as those of the previous method of estimating parameters without scaling factor. We developed R programs to estimate the dispersion parameter for both cases, i.e., without scaling factor and with scaling factor. For example, the estimated dispersion parameter for  $y = 14, 5, 12, 2, 9, 19$  was found to be 0.2747. For same  $y$  and scaling factor,  $s = 1.1, 1.3, 0.9, 1.4, 1.2, 0.85$ , the value of estimated dispersion parameter is 0.2390.

### **Power and sample size calculation based on negative binomial distribution**

The differential expression analysis in RNA-Seq data involves the calculation of fold change ( $FC = \delta = \frac{\mu_2}{\mu_1}$ ) for each feature such as gene. Therefore, for testing whether a feature is differentially expressed between two groups, we construct the hypothesis setting as given below:

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_1: \mu_1 \neq \mu_2$$

This can be alternatively written as

$$H_0: \Delta = \Delta_0 \text{ vs. } H_1: \Delta \neq \Delta_0$$

where  $\Delta = \mu_2 - \mu_1$ . If we have the null hypothesis that the population means are equal, then  $\Delta_0 = 0$ .

The above settings of the hypothesis can be alternatively written as

$$H_0: \ln(\delta) = \ln(\delta_0) \text{ vs. } H_1: \ln(\delta) \neq \ln(\delta_0)$$

where  $\delta = \frac{\mu_2}{\mu_1}$ . Here,  $\delta_0 = 1$  means the population means are equal in two groups.

The methods based on Wald test and log transformed Wald test have been done previously based on Poisson distribution and NB distribution [127, 128]. We used Wald test, log transformed Wald test and score test based on generalized linear model (GLiM) to estimate power and sample size calculation based on different constraints. These methods are discussed below.

1. Power and sample size calculation using Wald test based on negative binomial distribution

A. Method using Wald test statistic:

Given  $s_i = \sum_{j=1}^{N_i} s_{ij}$  and  $Y_i = \sum_{j=1}^{N_i} y_{ij}$  for  $i = 1, 2$ , the estimate of Wald test statistic (Wald 1943) [129] has been derived below. The statistical inference is based on the quantity

$$T = \hat{\mu}_2 - \hat{\mu}_1 = \frac{Y_2}{s_2} - \frac{Y_1}{s_1}$$

The variance of  $T$  is given by

$$\sigma_T^2 = var(T) = \frac{\mu_2}{s_2} + \frac{\mu_1}{s_1} + \frac{\mu_2^2 \phi_2}{N_2} + \frac{\mu_1^2 \phi_1}{N_1}$$

The estimate of  $\sigma_T^2$  (standard error under  $H_1$ ) is given by

$$S_T^2 = \frac{\hat{\mu}_2}{s_2} + \frac{\hat{\mu}_1}{s_1} + \frac{\hat{\mu}_2^2 \hat{\phi}_2}{N_2} + \frac{\hat{\mu}_1^2 \hat{\phi}_1}{N_1} = \frac{Y_2}{s_2^2} + \frac{Y_1}{s_1^2} + \frac{Y_2^2 \hat{\phi}_2}{s_2^2 N_2} + \frac{Y_1^2 \hat{\phi}_1}{s_1^2 N_1}$$

Let  $w = \frac{s_2}{s_1}$ . The Wald test statistic with unequal sample sizes and dispersion parameters is given by

$$Z_{w1} = \frac{T}{S_T} = \frac{\frac{Y_2}{s_2} - \frac{Y_1}{s_1}}{\sqrt{\frac{Y_2}{s_2^2} + \frac{Y_1}{s_1^2} + \frac{Y_2^2 \hat{\phi}_2}{s_2^2 N_2} + \frac{Y_1^2 \hat{\phi}_1}{s_1^2 N_1}}} = \frac{Y_2 - wY_1}{\sqrt{Y_2 + w^2 Y_1 + \frac{Y_2^2 \hat{\phi}_2}{N_2} + \frac{w^2 Y_1^2 \hat{\phi}_1}{N_1}}} \quad (7.15)$$



We reject null hypothesis when  $|z_{w1}| > z_{1-\alpha/2}$ . The power of the two-sided test is given by

$$\Pr[|z_{w1}| > z_{\alpha/2} | H_1 \text{ is true}] = 1 - \beta$$

$$\Pr[z_{w1} > z_{\alpha/2} | H_1 \text{ is true}] + \Pr[z_{w1} < -z_{\alpha/2} | H_1 \text{ is true}] = 1 - \beta$$

$$1 - \Phi\left[z_{\alpha/2} - \frac{\Delta_0 - \Delta}{S_T}\right] + \Phi\left[-z_{\alpha/2} - \frac{\Delta_0 - \Delta}{S_T}\right] = 1 - \beta$$

$$1 - \Phi\left[z_{\alpha/2} + \frac{\Delta - \Delta_0}{S_T}\right] + \Phi\left[-z_{\alpha/2} + \frac{\Delta - \Delta_0}{S_T}\right] = 1 - \beta$$

Alternatively, it can be written as

$$1 - \Phi\left[z_{\alpha/2} + \frac{\Delta_0 - \Delta}{S_T}\right] + \Phi\left[-z_{\alpha/2} + \frac{\Delta_0 - \Delta}{S_T}\right] = 1 - \beta$$

The term  $\Phi\left[-z_{\alpha/2} + \frac{\Delta_0 - \Delta}{S_T}\right]$  has very little contribution to the power. Therefore, we can ignore this term.

$$\Phi\left[-z_{\alpha/2} + \frac{\Delta_0 - \Delta}{S_T}\right] = \beta$$

$$-z_{\alpha/2} + \frac{\Delta_0 - \Delta}{S_T} = z_{\beta}$$

$$\left(\frac{\Delta_0 - \Delta}{S_T}\right)^2 = (z_{\alpha/2} + z_{\beta})^2$$

$$(\Delta_0 - \Delta)^2 = (z_{\alpha/2} + z_{\beta})^2 S_T^2$$

$$(\Delta_0 - \Delta)^2 = (z_{\alpha/2} + z_{\beta})^2 \left( \frac{\hat{\mu}_2}{s_2} + \frac{\hat{\mu}_1}{s_1} + \frac{\hat{\mu}_2^2 \hat{\phi}_2}{N_2} + \frac{\hat{\mu}_1^2 \hat{\phi}_1}{N_1} \right) \quad (7.16)$$

(B) Method using logarithmic transformation of Wald test statistic:

The logarithmic transformation is usually applied for skewness correction and variance stabilization. The estimate of log transformed Wald test statistic [129] is based on the quantity

$$U = \ln\left(\frac{\hat{\mu}_2}{\hat{\mu}_1}\right) = \ln(\hat{\mu}_2) - \ln(\hat{\mu}_1) = \ln\left(\frac{Y_2}{s_2}\right) - \ln\left(\frac{Y_1}{s_1}\right)$$

$\frac{Y_1}{s_1}$  and  $\frac{Y_2}{s_2}$  have asymptotic normal distributions,  $N\left(\mu_1, \frac{\mu_1}{s_1} + \frac{\mu_1^2 \phi_1}{N_1}\right)$  and  $N\left(\mu_2, \frac{\mu_2}{s_2} + \frac{\mu_2^2 \phi_2}{N_2}\right)$ , respectively. Therefore, by using Delta method,  $\ln\left(\frac{Y_1}{s_1}\right)$  and  $\ln\left(\frac{Y_2}{s_2}\right)$  have respectively asymptotic normal distributions,  $N\left(\ln(\mu_1), \frac{1}{s_1 \mu_1} + \frac{\phi_1}{N_1}\right)$  and  $N\left(\ln(\mu_2), \frac{1}{s_2 \mu_2} + \frac{\phi_2}{N_2}\right)$ . The variance of  $U$  is given by

$$\sigma_U^2 = \text{var}(U) = \frac{1}{s_2 \mu_2} + \frac{1}{s_1 \mu_1} + \frac{\phi_2}{N_2} + \frac{\phi_1}{N_1}$$

The estimate of  $\sigma_U^2$  (standard error under  $H_1$ ) is given by

$$S_U^2 = \frac{1}{s_2 \hat{\mu}_2} + \frac{1}{s_1 \hat{\mu}_1} + \frac{\hat{\phi}_2}{N_2} + \frac{\hat{\phi}_1}{N_1} = \frac{1}{Y_2} + \frac{1}{Y_1} + \frac{\hat{\phi}_2}{N_2} + \frac{\hat{\phi}_1}{N_1}$$

Then, the log transformed Wald test with unequal sample sizes and dispersion parameters is given by

$$z_{w2} = \frac{U}{S_U} = \frac{\ln\left(\frac{Y_2}{Y_1}\right) - \ln(w)}{\sqrt{\frac{1}{Y_2} + \frac{1}{Y_1} + \frac{\hat{\phi}_2}{N_2} + \frac{\hat{\phi}_1}{N_1}}} \quad (7.17)$$

We reject null hypothesis when  $|z_{w2}| > z_{1-\alpha/2}$ . The power of the two-sided test is given by  $1 - \beta = \Pr[|z_{w2}| > z_{\alpha/2} | H_1 \text{ is true}]$ .

$$[\ln(\delta_0) - \ln(\delta)]^2 = (z_{\alpha/2} + z_{\beta})^2 \left( \frac{1}{s_2 \hat{\mu}_2} + \frac{1}{s_1 \hat{\mu}_1} + \frac{\hat{\phi}_2}{N_2} + \frac{\hat{\phi}_1}{N_1} \right) \quad (7.18)$$

The above equation can be used to estimate power for the different input parameters. For example, if  $\mu_1 = \mu_2 = 20$ ,  $\phi_1 = \phi_2 = 0.4$ ,  $s_1 = 18.5$ ,  $s_2 = 21.5$ ,  $N_1 = N_2 = 20$ , then power achieved is 0.9043. To find the optimal allocation of samples, the method may not be appropriate. However, if  $s_{ij} = 1$ , then  $s_i = N_i$ , and the above equation can be written as

$$[\ln(\delta_0) - \ln(\delta)]^2 = (z_{\alpha/2} + z_{\beta})^2 \left( \frac{1}{N_2 \hat{\mu}_2} + \frac{1}{N_1 \hat{\mu}_1} + \frac{\hat{\phi}_2}{N_2} + \frac{\hat{\phi}_1}{N_1} \right) \quad (7.19)$$

In this case, the above equation can be used to find optimal sample allocation with sample size ratio fixed as well as optimal sample allocation for fixed cost to get maximum power and minimum cost for a fixed power. The method using log-transformed Wald test will be equivalent to the method discussed in next section. The methods using Wald test and log-transformed Wald test have been used for testing the significance of single feature. These methods can be extended for testing the significance of multiple features.

## 2. Sample size calculation using generalized linear model based on negative binomial distribution

### A. Sample size calculation for testing a single feature

The generalized linear model [130, 131] theory has been used to estimate sample size using negative binomial distribution [132, 133]. The derivation of sample size formula has been discussed previously in many works. For example, score test has been used for power and sample size calculation in Hart et al. [103]. The statistical properties of the test satisfy the following formula

$$[\ln(\delta)]^2 = (z_{\alpha/2} + z_{\beta})^2 \left[ \frac{\left(\frac{1}{\mu_1} + \phi_1\right)}{N_1} + \frac{\left(\frac{1}{\mu_2} + \phi_2\right)}{N_2} \right] \quad (7.20)$$

where  $\ln(\delta)$  is the desired log fold change;  $\mu_1$  and  $\mu_2$  are the average expected count in groups  $G_1$  and  $G_2$ , respectively;  $\phi_1$  and  $\phi_2$  are the dispersion parameters in groups  $G_1$  and  $G_2$ , respectively;  $N_1$  and  $N_2$  are the number of samples in groups  $G_1$  and  $G_2$ , respectively;  $z_p$  is the upper 100(p)th percentile of standard normal distribution. Biological coefficient of variation is the square root of dispersion parameter.

We followed similar approaches as discussed in Chapter 6. We considered the following two aspects for calculating sample size:

**A1. Sample size allocation without cost constraint**

Let the sample size ratio  $\left(\frac{N_2}{N_1} = r \geq 1\right)$  between two groups is fixed in advance.

Then, we use the starting sample size  $N_1$  that would be the smallest integer satisfying the inequality

$$N_{1Z} = \left[ \left( \frac{1}{\mu_1} + \phi_1 \right) + \left( \frac{1}{\mu_2} + \phi_2 \right) / r \right] (z_{\alpha/2} + z_{\beta})^2 / (\log \Delta)^2 \quad (7.21)$$

Then, an incremental search can be done to obtain the target power. An example of sample sizes and exact power obtained for different sample size ratio (1, 2 and 3) and different fold change (1.5, 2, 2.5 and 3) are shown in Table 7.1. We assumed that the expected read counts ( $\mu_1 = \mu_2 = 20$ ) and dispersion parameter ( $\phi_1 = \phi_2 = 0.4$ ) for both the groups are same. The value of  $\alpha$  chosen is 0.05 and target power is 0.9.

**Table 7.1.** Sample sizes ( $N_1, N_2$ ) and power computed with parameters  $\mu_1 = \mu_2 = 20$ ,  $\phi_1 = \phi_2 = 0.4$ ,  $\alpha = 0.05$ ,  $1 - \beta = 0.90$  for different combinations of  $r$  and  $\Delta$

		$\Delta$											
		1.5			2			2.5			3		
$r$		$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power
1		58	58	0.9023	20	20	0.9045	12	12	0.9171	8	8	0.9058
2		44	88	0.9055	15	30	0.9045	9	18	0.9171	6	12	0.9058
3		39	117	0.9047	14	42	0.9175	8	24	0.9171	6	18	0.9350

**A2. Sample allocation with maximum power for a fixed cost**

A hypothetical example of cost model for case-control study has been illustrated.

Suppose there are  $N_1$  controls (group 1) and  $N_2$  cases (group 2). The overall cost of the study comprises of the following components:

(i) The sample procurement cost:  $c_1 = a_{11}N_1 + a_{12}N_2$

where  $a_{11}$  is the sample procurement cost per control sample and  $a_{12}$  is the sample procurement cost per case sample.

(ii) Cost for library preparation and quality control:  $c_2 = a_2(N_1 + N_2)$

(Assuming equal costs for both cases and controls,  $a_2$  is the cost for library and quality control per sample)

(iii) Sequencing cost:  $c_3 = a_3(N_1 + N_2)/m$

(Given the alignment rate is  $m$  and average cost per million reads mapped to the genes for one sample is  $a_3$ )

(iv) Cost of analysis:  $c_4 = a_4(N_1 + N_2)$  assuming  $a_4$  is the average cost per sample for data analysis.

The total cost can be written in the form of  $C = C_1N_1 + C_2N_2$ , where  $C_1$  and  $C_2$  are the average cost per sample in control and case, respectively. A hypothetical example to show various costs involved in conducting RNA-Seq experiments is shown below in Table 7.2.

**Table 7.2.** A hypothetical example of various costs involved in RNA-Seq experiment

Services	Price per sample (in USD)
RNA isolation from tissue	50
Library preparation	400
Sample and Library QC	50
Sequencing cost	250
Bioinformatics Analysis	250
<b>Cost per sample</b>	<b>1000</b>

We followed the same procedure as given in previous chapter (A2.1 of Chapter

6). We used  $\sqrt{\frac{1}{\mu_1} + \phi_1}$  and  $\sqrt{\frac{1}{\mu_2} + \phi_2}$  in place of  $\sigma_1$  and  $\sigma_2$ , respectively. We find

the optimal allocation giving the maximum power. A hypothetical example of

sample sizes and power obtained for a fixed cost  $C$  (30000, 40000 and 50000) with different input parameters is given in Table 7.3.

**Table 7.3.** Sample sizes ( $N_1, N_2$ ) and the power obtained with parameters  $\mu_1 = \mu_2 = 20$ ,  $\phi_1 = \phi_2 = 0.4$ ,  $\alpha = 0.05$ ,  $C_1 = 1000$ ,  $C_2 = 1100$  for different values of fixed cost  $C$  and fold change

C	$\Delta$											
	1.5			2			2.5			3		
	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power
30000	14	14	0.3591	14	14	0.7805	14	14	0.9509	14	14	0.9912
40000	19	19	0.4614	19	19	0.8897	19	19	0.9878	19	19	0.9990
50000	23	24	0.5444	23	24	0.9431	23	24	0.9967	23	24	0.9999

#### B. Sample size calculation for testing multiple features

The sample size formula given in previous section is applicable to testing the significance of single feature such as gene. However, the experimenters are usually interested in testing the significance of multiple genes. The possible outcomes for testing multiple hypotheses has been given in Table 6.5. The sample size calculation method for testing single feature has been extended for the multiple features. We have used a different method for controlling FDR [128, 134]. The marginal type I error over all the genes is given by

$$\alpha^* = \frac{m_1 q}{m_0(1-q)} \quad (7.22)$$

where,  $m_1$  is the expected number of significant features,  $m_0$  is the number of true null hypotheses (unknown) and  $q$  is the FDR. We use  $\alpha^*$  in place of  $\alpha$  in the sample size formula for multiple feature testing. Therefore, we need extra input parameters, namely, FDR, number of features ( $m$ ) and expected number of DE features ( $m_1$ ). The above equation can be rewritten as

$$\alpha^* = \frac{\pi_1 q}{\pi_0(1-q)} \quad (7.23)$$

where  $\pi_1$  is the expected proportion of significant features,  $\pi_0$  is the proportion of true null hypotheses.

#### B1. Sample size allocation without cost constraint

We have calculated sample sizes and exact power obtained with different sample size ratio (1, 2 and 3) and fold change (1.5, 2, 2.5 and 3) for RNA-Seq experiment (Table 7.4). We have assumed that there are 10000 features and the expected number of significant features is 100. Further, we assumed that the expected read counts ( $\mu_1 = \mu_2 = 20$ ) and dispersion parameter ( $\phi_1 = \phi_2 = 0.4$ ) for both the groups are same. The value of  $\alpha$  and power are respectively 0.05 and 0.9. After controlling the FDR, we obtained the value of  $\alpha^*$  ( $\alpha^* = 0.00053$ ).

**Table 7.4.** Sample sizes ( $N_1, N_2$ ) and power computed with parameters  $m_1 = 100, m = 10000, \mu_1 = \mu_2 = 20, \phi_1 = \phi_2 = 0.4, \alpha = 0.05, q = 0.05, 1 - \beta = 0.90$  for different combinations of  $r$  and  $\Delta$

	$\Delta$											
	1.5			2			2.5			3		
$r$	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power
1	124	124	0.9023	43	43	0.9077	25	25	0.9139	17	17	0.9050
2	93	186	0.9023	32	64	0.9046	19	38	0.9188	13	26	0.9126
3	83	249	0.9040	29	87	0.9122	17	51	0.9212	12	36	0.9263

#### B2. Sample allocation with maximum power for a fixed cost

A hypothetical example of sample sizes and power obtained for a fixed cost  $C$  for multiple features is given in Table 7.5.

**Table 7.5.** Sample sizes ( $N_1, N_2$ ) and the power obtained with parameters  $m_1 = 100, m = 10000, \mu_1 = \mu_2 = 20, \phi_1 = \phi_2 = 0.4, \alpha = 0.05, q = 0.05, C_1 = 1000, C_2 = 1100$  for different values of fixed cost  $C$  and fold change

	$\Delta$											
	1.5			2			2.5			3		
<b>C</b>	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power	$N_1$	$N_2$	Power
30000	14	14	0.031083	14	14	0.232548	14	14	0.559463	14	14	0.807493
40000	19	19	0.054654	19	19	0.389931	19	19	0.772098	19	19	0.943344
50000	23	24	0.081829	23	24	0.530618	23	24	0.888161	23	24	0.984153

We have developed shiny apps for calculating sample size and power based on the above discussed methods. The apps will be useful to the experimenters in the designing of their experiments.

### **A shiny app for sample size estimation based on Poisson-log normal distribution**

There are many tools and applications available for sample size calculations for RNA-Seq experiments. One of the examples is Scotty [104], that performs sample size calculation and power analysis under cost constraint for RNA-Seq experiments. In this method, the read counts are assumed to follow Poisson-Log normal distribution. The original programs were written in MATLAB. However, MATLAB is a proprietary software. Therefore, on a similar line, we have made an improved user defined shiny application using R which is a freely available software. We have implemented the method in R and C++ and, used shiny for making the application. The program is more efficient in terms of computational time. It will be easier for the experimenters to calculate the sample size required for conducting the experiments according to the budget. The researchers can



use the app for writing grants and conducting research projects, that will save resources in terms of cost and time.

The experimenter can design the RNA-Seq experiment based on the pilot experimental data. In this method, we assume the count data is modelled by Poisson-log normal distribution. The data is normalized to the median value of all samples. Then, the estimates of the sequencing depth parameters are obtained by fitting Poisson log normal model. There are two sources of variation: biological and technical (Non-Poisson and Poisson variance). It optimizes the read depth and number of replicates. A screenshot of the app is shown below:

The screenshot displays the 'Sample Size Calculator for RNA-Seq Experiments using Pilot Data' Shiny application. The interface is split into two main areas. On the left, there is a form for inputting parameters, and on the right, there are navigation tabs for different analysis outputs.

**Input Parameters (Left Panel):**

- Choose file to upload data:** A file upload button labeled 'Browse...' with the text 'No file selected' below it.
- Cost per sample in group 1:** Input field with value '0'.
- Cost per sample in group 2:** Input field with value '0'.
- Cost per million reads:** Input field with value '0'.
- Total budget:** Input field with value 'Inf'.
- Fold change:** Input field with value '2'.
- P value cut off:** Input field with value '0.01'.
- Minimum genes detected:** Input field with value '50'.
- Maximum replication:** Input field with value '10'.
- Minimum reads per replication:** Input field with value '0'.
- Maximum reads per replication:** Input field with value '0'.
- Minimum % unbiased genes:** Input field with value '50'.
- Power bias cutoff (in %):** Input field with value '50'.
- Rate of alignment (in %):** Input field with value '50'.
- Submit:** A button at the bottom of the form.

**Navigation Tabs (Right Panel):**

- Result Summary
- MDS plot
- Rarefaction Plot
- Power Plot
- Excluded Plot
- Bias Plot
- Allowed Plot
- Power Plot (Cheapest Allowed Experiment)
- Power Plot (Most Powerful Experiment)
- Scatter Plot

**Figure 7.1.** Shiny application to calculate sample size for RNA-Seq experiments using pilot data

The inputs to be provided by the user are as follows:

(i) Pilot data: The user must upload the data in a prescribed format as discussed in the supplementary section. We have also provided some datasets that can be used as pilot data. We encourage the user to provide the case-control data. If the

data consists of either only control or only case, there must be at least two replicates. After the data is uploaded, please specify the other inputs.

(ii) Additional information of data: The user has to upload the additional information of the data containing the names of samples and group variables.

(iii) After the file is uploaded, the user has to select the name of variable for comparison.

(iv) Cost per control sample

(v) Cost per case sample

(vi) Cost per million reads

(vii) Total budget: Please specify the budget constraint. We will calculate the power achieved under the given budget constraint. The default value in “Inf” meaning no budget constraint.

(viii) Desired fold change

(ix) The significance level (the default value is 0.05)

(x) Minimum number of genes to be detected

(xi) Maximum replication

(xii) Minimum reads per replication

(xiii) Maximum reads per replication

(xiv) Minimum percent unbiased genes

(xv) Power bias cut off

(xvi) Alignment rate

After specifying all the input parameters, we obtain the results showing the experimental design with maximum power as well as cheapest experiment to

achieve a desirable power. We obtain various exploratory plots such as multidimensional scaling plot, rarefaction plot, power analysis curves, power and cost plots for different experimental designs.

We used RNA-Seq data “HumanLiverBleckman” [104, 135] obtained from <http://scotty.genetics.utah.edu/scottyDatasets.php>. This data has three control (female) and three test (male) samples. Each sample was run in two technical replicates. The count data of the technical replicates for each sample were added. We used the following inputs for sample size and power calculation:

**Table 7.6.** The inputs provided for sample size and power calculation

Cost per control replicate	500
Cost per test replicate	600
Significance level	0.05
FC	2
Cost Per Million Reads	1000
Total budget	Inf
Minimum % detected	50
Max number of replicates	10
Minimum reads per replicate	10000000
Maximum reads per replicate	100000000
Minimum % unbiased genes	50
Power bias cutoff (%)	50
Alignment rate (%)	50

The estimates of dispersion parameters in control and test condition are 0.2454 and 0.2539 respectively. Total 90 experimental designs were tested. The least expensive experiment is having 5 replicates with sequencing depth of 10 million reads per replicate (power = 0.55). The most powerful experiment is having 10 replicates with sequencing depth of 100 million reads aligned per replicate (power = 0.94).

## CHAPTER 8

### DISCUSSION AND CONCLUSION

We investigated the design, analysis and sample size estimation methods for high-throughput proteomics and RNA-Seq experiments. We developed various approaches to analyze the kidney proteomics data and studied the heterogeneity issues due to the technical steps in the presence of missing values. Furthermore, we developed an application to standardize proteomics workflow for LC-MS data that will aid in choosing the most appropriate technical methods. We studied the impact of the technical variability on the study design of proteomics experiments. We developed an interactive application for differential expression analysis of label-free LC-MS proteomics data. The application can analyze the data at protein as well as peptide level using various statistical tests. It can also handle the missing values and adjust the effects of additional covariates. Furthermore, we proposed sample size calculation methods under allocation and budget constraints for detecting differentially expressed features in proteomics experiments. We developed apps to compute sample sizes based on various input parameters provided. In future, we will come up with more methods of sample size calculation methods applicable to more than two class comparison including additional covariates. We studied various methods of sample size calculations in RNA-Seq experiments based on different models. We investigated

the estimation of dispersion parameter and sample size methods. We developed different apps to compute sample sizes for conducting future RNA-Seq experiments under different constraints. In future, we will develop the design, analysis and sample size calculation methods for other biological studies such as single-cell sequencing experiments.

## REFERENCES

1. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. Vienna, Austria.
2. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, 12(2):115-21.
3. Winston C, Cheng J, Allaire JJ, Xie Y, McPherson J (2018). shiny: Web Application Framework for R. R package version 1.1.0. <https://CRAN.R-project.org/package=shiny>.
4. Srivastava S, Merchant M, Rai A, Rai SN (2019). Standardizing Proteomics Workflow for Liquid Chromatography-Mass Spectrometry: Technical and Statistical Considerations. *Journal of Proteomics & Bioinformatics*, 12:048-55.
5. Srivastava S, Merchant M, Rai A, Rai SN (2019). Interactive Web Tool for Standardizing Proteomics Workflow for Liquid Chromatography-Mass Spectrometry Data. *Journal of Proteomics and Bioinformatics*, 12:085-7.
6. Dell RB, Holleran S, Ramakrishnan R (2002). Sample size determination. *ILAR J*, 43(4):207-13.
7. Gogtay NJ (2010). Principles of sample size calculation. *Indian J Ophthalmol*, 58(6):517-8.
8. Kadam P, Bhalerao S (2010). Sample size calculation. *Int J Ayurveda Res*, 1(1):55-7.

9. Leek JT, Storey JD (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724-35.
10. Kukurba KR, Montgomery SB (2015). RNA Sequencing and Analysis. *Cold Spring Harb Protoc*, 2015(11):951-69.
11. Anderson NL, Anderson NG (1998). Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis*, 19(11):1853-61.
12. Wilkins MR (2009). Hares and tortoises: the high- versus low-throughput proteomic race. *Electrophoresis*, 30 Suppl 1:S150-5.
13. Fliser D, Novak J, Thongboonkerd V, Argilés A, Jankowski V, Girolami MA (2007). Advances in urinary proteome analysis and biomarker discovery. *J Am Soc Nephrol*, 18.
14. Hanash S (2003). Disease proteomics. *Nature*, 422.
15. McGregor E, Dunn MJ (2006). Proteomics of the heart: unraveling disease. *Circ Res*, 98.
16. Neilson KA, Gammulla CG, Mirzaei M, Imin N, Haynes PA (2010). Proteomic analysis of temperature stress in plants. *Proteomics*, 10(4):828-45.
17. Komatsu S, Mock H-P, Yang P, Svensson B (2013). Application of proteomics for improving crop protection/artificial regulation. *Front Plant Sci*, 4.
18. Liu W, Gray S, Huo Y, Li L, Wei T, Wang X (2015). Proteomic analysis of interaction between a plant virus and its vector insect reveals new functions of hemipteran cuticular protein. *Mol Cell Proteomics*, 14.
19. Wang H, Wu K, Liu Y, Wu Y, Wang X (2015). Integrative proteomics to understand the transmission mechanism of Barley yellow dwarf virus-GPV by its insect vector *Rhopalosiphum padi*. *Sci Rep*, 5.

20. Ducret A, Van Oostveen I, Eng JK, Yates JR 3rd, Aebersold R (1998). High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci*, 7(3):706-19.
21. Mallick P, Kuster B (2010). Proteomics: a pragmatic perspective. *Nat Biotechnol*, 28(7):695-709.
22. Van Oudenhove L, Devreese B (2013). A review on recent developments in mass spectrometry instrumentation and quantitative tools advancing bacterial proteomics. *Appl Microbiol Biotechnol*, 97(11):4749-62.
23. Washburn MP, Wolters D, Yates JR 3rd (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19(3):242-7.
24. Wysocki VH, Resing KA, Zhang Q, Cheng G (2005). Mass spectrometry of peptides and proteins. *Methods*, 35(3):211-22.
25. Calderon-Gonzalez KG, Hernandez-Monge J, Herrera-Aguirre ME, Luna-Arias JP (2016). Bioinformatics Tools for Proteomics Data Interpretation. *Adv Exp Med Biol*, 919:281-341.
26. Blattmann P, Heusel M, Aebersold R (2016). SWATH2stats: An R/Bioconductor Package to Process and Convert Quantitative SWATH-MS Proteomics Data for Downstream Analysis Tools. *PLoS One*, 11(4):e0153160.
27. Tyanova S, Temu T, Carlson A, Sinitcyn P, Mann M, Cox J (2015). Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics*, 15(8):1453-6.
28. Gatto L, Breckels LM, Naake T, Gibb S (2015). Visualization of proteomics data using R and Bioconductor. *Proteomics*, 15(8):1375-89.
29. Gatto L, Christoforou A (2014). Using R and Bioconductor for proteomics data analysis. *Biochim Biophys Acta*, 1844(1 Pt A):42-51.



30. Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics*, 13(9):2513-26.
31. Neuhauser N, Michalski A, Cox J, Mann M (2012). Expert system for computer-assisted annotation of MS/MS spectra. *Mol Cell Proteomics*, 11(11):1500-9.
32. Cox J, Mann M (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26(12):1367-72.
33. Wang X, Wu ZP, Zhang XG (2010). Isoform Abundance Inference Provides A More Accurate Estimation of Gene Expression Levels in Rna-Seq. *Journal of Bioinformatics and Computational Biology*, 8:177-92.
34. Almeida AM, Bassols A, Bendixen E, Bhide M, Ceciliani F, Cristobal S, et al. (2015). Animal board invited review: advances in proteomics for animal and food sciences. *Animal*, 9(1):1-17.
35. Hu J, Rampitsch C, Bykova NV (2015). Advances in plant proteomics toward improvement of crop productivity and stress resistance. *Front Plant Sci*, 6:209.
36. Lippolis JD, Reinhardt TA (2010). Utility, limitations, and promise of proteomics in animal science. *Vet Immunol Immunopathol*, 138(4):241-51.
37. Vanderschuren H, Lentz E, Zainuddin I, Gruissem W (2013). Proteomics of model and crop plant species: status, current limitations and strategic advances for crop improvement. *J Proteomics*, 93:5-19.
38. McLeish KR, Merchant ML, Klein JB, Ward RA (2013). Technical note: proteomic approaches to fundamental questions about neutrophil biology. *J Leukoc Biol*, 94(4):683-92.

39. Gstaiger M, Aebersold R (2009). Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet*, 10(9):617-27.
40. Hathout Y (2007). Approaches to the study of the cell secretome. *Expert Rev Proteomics*, 4(2):239-48.
41. Zhang G, Annan RS, Carr SA, Neubert TA (2014). Overview of peptide and protein analysis by mass spectrometry. *Curr Protoc Mol Biol*, 108:10 21 1-30.
42. Pitt JJ (2009). Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin Biochem Rev*, 30(1):19-34.
43. Xu F, Zou L, Liu Y, Zhang Z, Ong CN (2011). Enhancement of the capabilities of liquid chromatography-mass spectrometry with derivatization: general principles and applications. *Mass Spectrom Rev*, 30(6):1143-72.
44. Chait BT (2006). Chemistry. Mass spectrometry: bottom-up or top-down? *Science*, 314(5796):65-6.
45. Ong SE, Mann M (2005). Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*, 1(5):252-62.
46. Bantscheff M, Lemeer S, Savitski MM, Kuster B (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*, 404(4):939-65.
47. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007). Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*, 389(4):1017-31.
48. Xie F, Liu T, Qian WJ, Petyuk VA, Smith RD (2011). Liquid chromatography-mass spectrometry-based quantitative proteomics. *J Biol Chem*, 286(29):25443-9.

49. Toby TK, Fornelli L, Kelleher NL (2016). Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem (Palo Alto Calif)*, 9(1):499-519.
50. Clough T, Key M, Ott I, Ragg S, Schadow G, Vitek O (2009). Protein quantification in label-free LC-MS experiments. *J Proteome Res*, 8(11):5275-84.
51. Clough T, Thaminy S, Ragg S, Aebersold R, Vitek O (2012). Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinformatics*, 13 Suppl 16:S6.
52. Serang O, Kall L (2015). Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. *J Proteome Res*, 14(10):4099-103.
53. Rubin DB (1976). Inference and missing data. *Biometrika*, 63(3):581–92.
54. Schwammle V, Leon IR, Jensen ON (2013). Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates. *J Proteome Res*, 12(9):3874-83.
55. Webb-Robertson BJ, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, et al. (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res*, 14(5):1993-2001.
56. Piehowski PD, Petyuk VA, Orton DJ, Xie F, Moore RJ, Ramirez-Restrepo M, et al. (2013). Sources of technical variability in quantitative LC-MS proteomics: human brain tissue sample analysis. *J Proteome Res*, 12(5):2128-37.
57. Glaab E, Schneider R (2015). RepExplore: addressing technical replicate variance in proteomics and metabolomics data analysis. *Bioinformatics*, 31(13):2235-7.
58. Hobeika L, Barati MT, Caster DJ, McLeish KR, Merchant ML (2017). Characterization of glomerular extracellular matrix by proteomic analysis of laser-captured microdissected glomeruli. *Kidney Int*, 91(2):501-11.

59. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol*, 32(3):223-6.
60. Fox J, Weisberg S. (2011) An {R} Companion to Applied Regression. Second ed. Thousand Oaks CA: Sage.
61. Karpievitch YV, Dabney AR, Smith RD (2012). Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, 13 Suppl 16:S5.
62. Lazar C (2015). imputeLCMD: A collection of methods for left-censored missing data imputation. R package version 2.0. <https://CRAN.R-project.org/package=imputeLCMD>.
63. Ported to R by Novo AA. Original by Schafer JL (2013). norm: Analysis of multivariate normal datasets with missing values. R package version 1.0-9.5. <https://CRAN.R-project.org/package=norm>.
64. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007). pcaMethods--a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164-7.
65. Hastie T, Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. and Botstein, D. (1999). Imputing Missing Data for Gene Expression Arrays. Technical Report. Stanford University Statistics Department.
66. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520-5.
67. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185-93.
68. Bolstad B (2017). preprocessCore: A collection of pre-processing functions. R package version 1.40.0. <https://github.com/bmbolstad/preprocessCore>.

69. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96-104.
70. Eriksson J, Fenyo D (2010). Modeling experimental design for proteomics. *Methods Mol Biol*, 673:223-30.
71. Bittremieux W, Tabb DL, Impens F, Staes A, Timmerman E, Martens L, et al. (2018). Quality control in mass spectrometry-based proteomics. *Mass Spectrom Rev*, 37(5):697-711.
72. Ivanov AR, Colangelo CM, Dufresne CP, Friedman DB, Lilley KS, Mechtler K, et al. (2013). Interlaboratory studies and initiatives developing standards for proteomics. *Proteomics*, 13(6):904-9.
73. Alkhas A, Hood BL, Oliver K, Teng PN, Oliver J, Mitchell D, et al. (2011). Standardization of a sample preparation and analytical workflow for proteomics of archival endometrial cancer tissue. *J Proteome Res*, 10(11):5264-71.
74. Dogu E, Taheri SM, Olivella R, Marty F, Lienert I, Reiter L, et al. (2019). MSstatsQC 2.0: R/Bioconductor Package for Statistical Quality Control of Mass Spectrometry-Based Proteomics Experiments. *J Proteome Res*, 18(2):678-86.
75. Stanfill BA, Nakayasu ES, Bramer LM, Thompson AM, Ansong CK, Clauss TR, et al. (2018). Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data. *Mol Cell Proteomics*, 17(9):1824-36.
76. Chiva C, Olivella R, Borrás E, Espadas G, Pastor O, Sole A, et al. (2018). QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories. *PLoS One*, 13(1):e0189209.
77. Chang C, Xu K, Guo C, Wang J, Yan Q, Zhang J, et al. (2018). PANDA-view: An easy-to-use tool for statistical analysis and visualization of quantitative proteomics data. *Bioinformatics*.

78. Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, et al. (2014). MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30(17):2524-6.
79. Goeminne LJE, Gevaert K, Clement L (2018). Experimental design and data-analysis in label-free quantitative LC/MS proteomics: A tutorial with MSqRob. *J Proteomics*, 171:23-36.
80. Sievert C (2018). plotly for R. URL: <https://plotly-r.com>.
81. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47.
82. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK (2016). Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression. *Ann Appl Stat*, 10(2):946-63.
83. Smyth GK (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3.
84. Bates D, Machler M, Bolker BM, Walker SC (2015). Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw*, 67(1):1-48.
85. Singmann H, Bolker B, Jake; W, Frederik; A (2019). afex: Analysis of Factorial Experiments.
86. Lenth R (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means.
87. Biomarkers Definitions Working G (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*, 69(3):89-95.
88. Cairns DA, Barrett JH, Billingham LJ, Stanley AJ, Xinarianos G, Field JK, et al. (2009). Sample size determination in clinical proteomic profiling

- experiments using mass spectrometry for class comparison. *Proteomics*, 9(1):74-86.
89. Zhou C, Simpson KL, Lancashire LJ, Walker MJ, Dawson MJ, Unwin RD, et al. (2012). Statistical considerations of optimal study design for human plasma proteomics and biomarker discovery. *J Proteome Res*, 11(4):2103-13.
  90. Oberg AL, Vitek O (2009). Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res*, 8(5):2144-56.
  91. Levin Y (2011). The role of statistical power analysis in quantitative proteomics. *Proteomics*, 11(12):2565-7.
  92. Welch BL (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3-4):350-62.
  93. Jan SL, Shieh G (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behav Res Methods*, 43(4):1014-22.
  94. van Belle G, Martin DC (1993). Sample Size as a Function of Coefficient of Variation and Ratio of Means. *The American Statistician*, 47(3):165-7.
  95. Smith HF (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, 9:211-2.
  96. Satterthwaite FE (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2(6):110-4.
  97. Guo J-H, Luh W-M (2013). Efficient sample size allocation with cost constraints for heterogeneous-variance group comparison. *Journal of Applied Statistics*, 40(12):2549-63.
  98. Benjamini Y, Hochberg Y (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B*, 57(1):289-300.

99. Marioni J, Mason C, Mane S, Stephens M, Gilad Y (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18:1509 - 17.
100. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344 - 9.
101. Nagalakshmi U, Waern K, Snyder M (2010). RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol*, Chapter 4:Unit-13.
102. Ching T, Huang S, Garmire LX (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *RNA*, 20(11):1684-96.
103. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP (2013). Calculating sample size estimates for RNA sequencing data. *J Comput Biol*, 20(12):970-8.
104. Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT (2013). Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, 29(5):656-7.
105. Wu H, Wang C, Wu Z (2015). PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*, 31(2):233-41.
106. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB (2013). ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *Plos One*, 8(7):e67019.
107. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621-8.
108. Robinson M, McCarthy D, Smyth G (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139 - 40.



109. Robinson M, Smyth G (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881 - 7.
110. Jiang H, Wong WH (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026-32.
111. Robinson M, Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 11:R25.
112. Srivastava S, Chen L (2010). A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res*, 38(17):e170.
113. Wang L, Feng Z, Wang X, Wang X, Zhang X (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136-8.
114. Robinson MD, Smyth GK (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9(2):321-32.
115. Anders S, Huber W (2010). Differential expression analysis for sequence count data. *Genome Biol*, 11:R106.
116. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc*, 8(9):1765-86.
117. Anscombe FJ (1949). The Statistical Analysis of Insect Counts Based on the Negative Binomial Distribution. *Biometrics*, 5(2):165-73.
118. Lawless JF (1987). Negative Binomial and Mixed Poisson Regression. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 15(3):209-25.
119. Fisher RA (1941). The Negative Binomial Distribution. *Annals of Eugenics*, 11(1):182-7.
120. Piegorsch WW (1990). Maximum Likelihood Estimation for the Negative Binomial Dispersion Parameter. *Biometrics*, 46(3):863-7.

121. Clark S, Perry J (1989). Estimation of the negative binomial parameter  $\kappa$  by maximum quasi-likelihood. *Biometrics*, 45(1):309-16.
122. Smyth GK (2003) Pearson's goodness of fit statistic as a score test statistic. In: Goldstein DR, editor. *Statistics and science: a Festschrift for Terry Speed*. Lecture Notes--Monograph Series. Volume 40. Beachwood, OH: Institute of Mathematical Statistics. p. 115-26.
123. Lu J, Tomfohr JK, Kepler TB (2005). Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6:165.
124. Nelder JA (2000). Quasi-likelihood and pseudo-likelihood are not the same thing. *Journal of Applied Statistics*, 27(8):1007-11.
125. Smyth GK, Verbyla AP (1996). A Conditional Likelihood Approach to Residual Maximum Likelihood Estimation in Generalized Linear Models. *Journal of the Royal Statistical Society Series B (Methodological)*, 58(3):565-72.
126. Cox DR, Reid N (1987). Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society Series B (Methodological)*, 49(1):1-39.
127. Ng HKT, Tang M-L (2005). Testing the equality of two Poisson means using the rate ratio. *Stat Med*, 24(6):955-65.
128. Li X, Cooper NGF, Shyr Y, Wu D, Rouchka EC, Gill RS, et al. (2017). Inference and Sample Size Calculations Based on Statistical Tests in a Negative Binomial Distribution for Differential Gene Expression in RNAseq Data. *Journal of Biometrics & Biostatistics*, 08(01).
129. Wald A (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, 54(3):426-82.
130. McCullagh P (1984). Generalized Linear-Models. *Eur J Oper Res*, 16(3):285-92.

131. McCullagh P, Nelder JA. (1989) *Generalized Linear Models*. 2nd Edition ed: Chapman and Hall/CRC.
132. Zhu HY, Lakkis H (2014). Sample size calculation for comparing two negative binomial rates. *Stat Med*, 33(3):376-87.
133. Cundill B, Alexander NDE (2015). Sample size calculations for skewed distributions. *Bmc Med Res Methodol*, 15.
134. Jung SH (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics*, 21(14):3097-104.
135. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y (2010). Sex-specific and lineage-specific alternative splicing in primates. *Genome Res*, 20(2):180-9.

APPENDIX A  
ACRONYMS USED

LC: Liquid chromatography  
MS: Mass spectrometry  
MCAR: Missing completely at random  
MAR: Missing at random  
MNAR: Missing not at random  
MVs: Missing values  
TSM: Tissue storage method  
FFPE: Formalin-fixed paraffin embedded  
FR: Frozen  
TEM: Tissue extraction method  
MAX: Protease MAX  
TX: Triton X-100  
SDS: Sodium dodecylsulfate  
LCMD: Laser capture microdissection  
ETD: Electron-transfer dissociation  
CID: Collision-induced dissociation  
cRAP: Common Repository of Adventitious Proteins  
FDR: False discovery rate  
CV: Coefficient of variation

GLM: General linear model  
ANOVA: Analysis of variance  
ANCOVA: Analysis of covariance  
SS: Sum of squares  
SD: Standard deviation  
PWST: Proteomics Workflow Standardization Tool  
DE: Differential expression/ Differentially expressed  
SSCP: Sample Size Calculator for Proteomics Experiment  
SATP: Statistical Analysis Tool for Proteomics  
NGS: Next-generation sequencing  
ECM: Extracellular matrix  
BH: Benjamini-Hochberg  
BY: Benjamini-Yekutieli  
MDS: Multidimensional scaling  
SVD: Singular value decomposition  
NB: Negative binomial  
MME: Method of moments estimation  
MLE: Maximum likelihood estimation  
MQLE: Maximum quasi-likelihood estimation  
CML: Conditional maximum likelihood  
RPKM: Reads aligned per kilobase of exon per million reads mapped  
FPKM: Fragments per kilobase of exon per million fragments mapped  
TPM: Transcripts per kilobase million  
FC: Fold change  
GLiM: Generalized linear model

## CURRICULUM VITAE

Sudhir Srivastava

**email:** sudhir.srivastava@louisville.edu;

Sudhir.Srivastava@icar.gov.in; sudhir0401bm@gmail.com

### **Education**

Ph.D., Interdisciplinary Studies with Specialization in Bioinformatics, University of Louisville, USA, 2015-2019

M.Sc., Agricultural Statistics, Post-Graduate School, Indian Agricultural Research Institute, New Delhi, India, 2008-2010

B.Sc., Agriculture, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, India, 2004-2008

### **Positions and Employment**

Scientist 2011-Present

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute (IASRI), Indian Council of Agricultural Research (ICAR), New Delhi, India

Student/Research Assistant 2018-19  
Biostatistics and Bioinformatics Facility, James Graham Brown Cancer Center,  
University of Louisville, USA

Executive Member 2015-17  
Indian Society of Agricultural Statistics, New Delhi, India

### **Professional Membership**

Life member (from 2012 onwards)  
Indian Society of Agricultural Statistics, New Delhi, India

### **Honors**

Junior Research Fellowship Award for M.Sc. Program in 2008-2010  
Agricultural Statistics, ICAR, New Delhi, India

Qualified National Eligibility Test in the discipline of 2010-2011  
Statistics and Computer Application, Agricultural  
Scientists Recruitment Board (ASRB), ICAR, New Delhi,  
India

Qualified Agriculture Research Service Examination in 2010-2011  
the discipline of Agricultural Statistics, ASRB, ICAR,  
New Delhi, India

Received the "Certificate of Merit" with Grade 'A' for the Sep 15-Dec 13,  
successful completion of 94th Foundation Course for 2011  
Agricultural Research Service at National Academy of

Agricultural Research Management, Hyderabad, India

Received “ICAR International Fellowship 2014-15” to 2015-2018  
pursue Ph.D. Program in Interdisciplinary Studies with  
Specialization in Bioinformatics at University of  
Louisville, USA

### Research Projects

Project Title	Role	Funding Agency
Methodology for protein structure comparison and its web implementation	Principal Investigator (PI)	ICAR-IASRI, New Delhi, India
A new distributed computing framework for data mining	Co-PI	Department of Electronics and Information Technology, Ministry of Communications and Information Technology, Government of India
Development of a tool for comparison of protein 3D structure using graph theoretic approach	Co-PI	ICAR-IASRI, New Delhi, India
Multilabel functional classification of abiotic stress related proteins in <i>Poaceae</i>	Co-PI	ICAR-IASRI, New Delhi, India



## Papers Published

**Srivastava, S.**, Merchant, M., Rai, A. and Rai, S.N. (2019). Standardizing Proteomics Workflow for Liquid Chromatography-Mass Spectrometry: Technical and Statistical Considerations. *Journal of Proteomics and Bioinformatics*, 12: 048-055. doi: 10.4172/0974276X.1000496.

**Srivastava, S.**, Merchant, M., Rai, A. and Rai, S.N. (2019). Interactive Web Tool for Standardizing Proteomics Workflow for Liquid Chromatography-Mass Spectrometry Data. *Journal of Proteomics and Bioinformatics*, 12: 085-087. doi: 10.4172/0974-276X.1000500.

Grover, M., Mishra, D.C., Sharma, N., **Srivastava, S.**, & Rai, A. (2017). The maximum computational capacity of proteins involved in abiotic stress differs significantly from the proteins not involved in abiotic stress. *National Academy Science Letters*, **40(4)**, 233-235.

**Srivastava, S.**, Lal, S.B., Mishra, D.C., Angadi, U.B., Chaturvedi, K.K., Rai, S.N., & Rai, A. (2016). An efficient algorithm for protein structure comparison using elastic shape analysis. *Algorithms for Molecular Biology*, **11(1)**, 27.

**Srivastava, S.**, Varghese, C., Jaggi, S., & Varghese, E. (2015). Augmented partial diallel cross plans involving two sets of parental lines. *The Indian Journal of Genetics and Plant Breeding*, **75(1)**, 105-109.

Grover, M., Mishra, D.C., Kumar, R., Trivedi, A.K., & **Srivastava, S.** (2014). Computation, Mathematics or Aesthetic Realism: Revisiting the foundations of

modern biology and agriculture. *International Journal of Current Research and Academic Review*, **2(8)**, 175-177.

**Srivastava, S.**, Varghese, C., Jaggi, S., & Varghese, E. (2013). Diallel cross designs for test versus control comparisons. *The Indian Journal of Genetics and Plant Breeding*, **73(2)**, 186-193.

Sharma, A., Lal, S.B., Mishra, D.C., **Srivastava, S.**, & Rai A. (2013). A web Based Software for Synonymous Codon Usage Indices. *International Journal of Information and Computation Technology*, **3(3)**, 147-152.

#### **Papers Accepted for Publication**

Angadi, U.B., Chaturvedi, K.K., **Srivastava, S.** and Rai, A. (2019). A Novel Way of Comparing Protein 3D Structure Using Graph Partitioning Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. [Accepted on January 17, 2019]

Rai, S.N., **Srivastava, S.**, Pan, J., Wu, X., Rai, S.P., Mekmaysy, C.S., DeLeeuw, L., Chaires, J.B., Garbett, N.C (2019). Multi-group diagnostic classification of high-dimensional data using differential scanning calorimetry plasma thermograms. *PLoS One*. [Accepted on August 2, 2019]

#### **Book Chapter Published**

Mishra, D.C., **Srivastava, S.**, Kumar, S., & Rai, A. (2013). Machine Learning Techniques and its Application in Bioinformatics. *Information and Knowledge*

*Management: Tools, Techniques and Practices*, 155-168, New India Publishing Agency.

### **Popular Article Published**

Mishra, D.C., Singh, I., Kumar, S., & **Srivastava, S.** (2015-16). Protein-Protein Interaction. *Sankhiki Vimarsh*, pg. 69-73.

### **Research Project Reports Published**

Angadi, U.B., Chaturvedi, K.K., Grover, M., & **Srivastava, S.** (2017). Development of a Tool for Comparison of Protein 3D Structure using graph theoretic approach. ICAR-Indian Agricultural Statistics Research Institute Publication, IPC: AGENIASRIL201400500024I I.C.A.R.-I.A.S.R.I./P.R.-01/2017; PIMS: XX10767.

**Srivastava, S.**, Lal, S.B., & Mishra, D.C. (2015). Methodology for Protein Structure Comparison and its Web Implementation. ICAR-Indian Agricultural Statistics Research Institute Publication, IPC: AGENIASRISIL201300600007 I.A.S.R.I./P.R.-03/2015.

### **Trainings/ Workshops**

Trainings/ Workshops attended in USA:

- September 12, 2018: The PLAN Workshop “Organizing and Writing a Large-Scale Writing Project such as a Dissertation or Thesis” organized by the

School of Interdisciplinary and Graduate Studies (SIGS) at the Houchens building, University of Louisville.

- April 17-18, 2018: The Targeted Quantitative Proteomics Workshop hosted by the Oklahoma Medical Research Foundation (OMRF) and the University of Oklahoma Health Sciences Center (OUHSC), Oklahoma City, Oklahoma, USA [The workshop was supported by supplements from the NIH to the Oklahoma IDeA Network of Biomedical Research Excellence (INBRE) (P20GM103447-17S1) and Arkansas INBRE (P20GM103429-15S1)]
- March 2-3, 2018: The annual Graduate Student Regional Research Conference (GSRRC) conducted by the School of Interdisciplinary and Graduate Studies at the University of Louisville, USA
- April 21-23, 2017: 16th Annual UT-KBRIN Bioinformatics Summit 2017, jointly sponsored by the University of Tennessee (UT) [Center for Integrative and Translational Genomics, UT Molecular Resource Center], the University of Memphis and the Kentucky Biomedical Research Infrastructure Network at Montgomery Bell State Park, Burns, Tennessee, USA
- July 11-15, 2016: The Sixth NIGMS-funded Short Course on Statistical Genetics and Genomics during at the University of Alabama, Birmingham, USA
- April 8-10, 2016: The UT-KBRIN Bioinformatics Summit 2016, jointly sponsored by the UT, University of Memphis and the KBRIN at Lake Barkley State Resort Park in Cadiz, KY, USA

- October 13-14, 2015: UCSC Genome Browser Workshop at University of Louisville, USA

Trainings attended at Centre for Agricultural Bioinformatics (CABin), ICAR-Indian Agricultural Statistics Research Institute (IASRI), New Delhi, India:

- Jun 17-21, 2014: Advance Training on Discovery Studio software
- Jun 11-13, 2014: Managing SAS Analytical Models Using SAS Model Manager
- Jun 9-10, 2014: Text Analytics Using SAS Text Miner
- Jun 2-6, 2014: Data Flux Data Management Studio: Fast Track
- May 26-30, 2014: SAS Data Integration Studio: Fast Track
- March 5-7, 2014: PBS PRO Training
- Feb 17-21, 2014: Linux and High-Performance Computing Training
- Oct 29–Nov 1, 2013: Advanced Workbench Module of CLC Bio software
- Oct 21–25, 2013: Administrative Module of High Performance Computing
- Aug 26–30, 2013: CLC Bio's Developer Module
- July 2-6, 2013: Working with CLC Bio software
- Nov 6-26, 2012: Winter school on "Recent advances in Quantitative Genetics and Statistical Genomics"

Trainings attended at ICAR-National Bureau of Agriculturally Important Microorganisms (NBAIM), Mau, Uttar Pradesh, India:

- Mar 04-May 22, 2012: Subject Matter Exposure in relation with protein structure prediction and codon usage analysis in bioinformatics

- Feb 23-Mar 03, 2012: National Subject Training on “Bioinformatics in Multi-omics Era: A Microbial Genomics Perspective” under the National Agricultural Bioinformatics Grid (NABG) project of National Agricultural Innovation Project (NAIP)

### **Abstracts/Oral presentations/Poster presentations**

- “Race as an independent factor for survival in breast cancer patients according to analysis of the National Cancer Database (NCDB)” authored by Drew Carl Drennan Murray\*, Shruti Bhandari, Phuong Ngo, Sarah Mudra, Rachana Shirish Lele, **Sudhir Srivastava**, Xiaoyong Wu, Shesh Rai, Mounika Mandadi, Elizabeth Carloss Riley in the ASCO Annual Meeting during May 31 - June 4, 2019 at McCormick Place, Chicago, Illinois, USA: *J Clin Oncol* **37**, 2019 (suppl; abstr e18155).
- “Treatment delays in localized breast cancer, a NCDB analysis” authored by Shruti Bhandari, Phuong Ngo, Sarah Mudra, Drew Carl Drennan Murray, Rachana Shirish Lele, **Sudhir Srivastava**, Xiaoyong Wu, Shesh Rai, Mounika Mandadi, Elizabeth Carloss Riley in the ASCO Annual Meeting during May 31 - June 4, 2019 at McCormick Place, Chicago, Illinois, USA: *J Clin Oncol* **37**, 2019 (suppl; abstr e18023).
- “Impact of Surgery Type on Survival in Breast Cancer Patients” authored by Phuong Tuyet Ngo, Shruti Bhandari, Drew Murray, Sarah Mudra, Xiaoyong Wu, Rachana Shirish Lele, **Sudhir Srivastava**, Mounika Mandadi, Elizabeth Carloss Riley in the ASCO Annual Meeting during May 31 - June 4, 2019 at

McCormick Place, Chicago, Illinois, USA: *J Clin Oncol* **37**, 2019 (suppl; abstr e12072).

- “Statistical issues and challenges in analyzing proteomics data in the presence of missing values and heterogeneity” authored by Shesh N. Rai\*, **Sudhir Srivastava**, Michael Merchant and Anil Rai\* in the 72nd Annual Conference of Indian Society of Agricultural Statistics “Statistics, Informatics, Engineering Interventions and Business opportunities: A Road-Map to Transform Indian Agriculture Towards Prosperity” during December 13-15, 2018 at ICAR-CIAE, Bhopal, India.
- The summary slide entitled “Novel biomarkers in alcoholic hepatitis: analysis of the plasma peptidome/degradome” authored by CE Dolin\*, DW Wilkey, V Vatsalya, **S Srivastava**, CJ McClain, SN Rai, ML Merchant, GE Arteel was selected in the Best of The Liver Meeting 2018 organized by the American Association for the Study of Liver Diseases (AASLD) during November 9-13, 2018 at San Francisco, California, USA. Abstract and poster of the same was presented in the Ohio Valley Chapter of the Society of Toxicology on November 30, 2018 at Kosair Charities Clinical & Translational Research Building, University of Louisville, Louisville, Kentucky, USA.
- Poster presentation entitled “Chronic Arsenic Exposure in a HaCaT Cell Model of Squamous Cell Carcinoma: Altered Splicing Events or Selection of Clones with Specific Isoforms?” by Mayukh Banerjee, Ana P. F. Cardoso, Laila Al-Eryani\*, M. Sayed, Juw W. Park, Shesh N. Rai, Jianmin Pan, **Sudhir Srivastava**, J. Christopher States in the Ohio Valley Chapter of the Society of

Toxicology on November 30, 2018 at Kosair Charities Clinical & Translational Research Building, University of Louisville, Louisville, Kentucky, USA.

- Presentation entitled “Prediction of Hospital Readmission of Diabetic Patients Using Classification Algorithms” by Somesh P. Rai, Arinjita Bhattacharyya, **Sudhir Srivastava**, Samarendra Das, Divya Srivastava, Xiaoyong Wu, Marion McClain, Anand Seth, Shesh N. Rai, Mehmed Kantardzic and Aruni Bhatnagar in the International Conference on Emerging Innovations in Statistics & Operations Research (EISOR) in conjunction with 38th Annual Convention of Indian Society for Probability and Statistics (ISPS) & 4th Convention of Indian Association for Reliability and Statistics (IARS) during December 27-30, 2018 organized by Department of Statistics, Maharshi Dayanand University, Rohtak, Haryana, India.
- “Differential expression of long non-coding RNA in colon adenocarcinoma RNA-sequence data set” authored by Stephen J. O'Brien\*, Theodore Kalbfleisch, **Sudhir Srivastava**, Shesh Rai, and Susan Galandiuk. In: Proceedings of the 110th Annual Meeting of the American Association for Cancer Research; March 29 - April 3, 2019; Atlanta, Georgia (GA). Philadelphia (PA), USA: AACR; 2019. Abstract nr 1817.
- **Sudhir Srivastava\***, D.C. Mishra, S.B. Lal and U.B. Angadi (2015). “A Tool for Protein Structure Comparison using Elastic Shape Analysis” in 68th Annual Conference on Statistics and Informatics in Agricultural Research of the Indian Society of Agricultural Statistics during Jan 29-31, 2015 at ICAR-IASRI, New Delhi.



- D.C. Mishra\*, Veena Rajan, **Sudhir Srivastava**, Sanjeev Kumar and Anil Rai (2015). "Support Vector Machine based Prediction Model for Protein-protein Interaction using Protein 3D Structure and Physicochemical Properties" in 68th Annual Conference on Statistics and Informatics in Agricultural Research of the Indian Society of Agricultural Statistics during Jan 29-31, 2015 at ICAR-IASRI, New Delhi.
- Anu Sharma\*, S.B. Lal, D.C. Mishra, **Sudhir Srivastava** and Anil Rai (2013). "A web Based Software for Synonymous Codon Usage Indices" in the INTERNATIONAL CONFERENCE on Advancements in Computing Sciences, Information Techniques & Emerging E-Learning Technologies during Oct 5-6, 2013 at JNU, New Delhi, India.
- **Sudhir Srivastava\***, Cini Varghese, Seema Jaggi and Eldho Varghese (2012). "Augmented partial diallel cross plans involving two sets of parental lines" in the 66<sup>th</sup> International Conference on Statistics and Informatics in Agricultural Research during Dec 18-20, 2012 at Indian Agricultural Statistics Research Institute (IASRI), New Delhi, India.
- **Sudhir Srivastava**, Cini Varghese, Seema Jaggi and Eldho Varghese\* (2010). "Diallel cross designs for test versus control comparisons" in International conference on Development and Applications of Statistics in Emerging Areas of Science & Technology (ICDASEAST) along with XXX Annual Convention of Indian Society for Probability and Statistics (ISPS) during Dec 8-10, 2010 hosted by Department of Statistics, University of Jammu, Jammu, India.

\* denotes the presenter

### **Technical Knowledge**

Working Expertise: Statistical data analysis, NGS Data Analysis, Proteomics Data Analysis, GWAS Data Analysis, Data Mining

Software Packages and Programming Languages: R, SAS, SPSS, MATLAB, CLC-Bio, Discovery Studio