1-2020

Statistical approaches of gene set analysis with quantitative trait loci for high-throughput genomic studies.

Samarendra Das
University of Louisville

STATISTICAL APPROACHES OF GENE SET ANALYSIS WITH QUANTITATIVE TRAIT LOCI FOR HIGH-THROUGHPUT GENOMIC STUDIES

By

Samarendra Das
B.Sc. (Agriculture), Orissa University of Agriculture and Technology, 2005-09
M.Sc. (Agricultural Statistics), Indian Agricultural Research Institute, 2009-11

A Dissertation
Submitted to the Faculty of the
Graduate School at the University of Louisville
In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy
in Interdisciplinary Studies: Specialization in Bioinformatics

Interdisciplinary Studies
University of Louisville
Louisville, Kentucky, USA

December 2020

STATISTICAL APPROACHES OF GENE SET ANALYSIS WITH QUANTITATIVE TRAIT LOCI FOR HIGH-THROUGHPUT GENOMIC STUDIES

By

Samarendra Das
B.Sc. (Agriculture), Orissa University of Agriculture and Technology, 2005-09
M.Sc. (Agricultural Statistics), Indian Agricultural Research Institute, 2009-11

A Dissertation Approved on

November 20, 2020

by the following Dissertation Committee:

_____

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

_____

Eric C. Rouchka, D.Sc.

_____

Craig J. McClain, M.D.

_____

Michael L. Merchant, Ph.D.

_____

Subhadip Pal, Ph.D.

DEDICATION

This dissertation is dedicated

to my parents

late Mr. Bishnu Charan Das and Mrs. Sashirekha Das

to my uncle

late Mr. Sridhar Charan Das

to all my teachers,

and to my wife

Rupali Das

for their love, selfless support, constant guidance, and encouragement

in all my endeavors.

ACKNOWLEDGEMENTS

I would like to express my sincere respect and gratitude to my thesis advisor, Dr. Shesh N. Rai, for his patience, guidance, and continuous support throughout this period. Dr. Rai provided opportunities and encouragement for me to perform independent analytical thinking in research and to further formulate different research problems. Dr. Rai is a true mentor who has included me in several projects with colleagues from the University of Louisville, USA, which helped me to learn about the multidisciplinary approaches to solve research problems. I would like to thank my mentors, Dr. Eric C. Rouchka, Dr. Craig J. McClain, Dr. Michael L. Merchant, and Dr. Subhadip Pal for their continuous guidance, motivation, and support.

I would like to thank the Education Division, Indian Council of Agricultural Research (ICAR), India, for providing financial support through the Netaji Subhas ICAR-International fellowship, OM No. 18(02)/2016-EQR/Edn., to pursue my Ph.D. in University of Louisville, USA. I would like to acknowledge the financial support obtained from ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India, and James Graham Brown Cancer Center (through Student Assistantship), University of Louisville, USA. I am thankful to Dr. McClain for covering the Ph.D. tuition fees through National Institutes of Health (NIH), USA grant (5P20GM113226, PI: McClain). I would like to thank the University of Louisville,

USA, for providing me a great environment for learning, and facilities for conducting my Ph.D. studies and research work. I am extremely thankful to my wonderful teachers from various departments at the University of Louisville for their selfless support, and encouragement during my course work. It was my utmost pleasure to be a part of this great University. I would like to give thanks to my cohort mates for making the course work more enjoyable. I would also like to extend my heartfelt thanks to Ms. Marion McClain for her English language editing.

I would like to thank my friends, and lab members at the Biostatistics and Bioinformatics Facility, Kentucky Biomedical Research Infrastructure Network (KBRIN), and Bioinformatics Journal Club, University of Louisville, USA, for their support. Further, I am extremely thankful to my colleagues and friends at my parent institute, the ICAR-Indian Agricultural Statistics Research Institute, India, for giving me moral support, and encouragement during this period. Special thanks to Dr. Rajender Parsad, Dr. Anil Rai, Dr. L. M. Bhar, Dr. A. K. Paul, Dr. Sudhir Srivastava, Dr. U. K. Pradhan, and Dr. D. C. Mishra for their encouragement and motivation during my Ph.D. studies.

Last, but not the least, I would like to thank my parents and all my teachers, since my primary schooling, who have given me invaluable education, and support during my entire life. I would also like to express my thanks to my wife, Rupali, for her understanding, and patience during these times. She encouraged me and made me stick with it. Also, many thanks to the members of my family in Jajpur, Odisha, India: Sasmita (Aunty), Sunachand (Bhai), Pratima (Bhauja), Jitendra (Kuna), Upendra (Runa) and Sahil for their patience, moral and loving support.

ABSTRACT

STATISTICAL APPROACHES OF GENE SET ANALYSIS WITH QUANTITATIVE TRAIT LOCI FOR HIGH-THROUGHPUT GENOMIC STUDIES

Samarendra Das

November 20, 2020

Recently, gene set analysis has become the first choice for gaining insights into the underlying complex biology of diseases through high-throughput genomic studies, such as Microarrays, bulk RNA-Sequencing, single cell RNA-Sequencing, *etc*. It also reduces the complexity of statistical analysis and enhances the explanatory power of the obtained results. Further, the statistical structure and steps common to these approaches have not yet been comprehensively discussed, which limits their utility. Hence, a comprehensive overview of the available gene set analysis approaches used for different high-throughput genomic studies is provided. The analysis of gene sets is usually carried out based on gene ontology terms, known biological pathways, *etc*., which may not establish any formal relation between genotype and trait specific phenotype. Further, in plant biology and breeding, gene set analysis with trait specific Quantitative Trait Loci data are considered to be a great source for biological knowledge discovery. Therefore, innovative statistical approaches are developed for analyzing, and

interpreting gene expression data from Microarrays, RNA-sequencing studies in the context of gene sets with trait specific Quantitative Trait Loci. The utility of the developed approaches is studied on multiple real gene expression datasets obtained from various Microarrays and RNA-sequencing studies.

The selection of gene sets through differential expression analysis is the primary step of gene set analysis, and which can be achieved through using gene selection methods. The existing methods for such analysis in high-throughput studies, such as Microarrays, RNA-sequencing studies, suffer from serious limitations. For instance, in Microarrays, most of the available methods are either based on relevancy or redundancy measures. Through these methods, the ranking of genes is done on single Microarray expression data, which leads to the selection of spuriously associated, and redundant gene sets. Therefore, newer, and innovative differential expression analytical methods have been developed for Microarrays, and single-cell RNA-sequencing studies for identification of gene sets to successfully carry out the gene set and other downstream analyses. Furthermore, several methods specifically designed for single-cell data have been developed in the literature for the differential expression analysis. To provide guidance on choosing an appropriate tool or developing a new one, it is necessary to review the performance of the existing methods. Hence, a comprehensive overview, classification, and comparative study of the available single-cell methods is hereby undertaken to study their unique features, underlying statistical models and their shortcomings on real applications. Moreover, to address one of the shortcomings (*i.e.*, higher dropout events due to lower cell capture rates), an

improved statistical method for downstream analysis of single-cell data has been developed. From the users' point of view, the different developed statistical methods are implemented in various software tools and made publicly available. These methods and tools will help the experimental biologists and genome researchers to analyze their experimental data more objectively and efficiently. Moreover, the limitations and shortcomings of the available methods are reported in this study, and these need to be addressed by statisticians and biologists collectively to develop efficient approaches. These new approaches will be able to analyze high-throughput genomic data more efficiently to better understand the biological systems and increase the specificity, sensitivity, utility, and relevance of high-throughput genomic studies.

TABLE OF CONTENTS

LIST OF TABLES

TABLES PAGE

LIST OF FIGURES

FIGURES PAGE

CHAPTER 1

INTRODUCTION

Gene Set Analysis for High-Throughput Genomic Studies

Recent advancement in genome sequencing technologies, such as Microarrays, bulk RNA-sequencing (RNA-seq), single cell RNA-sequencing (scRNA-seq), *etc.* leads to generation of tremendous volume of biological data [1]. Further, exploiting these data and drawing valid biological knowledge has posed a great challenge to researchers across the globe. For instance, in a genome wide expression study, the expression levels of several thousand(s) of genes for a tissue sample are measured in a single experiment and further used for identifying the group of genes which are relevant to the condition under study. The selected genes are expected to have major causal role for the phenotypic trait under study [2,3]. Earlier, biologists considered this Differential Expression (DE) analysis as the end of their analysis [4]. However, such analysis is the starting point of a complex process of drawing valid biological insights into high-throughput genomic data [5]. Earlier, the Gene Expression (GE) studies focused on univariate gene analysis, *i.e.*, testing the role of a single gene in the phenotypic trait under study (single gene testing) [6,7]. The scope of such studies is limited as the genes do not act individually; rather, genes work as an intricate network of a set of genes [8]. Therefore, to study such phenomena, a

set of secondary tools have been developed that place results from the expression studies in a broader biological context. One such approach is Gene Set Analysis (GSA) and one of its popular forms is called pathway analysis [9].

Traditional GSA methods used annotation information *like* pathways, Gene Ontology (GO), DE score, co-expression [8,10,11]. The enrichment analysis of gene sets based on such annotations does not establish any link between the selected gene sets and phenotypic trait, under which the data are being generated. Therefore, performing analysis of gene sets based on trait specific Quantitative Trait Loci (QTLs) through a computational approach instead of traditional GO or pathways information will be very helpful in unraveling genotype-phenotype relationships in plants and complex disease biology. Hence, the purpose of this project is to develop statistical approach(s)/framework(s) for analyzing gene sets based on genetically trait enriched QTL information. At the outset, this framework will consist of two major steps a) selection of gene sets (or DE analysis); and b) analysis of gene sets with QTL data. Here, we used expression data from Microarrays, RNA-seq and scRNA-seq studies. As the nature and underlying distributional properties of these datasets are significantly different, so, we have developed separate innovative statistical approaches for performing GSA with QTL for Microarrays and for RNA-seq/scRNA-seq studies. As, no web tools/R packages are available so far for performing analysis of gene sets with QTLs, we have developed the R packages for each of the considered high-throughput genomic studies, such as Microarrays, RNA-seq/scRNA-seq.

Innovative Aspects of the Project

This project is innovative in proposing the development of different statistical approaches for performing GSA with genetically rich trait specific loci data for various high-throughput genomic studies. It is well known that most of the traits (or diseases) are complex in nature because multiple genes (polygenes) contribute to the phenotype either individually or through interactions with each other or the environment. However, available univariate gene analysis approaches may not be helpful in drawing valid biological interpretations. Several statistical approaches, algorithms and tools have been developed to analyze gene sets instead of single genes. In the existing literature, gene sets are analyzed based on the annotation libraries like GO, KEGG pathways, DE score or MIPS functional categories. However, such approaches fail to tell the trait specific enrichment analysis of gene sets, which are essential for studying the biology of complex traits. Therefore, statistical approaches of GSA with QTL instead of traditional annotation categories will provide innovative ways to perform enrichment analysis of gene sets with highly popular QTL data. Further, this study will lead to identification of QTL candidate genes or QTL-enriched gene sets, helpful for plant biologists and genome researchers for framing further hypothesis to design crop breeding experiments or molecular designing of drugs. This project has provided an innovative and efficient platform for analyzing gene sets derived from wide range high-throughput studies, such as Microarrays, RNA-seq/scRNA-seq, with trait specific QTLs. Further, this study will provide valuable platforms for integrating various genomic datasets with QTL information.

Contributions and Layouts

Over the last decade, GSA approaches have been extensively used complex disease/plant biology to reduce the complexity of statistical analysis and enhance the explanatory power of the obtained results. Although a wide range of GSA approaches have been extensively reported in the literature, the statistical structures, and steps common to these approaches have not yet been comprehensively discussed, and this limits their utility. Therefore, Chapter 2 provides a comprehensive overview, statistical structure, steps, and generation wise evolution of GSA approaches used for Microarrays, RNA-seq and genome wide association data analysis. Further, the GSA approaches, and tools are classified based on the type of genomic study, null hypothesis, sampling model, and nature of the test statistic, *etc.*, along with their relative merits and limitations. Moreover, the Chapter 2 identifies the key biological, and statistical challenges in current gene set testing, which will be addressed by statisticians, and biologists collectively in order to develop the next generation of GSA approaches.

The preparation of ranked gene list is key part of the GSA, which involves the DE analysis of genes across the two conditions (case and control). For this purpose, several methods have been developed in the literature, which are either based on relevancy or redundancy measure(s). Through these methods the ranking of genes was done on a single high-dimensional expression data, which leads to the selection of spuriously associated and redundant genes. Hence, Chapter 3 provides a hybrid statistical approach for the selection of biologically relevant genes. Here, the genes are selected through statistical significance values

computed using a Non-Parametric (NP) test statistic under a bootstrap based subject sampling model. Further, the reported approach (Chapter 3) outperformed the competitive existing methods on multiple real datasets. After the preparation of the ranked gene list, gene sets are analysed with trait specific QTLs, which require further innovative statistical advancements. Such approaches may be considered as a great source for biological knowledge discovery in plant/disease biology and breeding. Hence, Chapter 4 proposes an innovative statistical approach called Gene Set Analysis with QTLs for interpreting gene expression data in the context of gene sets with traits. The reported approach (Chapter 4) was more innovative and effective in performing gene set analysis with underlying QTLs and identifying QTL candidate genes than the existing approach.

scRNA-seq is gradually replacing bulk RNA-seq and Microarrays for high-throughput studies of gene expression dynamics. The DE analysis or the ranking of genes is the major downstream analysis undertaken prior to GSA. The DE analysis in the presence of noise, from biological and technical sources, remains a key challenge in scRNA-seq. Several approaches have been reported in the literature to address this problem. Further, to provide guidance on choosing an appropriate tool or developing a new one, it is necessary to review, classify, evaluate, and compare the performance of DE analysis methods for scRNA-seq. Therefore, Chapter 5 provides a brief review of the existing practices in DE analysis of scRNA-seq data. Further, this Chapter also presents a detailed classification and comparative study of the available techniques. The shortcomings for each method, the best practices for DE analysis in single-cell

studies are well reported in Chapter 5. These findings are new. Hence, Chapter 5 provides a guideline for selecting the proper DE tool, best performing under particular experimental settings in the context of scRNA-seq. Among the reported challenges in scRNA-seq (Chapter 5), the presence of dropout events (excess zeros) due to low capture rates of cells severely biases the results, and this needs to be studied in detail. To address this problem, Chapter 6 presents an improved method for DE, and other downstream analysis that considers the molecular capture process of the cell in scRNA-seq data modeling. Further, the Chapter 6 also demonstrates that the reported method outperformed the existing methods on several public scRNA-seq datasets generated using different scRNA-seq protocols. The external spike-ins data can be used in the developed method to enhance its performance (Chapter 6).

After the DE analysis, gene sets selected from the ranked gene list need to be analyzed with the underlying trait specific QTLs for RNA-seq/scRNA-seq studies. This process requires newer and advanced statistical methods and tools. However, the GSAQ approach, reported in Chapter 3, only considers the selected gene set but ignores the DE scores of the genes present in the gene set. Therefore, Chapter 7 reports another innovative GSA approach for performing the analysis of gene sets with QTLs through considering the significant genes along with their respective DE scores for RNA-seq studies. To make all the developed approaches, reported in Chapters 3, 4, 6 and 7, user friendly, four different R software packages are developed and reported in Chapter 8. Finally, Chapter 9 provides the general discussion and conclusion of the findings reported in Chapters 2 – 8.

CHAPTER 2

FIFTEEN YEARS OF GENE SET ANALYSIS: A REVIEW OF STATISTICAL APPROACHES AND FUTURE CHALLENGES

Background

The term GSA refers to an analysis of set of genes and does not specifically mean modelling of the relations among genes in the gene set. Formally, the GSA is defined as a secondary statistical approach used to test the enrichment of the gene sets with any biological process or pre-existing bio-knowledge base or quantitative trait. In other words, genes are aggregated into gene sets based on shared biological or functional properties or any pre-existing bio-knowledge base [5]. These bio-knowledge bases include databases of molecular knowledge, *i.e*., molecular interactions, regulation, molecular product(s), and even phenotype associations. In other words, GE and Single Nucleotide Polymorphism (SNP) datasets are used as input for GSA (in the presence of a annotation database) to provide valid biological insights into various complex diseases (Figure 2.1) [9,12]. In fact, GSA can be used for all the genomic studies, where the output is a long list of genes or transcripts. For instance, that long list of genes can even come from any upstream analysis including signatures of co-expressed genes from weighted gene co-expression network analysis [3].

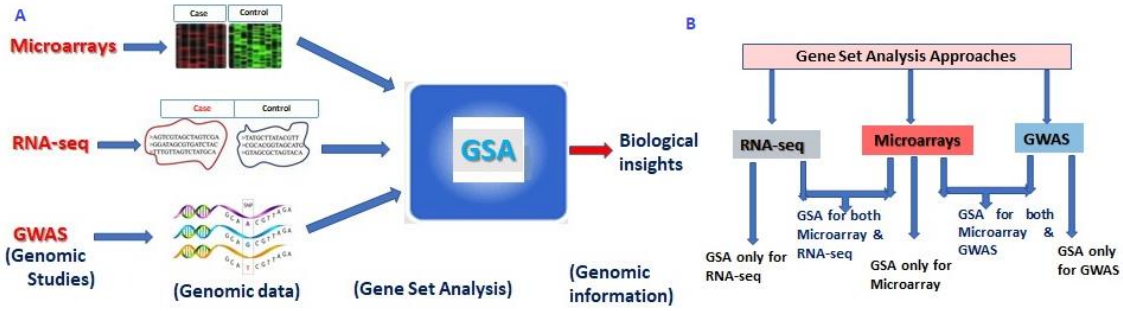


Figure 2.1. Outlines and classification of gene set analysis approaches. **A**: Outlines of gene set analysis approaches; **B**: Classification of gene set analysis approaches for high-throughput sequencing studies.

Units of Gene Set Analysis

The functional unit of GSA is the gene set, which can be defined as any group of genes that share a particular property, *i.e.*, involvement in a common biological process or any pre-existing bio-knowledge base [12,13]. Through GSA, a gene set that shares a common property is tested for its association with the trait or phenotype under study [8]. For this purpose, a wide range of GSA approaches and tools are available for high-throughput sequencing studies. These tools have differences in underlying statistical principles and practices, but there are similarities among the available tools in terms of statistical structure. For instance, GSA for GE studies has a two-tier structure [13,14]: a) computation of gene level statistic(s); and b) bi-variate statistical testing to compute the test statistic or *p-value* for the gene set. However, GSA for Genome Wide Association Study (GWAS) has a three-tier structure: a) computation of SNP level statistics; b) associating SNPs (linkage between SNPs) to genes and computing gene-level statistics from SNP statistics; and c) computation of enrichment statistic or *p-value or* False Discovery Rate (FDR) for the gene set.

Hypotheses in Gene Set Analysis

The available statistical approaches for GSA vary greatly with respect to underlying statistical tests, and hence depend on the formulation of the null hypothesis [5,9,15]. These null hypotheses can be grouped as self-contained and competitive [16]. In the usual set up of GE studies (or GWAS), genes (or SNPs) that are significantly associated with a trait/phenotype are identified and then evaluated, whether the significantly associated genes (or SNPs) tend to cluster in predefined gene sets or not. For instance, the self-contained null hypothesis can be framed as, H_0: genes/SNPs in predefined gene sets are not associated with the underlying trait (phenotype) against the alternate H_1: genes/SNPs in predefined gene sets are associated with the trait (phenotype). The statistical approaches with a self-contained null hypothesis are called as self-contained approaches of GSA and they only consider the genes (SNPs) in the predefined gene sets. Statistical tests of GSA with a competitive null hypothesis are known as competitive GSA approaches, and the underlying null hypothesis can be expressed as, H_0: genes/SNPs in predefined gene sets are associated with the underlying trait (phenotype) as much as are genes/SNPs outside the predefined gene set, against H_1: genes/SNPs in predefined gene sets are more associated with the trait (phenotype) than genes outside predefined gene set. Here, the competitive GSA approaches consider genes (SNPs) from both the predefined gene set and the outside gene set [5,17]. The self-contained null hypothesis is invariably more restrictive than the competitive null hypothesis.

Sampling Models in Gene Set Analysis

The enrichment significance of a gene set is assessed through *p-value or* adjusted *p-value or* FDR after multiple testing correction (*i.e.*, lower values indicate more enrichment and *vice-versa*) computed from a statistical test. Further, these statistical tests are commonly based on experimental designs having subjects/genes as units. On such statistical designs, different sampling procedures are rigorously used to obtain the distribution of the test statistic(s). Here, two types of sampling models are used in GSA: *i*) subject sampling model; and *ii*) gene sampling model.

Subject Sampling Model

Classical statistical tests are based on an experimental design having microarray/RNA-seq samples as subjects, where each subject has the same set of (GE) measurements [5,8,17]. In the usual supervised setting, the sampling model consists of M independent realizations (for M subjects) of (X_1, y_1), (X_2, y_2), …, (X_s, y_s), …, (X_M, y_M), where, X_s represents the N-dimensional vector (N: total number of genes) of the GE levels for *s-th* subject and y_s is the corresponding class label (e.g., case: +1 vs. control: −1), s=1, 2, …, M. Therefore, M expression levels of different subjects are assumed to be independently and identically distributed (iid), but expression levels of genes within the same subject may be correlated for a given condition. Usually, resampling procedures like bootstrap and permutation procedures are used on such models for gene [3,18] as well as gene set testing [5,19]. The statistical combination of subject sampling model and a self-contained

null hypothesis provides a reliable platform for valid computation of *p-values* with easy interpretation and close relation(s) with single gene (or SNP) testing [20].

Gene Sampling Model

In GSA, 2×2 tables are extensively used to statistically fit a Hypergeometric distribution [5,21]. The underlying model of a 2×2 table is a gene sampling model. Further, each cell of such a table is filled with a sample of genes, each of which is drawn at random from the gene space (*i.e.*, set of genes in the data). Here, in this sampling model, each sampling unit (*i.e*., gene) can be subjected to two fixed set of indicator measurements, *i.e.*, (*A*, *B*), where, (*i*) *A* (1 or 0) indicates whether the gene is a part of the predefined gene set or not and (*ii*) *B* (1 or 0) indicates whether that gene is in the list of DE genes or not [5,17]. Further, the gene space can be formalized into a population having *N* units (for *N* genes) and shown as: (A_1, B_1), (A_2, B_2), …, (A_i, B_i), …, (A_N, B_N). The competitive null hypothesis is popular and easy to formulate in a gene-sampling model setup [9]. Here, the gene sampling model may be considered as a mirror image of classical subject sampling model [18]. The gene sampling model considers the sampling units as iid, which assumes that genes are independent. Such assumptions are highly unrealistic, and the *p-values* computed using such models are statistically invalid for further interpretations. Hence, gene sampling models are quite complex and delicate as compared to a subject sampling model.

GSA Approaches for High-Throughput Genomic Studies

The GSA approaches can be grouped based on different high-throughput genomic studies, as the underlying nature and distributions of the datasets are different. A

classification of GSA approaches with respect to their application to genomic studies is shown in Figure 2.1. Initially, the GSA approaches were developed for Microarrays (*i.e.*, Microarrays GSA) and subsequently extended to RNA-seq and GWAS data analysis (Figure 2.1). For instance, gene set enrichment analysis (GSEA) was originally developed for Microarrays, and subsequent extensions of GSEA, *i.e.*, SeqGSEA and GSEA-SNP were introduced to analyze RNA-seq and SNP datasets, respectively.

Microarrays GSA

Huge amounts of GE data from Microarrays are available in public domain databases, which need to be analyzed for drawing valid biological insights into such datasets. Therefore, several GSA methodologies have been developed for this purpose. The classification of Microarrays GSA is shown in Figure 2.2, which illustrates the evolution of GSA approaches over time in terms of the requirements of annotation information, sampling model, and various null hypotheses under statistical tests. Moreover, the work on GSA started with the immediate need for functional analysis of Microarray data based on GO that gave rise to over representation analysis (ORA), which evaluates the statistical significance of gene sets in a particular pathway/functional category [22]. It is also referred to as a 2×2 table method [5], due to the fact that ORA approaches are mostly based on 2×2 tables and gene sampling models. The most commonly used statistical tests in ORA approaches/tools are hypergeometric, chi-square or binomial tests [23–25] . However, despite the extreme popularity and ease of execution, the ORA approaches also suffer from limitations, as listed in Table 2.1. The ORA form of analysis of gene sets can also be labelled as first generation of Microarrays GSA.

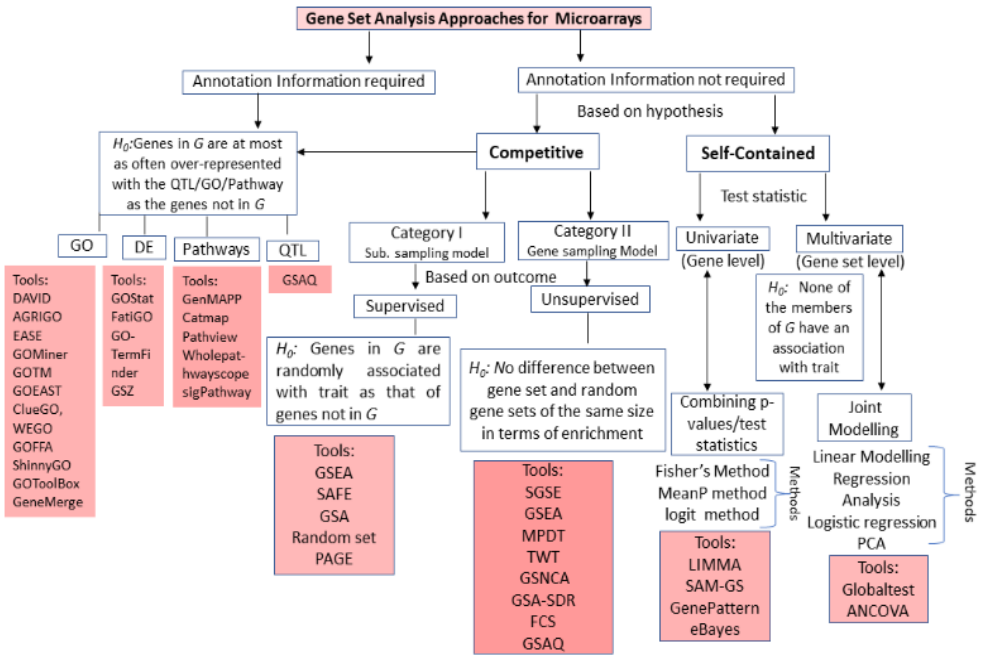


Figure 2.2 Classification of gene set analysis approaches and tools available for Microarrays. Schematic representation of the breakup of GSA methods available for microarrays data analysis based on statistical tests (i.e., null hypothesis, test statistic(s)) and requirement of annotation databases. *G*: Gene set. (LIMMA, SAM-GS, eBayes, etc. generate inputs to GSA, not GSA themselves)

In most of the cases, the gene annotation information is either incomplete or totally unavailable; therefore, another class of GSA approaches was developed. These approaches include the Enrichment Score (ES) form of GSA [11], starting with the landmark work on enrichment analysis of gene sets (i.e., GSEA) [8,26]. Subsequently, several other statistical approaches, and tools were developed for assessing the significance of gene sets in interpreting the high-throughput data. The ES based GSA approaches greatly vary among themselves with respect to underlying statistical tests and sampling models. The major steps for such approaches include initial computation of the gene-level statistic(s) using GE data under two contrasting conditions. For instance, correlation of expression measurements with phenotypes/traits [27], ANOVA [28], Q-statistic [16], signal-to-

noise ratio [8], t-statistic [6], Fold Change (FC) [29], Z-score [30], etc., are implemented in contemporary ES based tools. There is a wider choice for gene-level statistic(s), ranging from parametric to NP, for GSA. However, the selection of a gene-level statistic has a negligible effect on identification of significantly enriched gene sets [21]. When there are few biological replicates available, a regularized statistic may be preferred [21]. The second step is the aggregation of gene-level statistic(s) for all genes in a gene set into a single gene-set level statistic (Figure 2.3). This includes the computation of gene-set level statistic using multivariate or univariate techniques (Figure 2.2). The former accounts for interdependencies among genes, while the latter disregards the same among genes distributed across the gene set. The currently available ES based GSA approaches/tools include Kolmogorov-Smirnov (KS) statistic, weighted KS statistic [8,11], sum, mean, or median of gene-level statistic [31], Wilcoxon rank sum [32], Max-mean statistic [26], *etc*., under the univariate category. Moreover, the multivariate category includes global test, ANCOVA, *etc*., for computing the gene-set level statistic [16]. Interestingly, multivariate statistic(s) are expected to have higher statistical power, but univariate statistic(s) actually show more power at a higher level of significance (*e.g.,* 0.1%) in real biological data, and equal power as the former at lower level of significance (*e.g.*, 5%) [33]. The third step is computation of statistical significance (*p-value*) or adjusted *p-value* or FDR to assess the enrichment of gene sets (for gene-set level statistic). This step requires the formulation, as well as testing of the null hypothesis against alternate one. Based on the null hypothesis, the ES-based GSA approaches can be broadly

divided into: *i*) competitive approaches, and *ii*) self-contained approaches (Figure 2.2). Moreover, the competitive approaches can be further subdivided into two categories based on the available outcome information of class: *i*) supervised approaches and *ii*) unsupervised approaches (Figure 2.2).

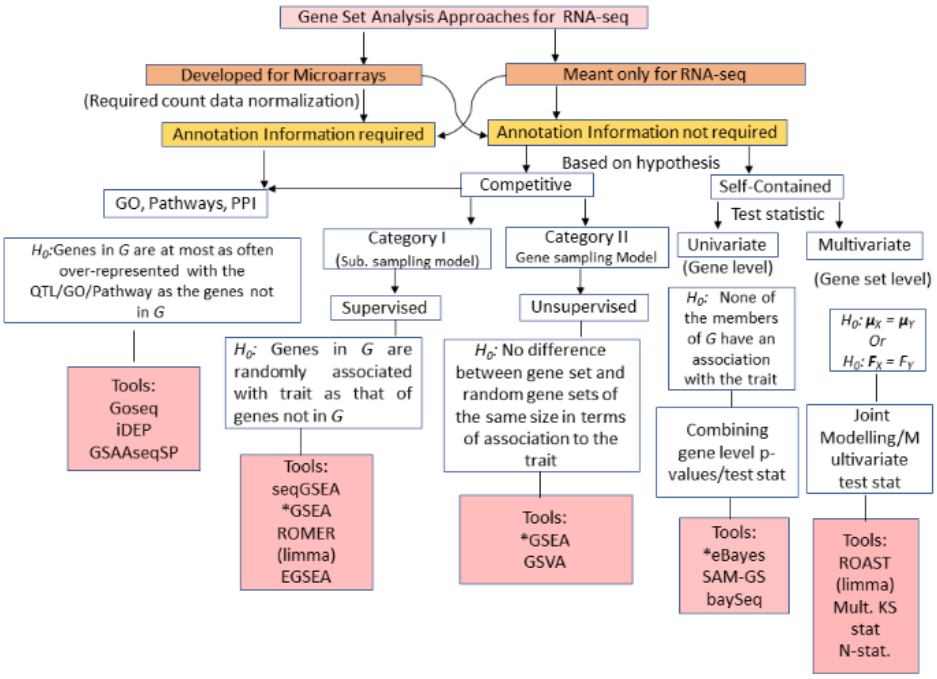


Figure 2.3. Classification of gene set analysis approaches and tools available for RNA-seq data analysis. Schematic representation of the breakup of GSA methods available for RNA-seq data analysis based on statistical tests and requirement of annotation databases. *G*: Gene set. *Tools require normalization of data prior to application.

Mostly, the supervised competitive approaches use the subject sampling model to randomly sample the class labels of each sample and compare the genes in the gene set with those of its complement. Here, it may be noted that the supervised term is used as the class labels are known and these approaches use these class labels for sampling purposes. However, unsupervised competitive approaches used the gene sampling model to compute the *p-value* through comparing genes in gene set with the genes outside gene set. But self-contained ES-based GSA approaches use the permutation procedure to compute the *p-*

values by permuting the class labels for each sample and comparing the genes in the gene set with itself, while ignoring the genes outside gene set. Here, it is evident that competitive ES-based GSA approaches have more statistical power as compared to self-contained approaches [26]. This may be due to the fact that competitive approaches require information on both genes in the gene set as well as genes not in the gene set [5]. The ES form of GSA may constitute the second generation of Microarrays GSA (Table 2.1).

Table 2.1. Generation-wise evolution of GSA approaches for microarray studies.

Approach	Methods	Advantages	Limitations	Tools/algorithms
ORA **(First generation microarray GSA)**	Hypergeometric distribution/Fisher's test Binomial distribution, Chi-square distribution, etc.	• Easiness in execution. • Assigns easily interpretable measure like p-values to the whole gene set.	• Highly dependent on threshold/cutoff value, which is at user's discretion and hard to determine. • Test statistic independent of genes differential expression score. • Uses only most significant genes based on hard threshold and discards others, lead to information loss. • Assumes each gene contribute equally to phenotype/trait. • Assumes each gene as independent and ignores the correlation or redundancy among genes in gene set. • Assumes that each predefined gene set is independent of others, which is erroneous.	DAVID [34], AgriGO [24], Onto-Express [22], GenMAPP [35], GoMiner [36], FatiGO [37], GOstat [25], FuncAssociate [38], GOToolBox [39], GeneMerge [40], GOEAST [41], ClueGO [42], FunSpec [43], GARBAN [44], GO:TermFinder [45], WebGestalt [46], GOFFA [47], WEGO [48], GOTM [49], EASE, GSAQ [17], Pathview [50], Wholepathwayscope [51], ShinnyGO
Enrichment Statistic Analysis **(Second generation microarray GSA)**	Wilcoxon signed rank test, Sum, Mean, or Median of gene-level statistic(s), Wilcoxon signed rank sum, Max-Mean Statistic	• Do not require a threshold/ cutoff value for dividing gene space into selected and non-selected part. • Considers dependence among genes in gene set. • Test statistic is based on the differential GE score of genes in gene set.	• Analyzes each gene set independently. • Considers only the number of genes in a gene set (pathway) for performing GSA but ignores the additional information available from the bio-knowledge bases. • Assumes the predefined gene sets mutually exclusive, but in biology, these gene sets are overlapping. • Most ESA methods use differential GE to rank genes/compute test statistic but discard this information from further analysis.	GSEA [8], SAFE [32], GSA [26], Random set [52], sigPathway, Category, GlobalTest [16], PCOT2 [53], SAM-GS [54], LIMMA [55], Catmap [56], T-profiler [57], FunCluster [58], GeneTrail [59], Gazer [60], GSAQ [17], ANCOVA test, CAMERA [61], PAGE [30], GAGE [62], SGSE [63], GSNCA [64], GSA-SDR [65], GenePattern [66], plantGSEA [67], GSAR [20]
Topology Analysis **(Third generation microarray GSA)**	**Graph/network theory**	• Considers both genes relation /dependency with other genes as well as experimental condition changes. • Considers the topology of the pathways/gene sets in modeling.	• Dependent on the type of cell due to cell-specific GE profiles and condition being studied, which is rarely available. • Not so popular as require more rarely available information and computationally intensive. • Unable to consider interactions between gene sets (pathways). • Heavily dependent on annotations.	PathwayExpress [68], ScorePAGE [69], SPIA [70], NetGSA [71], TopoGSA [72], CliPPER [73]

RNA-seq GSA

Recently, transcriptome deep sequencing has surpassed Microarrays by providing better quantification of GE for high expressed genes (in terms of read counts), and higher levels of accuracy and reproducibility [15,74,75]. Hence, it is highly pertinent to adapt the existing Microarrays GSA to RNA-seq data with the help of data transformation along with new approaches being developed (Figure 2.1B). The first approach of GSA for RNA-seq data (RNA-seq GSA), GOseq, was suggested by Young *et al*. a decade ago [76]. It performs over-representation of GO categories enriched with a long list of highly expressed genes in RNA-Seq data. Further, an easy-to-use web application, integrated differential expression and pathway (iDEP) analysis was developed for in-depth analysis of RNA-seq data [77]. Detailed descriptions of the available RNA-seq GSA approaches, background methodologies, execution tools, and their features are listed in Table 2.2. Moreover, the ORA-based RNA-seq GSA may be considered as the first generation of RNA-seq GSA.

To tackle the limitations of ORA approaches (Table 2.2), ES-based RNA-seq GSA approaches have been developed, and these constitute the second generation of RNA-seq GSA. Here, the read counts are given as input for computation of different test statistic(s) for GSA, which depend on the nature and distribution of the data. For instance, Microarrays GSA (*i.e*., ES-based GSA) deal with continuous data expected to follow a Gaussian distribution [74]. However, RNA-seq involves measurements that are non-negative counts ranging from zero to millions and are expected to follow Negative Binomial Distribution (NBD) [15,75].

Therefore, Microarrays GSA approaches may not be directly applicable to RNA-Seq data. Hence, some authors suggested normalization of the count data prior to the use of Microarrays GSA [15]. For instance, VOOM-normalization is used for normalizing the read counts for sequence-depths, then Microarrays GSA are applied on the normalized RNA-seq data [78]. The Goeman and Buhlmann formulation can be applied to classify the ES-based RNA-seq GSA approaches into either competitive or self-contained [5], based on the underlying null hypotheses (Figure 2.3). Further, a competitive GSA approach, *i.e*., gene set variation analysis (GSVA), was developed and demonstrated highly correlated results between Microarrays and RNA-Seq sets for samples of lympho-blastoids cell lines [79]. This high correlation may be due to the fact that GSVA as a NP approach does not depend on the distributional nature of data obtained from the studies. Fridley *et al*. proposed a GSA approach, Gamma method, with a soft truncation threshold to determine the significant gene set, while a generalized linear model is used to assess significance [80]. Subsequently, GSEA, the first ever competitive approach of RNA-seq GSA, was used for RNA-seq data analysis after normalization of the count data [80]. Thereafter, several modifications were made in GSEA by integrating both DE and differential splicing (DS) information in the analyses to develop SeqGSEA and has better performance over GSEA [19].

Table 2.2. Generation-wise evolution of GSA approaches for RNA-seq studies.

Approach	Methodology	Advantages	Limitations	Tools
ORA (First generation RNA-seq GSA)	Hypergeometric distribution, Fisher's exact test	• Simple to use. • Assigns easily interpretable measure like *p-value* to the whole gene set.	• Use hard threshold approach to select gene sets. • Assumes each transcript as independent and ignores the correlation or gene-gene interaction.	GOseq [76], iDEP [77]

		• Less time consuming to interpret huge RNA-seq data.	• Mostly dependent on annotation bases, but RNA-seq transcripts are not well annotated.	
GS Enrichment Analysis **(Second generation of RNA-seq GSA)**	Wilcoxon signed rank test, Max-Mean Statistic (with count normalization technique)	• Do not require a threshold for dividing gene space into selected and non-selected part. • Considers dependence among genes in gene set.	• Use normalization technique to get microarray like data, hence, loss of the count nature of RNA-seq data • Through data transformation, dispersion and other inherent nature of RNA-seq data are lost • ES based tools/algorithms use differential score to prepare ranked transcript list but ignore this information for gene set testing. • GSEA based tools like seqGSEA are computationally intensive, time consuming and and only offers the single gene set-level statistic. • GSVA is not designed for gene set-based differential expression analysis between two phenotypically distinct sample groups. • ES based GSA approaches do not consider the inherent zero inflation in the RNA-seq data.	AbsFilterGSEA [81], GSAAseqSP [82], seqGSEA [83], ssGSEA, EGSEA [84], GSVA [79], GSEPD [85], RNA-Enrich [86]

The self-contained GSA approaches can be divided into a) univariate or gene-level; and b) multivariate or gene set-level based on the distributional nature of the test statistic (Figure 2.3). The gene-level GSA approaches test a null hypothesis that the gene-set associated score does not differ between phenotypes/traits. Further, the univariate approaches are executed in two steps: *i*) computation of gene level statistic(s) from the count data; and *ii*) combining gene-level statistics to compute gene set level statistic or *p-value* or adjusted *p-value*. For the former case, the gene-level test statistic(s) of Microarrays GSA were used in a recent study for RNA-seq GSA [80], which is quite straight forward and easy to implement. For the latter step, the gene-level statistic(s) can be combined into a single gene set statistic/*p-value* through Fisher's method, Stoufer's method, Meanp, logit method, etc. [17]. Moreover, the self-contained multivariate GSA approaches jointly model the genes to compute the gene set-level statistic(s)

(Figure 2.3). These tests include multivariate generalization of the KS statistic [8,11], N-statistic [74], ROAST [78], etc. Further, the application of these tests requires the normalization of the RNA-seq data over varying sequencing depths [78]. Moreover, statistical significance is computed by comparing the observed statistics of gene sets with its null distribution, obtained by permuting the sample labels. Then, the enrichment significance of the gene set is assessed through the computed *p-value or* adjusted *p-value* or FDR after multiple testing correction.

GWAS GSA

GWAS has been successfully applied to identify many novel loci for complex traits, which are quantitative (polygenic) in nature. Therefore, to understand the underlying genetic architecture, GSA approaches that place GWAS results in a broader biological context have been used [87]. Initially, GSA methods for GWAS (i.e., GWAS GSA) were borrowed from Microarrays [8,11] and subsequent new approaches were developed exclusively for GWAS (Figure 2.1). The classification of GWAS GSA approaches is shown in Figure 2.4. The first step for classification of GWAS GSA approaches can be their source of origin, including: *i*) GSA Microarrays adapted to GWAS; and *ii*) those developed exclusively for GWAS. Further, based on the requirement of annotation libraries, the GWAS GSA approaches can also be classified as: *a*) GSA requiring pre-defined gene sets; or *b*) GSA not requiring pre-defined gene sets. These approaches are based on the principle of over-representation of genes in those predefined gene sets obtained from different bio-knowledge bases. Moreover, such ORA approaches constitute the first generation of GWAS GSA.

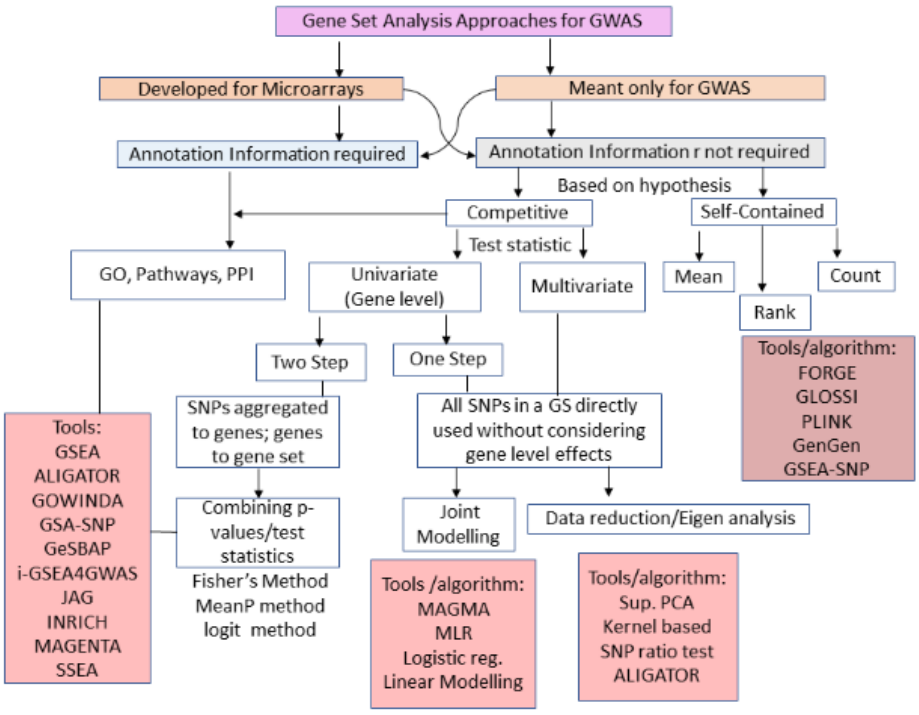


Figure 2.4. Classification of gene set analysis approaches and tools available for SNP data analysis. Schematic representation of the breakup of GSA methods available for SNP data analysis based on statistical tests and requirement of annotation databases.

Due to the limitations of ORA-based GWAS GSA approaches, ES-based GWAS GSA approaches came into use, which we may call the second generation of GSA in GWAS. Further, the second generation of GWAS GSA starts with the enrichment analysis of gene sets for SNP data, *i.e.*, GSEA-SNP [14,88] using weighted KS statistics [89]. Later approaches, based on other tests, *viz.* weighted-sum test [90], simple-sum test [91], collapsing test in combined multivariate and collapsing method [92] and sequence kernel association test [93], are used for computation of the gene-set enrichment score. Moreover, varieties of ES-based methods with similar ideas have been developed, such as the gene set based testing of polymorphism [94], GSA-SNP [88], SNP-ratio test [95], *etc.*

A class of GWAS GSA approaches have been developed by considering the topology of the gene sets/pathways, and this constitutes the third generation of GWAS GSA. This includes methods to parse the internal information of the pathway (e.g., signaling pathway impact analysis (SPIA) [70] and CliPPER [73]). Further, the second and third generation GWAS GSA methods focus on statistical results such as *p-values* or ES, as input rather than the original data. Thus, the fourth generation of GWAS GSA approaches have been developed by providing original data as input. Further, the underlying principle of these approaches is testing of the multivariate distribution of the multi-loci data or extracting the principal components from the original data. This includes linear combination test [96], supervised principal component analysis (SPCA) [96], Smoothed functional PCA [97], etc. Other model-based methods include LRpath [98], a logistic regression-based method, and MAGMA [99], linear model based method. Recently, the Generalized Berk-Jones statistic, a permutation-free parametric framework, was used for GSA [99], and this incorporates information from multiple signals in the same gene. The descriptions of the available GWAS GSA approaches, tools, their background methodologies pertaining to various generations are listed in Table 2.3.

Table 2.3. Generation-wise evolution of GWAS GSA approaches.

Approach	Method	Advantages	Limitations	Tools/Algorithm
ORA **(First generation GWAS GSA)**	Hypergeometric distribution, Fisher's exact test, Binomial test	• Simple to use and easy to interpret • Assigns statistically convincing measure like p-value for SNP set, which is biologically meaningful • Computationally not so expensive	• Hard threshold (arbitrary) divides the SNP list into selected and not selected SNP set. For instance, if threshold value for p-value is 0.05, means SNP with value 0.051 is not included in SNP list • Uses only most significant SNP and discards others, lead to information loss	SNPtoGO [100], ALIGATOR [101], ATRP [102], MetaCore [103], PARIS [104], SET SCREEN test [105], SNP ratio test [95], GLOSSI, GeSBAP [94], INRICH [106],

			• Test statistic is independent of SNP data (based on only SNP count), and ignores the strength of association • Considers each SNP independent and ignores the linkage disequilibrium • Assumes each SNP contribute equally, which is not true as there are common and rare variants • Dependent on pre-defined bio-knowledge base, which is mostly incomplete or unavailable	GeneSetDB [107], MAGENTA [108], KGG-HYST [109], PLINK [110], JAG [111], FORGE [112]
Enrichment Statistic(s) Analysis **(Second generation GWAS GSA)**	Wilcoxon signed rank test, Sum test, Weighted Sum test (Enrichment score like statistic)	• Do not require hard threshold for dividing SNP list into selected and non-selected part • Jointly consider multiple contributing factors in the same gene set, might complement the most-significant SNPs/genes approach • Test statistic is computed from the data considering linkage disequilibrium	• Analyzes each gene set independently. • Only considers data for selecting SNPs and after ignores the data from gene-set testing. • Treat all genes in a gene set independently and do not account for the relationships between genes.	GSA-SNP [88], GSA-SNP2, GSEA-SNP [113], GSEA-P [10] GenGen [114], ICSNPathway [115], i-GSEA4GWAS[116], i-GSEA4GWAS2 [117]
Topology Analysis **(Third generation GWAS GSA)**	Graph/Network theory	• Relationships between genes are used to assign different levels of "importance" to genes in the set • Helps in integrate gene set membership information with interaction data from a separate source	• Difficult to generalize • True topology is dependent on the type of cell and experimental condition, which are rarely available • Cannot model the dynamicity of the cellular system • Heavily dependent on annotations, which is either missing or incomplete	dmGWAS[118], Ingenuity Pathway Analysis (IPA)[119], PINBPA[120], PathVisio[121], Cytoscape[122]
Multivariate/ Model/ Regression Analysis **(Fourth generation GWAS GSA)**	Linear regression Model, Ridge regression, Logistic regression, Linear models	• Consider both SNP and gene set information simultaneously in same model • Jointly consider linkage disequilibrium and gene-gene interaction in gene set for modeling • Future behavior of the system can be predicted • Dynamicity of the biological system can also be modeled and studied	• Computationally intensive • High dimensionality of genomic data raises serious concerns • Ignores the non-linear interactions among biomolecules	LRpath [98], SPCA[96], SFPCA[97], MAGMA[123], GRASS, Generalized Berk-Jones statistic[99],

The formulations based on underlying statistical tests [5] can also be used for classifying GSA GWAS, *i.e.*, self-contained and competitive approaches (Figure 2.4). Self-contained GWAS GSA considers only the SNPs in the gene set and tests the null hypothesis that none of those SNPs are associated with the phenotype.

Competitive GSA considers all SNPs in the data and tests the null hypothesis that the genes in the gene set are no more strongly associated with the phenotype than other genes [124]. Further, the competitive GWAS GSA approaches can be divided into: *i*) two-step approach(s), in which SNPs (in each gene) are first used to evaluate association with the gene, then gene-level statistic(s) are aggregated to gene-set level enrichment value to test its association with the phenotype; and *ii*) a one-step approach, in which all SNPs in a gene set are simultaneously considered in the analysis without consideration of gene-level effects (e.g., MAGMA) (Figure 2.4). For the former categories, the univariate statistical approaches are used, while multivariate techniques such as joint modelling are used for latter. Moreover, the self-contained GWAS GSA approaches can also be grouped based on the type of gene-set test statistic used for testing (Figure 2.4). This can be broadly subdivided into three classes: *i*) mean-based, (i.e*.,* mean or sum of the gene-association scores); *ii*) count-based, (i.e., classifying genes as 'significant' or 'not significant' by applying a threshold to the gene-association scores and using the number of 'significant' genes in the gene set as a test statistic); and; *iii*) rank-based, first ranking the genes according to their gene-association score and computing overrepresentation of the gene-set genes at the top of that ranking.

Limitations and Future Challenges of GSA

Here, we report the existing limitations as well as the key challenges observed in the available GSA approaches that should be kept in mind while using them. These

existing limitations and challenges can be divided into two broad categories: *i*) biological annotation challenges and *ii*) methodological challenges.

Biological Annotation Challenges

The classification of GSA approaches for high-throughput genomic studies (Figures 2.2–2.4) shows that GSA approaches require annotation information for analyzing gene sets. It is expected that the next generation GSA will require improvement of the existing annotations as well as new high-throughput annotation information [21,53]. Therefore, it is important to create accurate, high resolution bio-knowledge bases with specific emphasis on cell dynamics and condition, along with tissue information to annotate genes studied in an experiment. These knowledge bases will allow us to model the inherent organism's response to any extraneous condition as a dynamic system and will help in predicting the system's behavior at different times as well as in relation to various factors (e.g., mutation, disease, environmental conditions, etc.).

Limited annotation information: The contemporary GSA approaches mostly use GO and pathways information for analyzing gene sets [24,25,34,36,37,67,76,100,101], but there is enough other annotation information available or will soon be available in public domain databases that can be effectively used for GSA to gain biological insights into the etiology of complex diseases in humans as well as other organisms. For instance, Das et al. used the QTL data as annotation information to develop a GSA approach to analyze the gene sets obtained from Microarrays [17]. This approach has immense use for performing trait/QTL enrichment analysis of gene sets and further, QTL enriched

gene sets can be used for molecular breeding programs for biotic/abiotic stress engineering in plants. Moreover, this annotation information can also be used in the future for developing new generation GSA approaches for analysis of RNA-seq and GWAS data. Such advances in GSA will open new avenues to understand the molecular complexity behind complex diseases in humans and other organisms including crop plants.

Low resolution knowledge bases: Recent advancements in genomics and proteomics lead to a paradigm shift in data generation, with unprecedented high resolution. At the same time, there is a demand for high resolution annotation bio-knowledge bases to perform GSA. For instance, during the early period of GE genomics, Microarrays were the key experiment to obtain a global view of GE in the human genome. To perform GSA, GO [125] and KEGG [126] annotation bases were developed in parallel and implemented in several web tools. Further, such databases specify which genes (in terms of probe id/Entrez id) are active in each GO category/pathway/ predefined gene sets. However, microarray technology has been replaced with RNA-seq and single cell RNA-seq (scRNA-seq) technologies. Hence, the current annotation databases need to be updated with respect to these high-resolution techniques. It is essential that they also begin specifying other information, such as transcripts (or scRNA-seq transcript) and SNPs that are active in each predefined pathway, GO category.

Missing or incomplete annotation: Although enormous annotation bases are available in the public domain, some annotations are either missing or incomplete for certain genes. For instance, the current release of GO contained entries for

19,649 human genes annotated with at least one GO term. Many of these genes are hypothetical, predicted or pseudogenes. For example, the number of protein-coding genes in the human genome is estimated to be 20,000–25,000, which shows that annotation information of hundred(s) of genes is still missing, and this may have a crucial role in various diseases. In addition to the missing annotations, most of the current databases have lower resolution (*i.e*., lesser information on transcript and SNP) [21,127], which leads to biased results from GSA. Further, current knowledge bases are built by curating experiments performed in different cell types at different time points under different conditions/locations. However, these details are typically not available in these knowledge bases. Thus, these databases need to be updated for future dynamic or cell specific GSA.

Methodological Challenges

Lack of benchmark/gold standard: In simulation, it is expected that multivariate approaches outperform the univariate counterparts, as the former considers inter-variable correlations. However, in biology, it is observed that univariate statistic(s) are equal to or better than multivariate statistic(s) [33]. This observation raises several questions about the performance assessment of GSA approaches using simulated datasets as a benchmark. It is likely that biology is more complicated than simulated scenarios and is influenced by factors such as the absence of exclusive division into classes, presence of outliers, experimental or technical hidden factors, environmental influence(s), random errors, etc. Therefore, one way to handle such a situation is to use benchmark/gold standard datasets with a valid biological basis. For instance, Ballard *et al*. (2010) compared two GSA methods

based on their applications to three Crohn's disease benchmark GWAS datasets with well-known biological basis [9,13,114]. Further, a combination of both benchmark biological datasets with statistically strong criteria can provide a suitable platform for comparative performance analysis of GSA approaches.

Criteria for comparing GSA approaches: When the performance of a GSA approach is assessed, it is expected to have certain proportions of false positives from the test. The ES-based GSA approaches compare the observed ES statistic with its null distribution as generated by random sampling/permuting the sample labels/disease outcomes or permuting genes/genotypes information [12,99]. Usually, through permutation, *p-values* are computed for assessing the enrichment significance of gene sets [5,16]. Then, -log10(*p-value*) and power of the statistical tests are used to assess the performance of GSA approaches [17]. However, alternate measures may also be used for comparative performance analysis of GSA approaches. In one such measure, the above computed *p-values* may be used to plot the histogram for the null gene sets, and that is expected to follow a uniform distribution. This phenomenon may be used to compute type-I error rates for GSA approaches, which can then be used as an efficient criterion for performance analysis of GSA approaches along with statistical power and FDR. In other words, GSA approaches with lower type-I error rates will be considered as better and *vice-versa*. These criteria can be computed on benchmark/gold standard datasets, which will provide a suitable platform to compare GSA approaches.

Improvement in terms of statistical power: In ORA-based GSA approaches, the test statistic(s) are computed by treating each gene equally. But in biology, some genes contribute more toward the disease/trait development. Treating all genes as equal in computing the test statistic reduces the statistical power of the GSA approach. Hence, one powerful strategy may be to consider the DE scores of genes [8,11,128,129] or ranks of the genes in a gene list while constructing the test statistic(s). This mechanism will attribute more statistical power to GSA approaches as compared to the existing ones. This approach needs to be well studied on benchmark data in the future to assess its rigor and reproducibility. Further, other *a priori* biological information, *viz.* eQTL, network topology, co-expression scores, *etc*., can be used as auxiliary information in GSA approaches to improve their performance.

Selection of null hypotheses: The competitive GSA approaches use a gene sampling model to compute the *p-values* for gene sets [5,16]. In the gene sampling model, it is assumed that genes are iid, which is highly unrealistic from a biological standpoint. Hence, the test statistic computed based on such assumptions from the gene sampling model leads to biased and misleading results. Therefore, methods, such as GSEA [8,11] and SAFE [32] use a hybrid concept, *i.e*., compute their test statistic(s) based on a gene-sampling model but calculate their *p-values* using the subject sampling model. The discrepancy between these two models makes the statistical properties of the test unclear and its interpretation very difficult. These problems are unavoidable, as the definition of the competitive null hypothesis is intimately tied to the gene-sampling model, whereas valid *p-values*

are easily available for subject sampling only. This type of problem may provide impetus to future research in GSA.

Inability to model and analyze a dynamic response: It is well known that biological systems are dynamic. There has been a long debate about the feasibility of using static models to model the inherent dynamics of biological systems. However, in GSA, only static approaches (linear, gamma, generalized linear and regression models) [80,98,99] have been used to date. This raises a serious concern about the use of GSA approach in assessing living systems. The lack of methods that analyze gene sets as a dynamic system is partly due to the limitations of current molecular measurement technologies. These technologies can only quantify a snapshot of a biological system because they are unable to: (*i*) determine the protein states in a high-throughput fashion, or are severely restricted in this regard; and (*ii*) detect signals that propagate without affecting GE. Therefore, we encourage researchers in the future to use dynamic models such as time-series models, auto-regressive models, dynamic Bayesian models, etc. for GSA from time-dependent GE or association data.

Redundancy among genes in gene sets: In GE data analysis, redundancy among genes (*i.e*., genes may not be related to a case/disease but ranked in the top due to high correlation with other top ranked genes) is a serious issue [18]. During the process of ranked gene list preparation, redundant genes may be included and further, do not give valid *p-values* for the gene set testing, as genes in gene lists are correlated. In other words, *p-values* may easily be falsely significant when the genes in the gene set are correlated, even when none of the

genes is truly significant. One strategy may be to use such a GE data analysis approach [18] which minimizes the redundancy among genes during the gene ranked list preparation. Other approaches may include avoiding the use of gene-sampling models in gene set testing for *p-value* computation [5].

Develop threshold-free approach(s): ORA based GSA approaches are mostly threshold dependent [14]. Further, other GSA methods such as mGSZ (based on Gene Set Z-scoring function) requires a threshold value for DE score to divide the ranked gene list into member genes and non-member genes (*i.e.,* two gene groups) [128]. Gene set testing (*e.g*., Z-test) is then performed on these gene groups [8,11,114,128]. The determination of an optimal threshold is often a cumbersome task. Therefore, the obtained analytical results from such approach are unstable and irreproducible [8,14,89]. Hence, researchers use a set of threshold values to compute enrichment significance of gene sets and then select the threshold that gives the most significant results [5,130]. This approach seems inelegant. A more comprehensive and computationally intensive approach for choosing a threshold will be a reasonable compromise among power, type I error, and reproducibility of results, using a cross validation technique. Another strategy may be development of threshold-free GSA approaches to improve the stability of results.

Proper permutation procedure: Current GSA approaches mostly use permutation procedures that compute *p-values* by comparing the observed test statistic with its null distribution generated from the permuted datasets [5,26,69,130]. It is expected to reflect chance-based confounding effects, including

biases introduced by the gene set. However, the permutation procedures (if not designed properly) can produce misleading results and introduce bias in the resulting inference. For instance, permutation of SNPs, which is often used in *p-value* based approaches, may disrupt the linkage disequilibrium pattern, and may not generate the correct null distribution. For gene-based approaches, permutation of sample labels may not generate the correct null distribution, as the samples are generated from tissues of same or related individuals [9,131]. Furthermore, whether the SNPs or genes or phenotypes are being permuted, the sampling units are assumed to be iid, which may not be the case; SNPs may be correlated due to linkage disequilibrium or gene-gene interactions. Therefore, proper care should be taken before choosing the permutation procedure for computing the *p-values* for gene sets.

GSA approach(s) for alternate annotations: The existing ORA based GSA approaches have mostly focused on whether the selected gene sets are over-represented by known pathways or GO terms [24,25,34,36,37,67,76,100,101]. However, in plant and complex disease biology, such approaches may not be able to establish any formal relation between the underlying genotypes and the trait/phenotype, as most of the traits are quantitative in nature and controlled by polygenes [13,17,132,133]. For this purpose, a statistical approach and R package of GSA with QTL has recently been developed [17], which is useful for obtaining QTL-enriched gene sets. Moreover, there is a lot of genomic annotation information, such as tissue information, QTL, *etc.,* available in the public domain, which can be used to develop new and innovative GSA approaches and tools.

Stability of gene set testing results: The statistical power and FDR are used for performance analysis of GSA approaches [12,15,26,74]. It is well known that different samples (on which the test is based) would give different results due to sampling errors. One way to deal with such a problem is to draw different sub-samples from a relative homogenous population, and the approach with small variance and uniform results over sub-samples can be termed as stable [134]. This principle can be applied to GSA, *i.e.,* first, sub-samples can be taken from all samples, and then GSA can be applied on each sub-sample to compute the *p-value* for the gene sets. Finally, one can evaluate the stability of the approach by comparing a change in ranks over different sub-samples. The approach with the least change in ranks can be termed as the stable approach and can be easily implemented in simulation analysis. In biology, several factors may be responsible for causing instabilities to the results; these include gene-gene correlations, genetic heterogeneity, and patient-to-patient variability. To address this problem, several researchers have hypothesized that testing gene sets rather than individual gene/marker will be more stable across different samples [26,135,136]. More relevant, and specialized studies and methodologies are needed to validate such claims.

"In 21st century all the biological investigations will be done *in silico…*"

Walter Gilbert, Nobel laureate

CHAPTER 3

STATISTICAL APPROACH FOR BIOLOGICALLY RELEVANT GENE SET SELECTION FROM GENE EXPRESSION DATA

Background

Emergence of high-throughput sequencing technologies exponentially increased the size of output data in biological sciences with respect to a number of features [137]. For example, GE studies generate the expression measurements of several thousand(s) of genes for tissue samples over two contrasting conditions in a single study [138,139]. These huge amounts of expression data being generated for complex traits, have been deposited in public domain databases, such as NCBI, ArrayExpress, *etc*., over the years by different researchers across the globe [3,140]. Moreover, these huge publicly available high-throughput datasets need to be analyzed for gaining valid biological insights. One such aspect of research may be to select genes which are highly relevant to the phenotype/trait under study from the several thousands of genes in the data. This is called feature selection in machine learning in general and gene selection in genomics [2,3,141]. Gene selection has been the focused area of research in functional genomics, and thus, several statistical, and machine learning approaches have been developed for this purpose [142,143]. The main aim of gene selection is to reduce the problem of high-dimensionality in expression data [2–4,144].

Further, these selected genes are used as predictors for further predictive analysis, *i.e.,* subjects/patients classification [141,142,144], regression modeling [145], gene network analysis [2,3], *etc*.

The gene selection methods can be grouped into: (*i*) filter; and (*ii*) wrapper methods [143,146]. Filter methods select individual genes or evaluate a gene subset based on a performance measure, *i.e.,* relevance or redundancy measure computed from the data with respect to class variables regardless of the predictive modeling algorithm [147]. Further, these methods include univariate approaches such as t-test [6,148], FC [148], F-score [149,150], Volcano plot [6], Wilcoxon's statistic (Wilcox) [151,152], Information Gain (IG) [143,153], Gain Ratio (GR) [143,153], symmetric uncertainty [19], *etc*. These methods select genes by only considering their relevance within a level of the experimental condition/trait. However, these approaches may not sufficient to discover some complex relationships among genes (*i.e.,* gene-gene interactions) for certain conditions/traits, under which the data are generated [4]. Multivariate filter approaches, *i.e.,* Pearson's Correlation (PCR), Spearman's rank correlation [143,153], Maximum Relevance and Minimum Redundancy (MRMR) [149,154,155], *etc*. have been developed to select genes from GE data [143,146]. Further, the wrapper methods select gene subsets through assessing the performance of the predictive modelling algorithm [156]. For instance, in classification, a wrapper will evaluate gene subsets based on the classifier's performance on GE data. Wrapper methods of gene selection are embedded in the classification process, better in performance over filter methods [143,146], but

are more complex and computationally expensive [156]. Support Vector Machine-Recursive Feature Elimination (SVM-RFE) [142,157], Multiple SVM-RFE (MSVM-RFE) [158] and Random Forest (RF) [144], *etc.* are classic examples of wrapper methods. Furthermore, hybrids of filter and wrapper methods are also reported in the literature (known as embedded methods [143]) such as linear combination of SVM-RFE and MRMR weights (SVM-MRMR) to select relevant genes from GE data [159]. In addition, other embedded methods were also developed by hybridizing SVM with other gene selection methods [150].

Gene selection methods were used to select cancer responsible genes from GE datasets, and subsequently used for patient classification (*e.g.,* with and without cancer) [18,141,142,144]. However, it is important and highly necessary to systematically explore the performance of gene selection methods on crop GE datasets. Further, the existing methods (*e.g.,* SVM-RFE, MRMR, SVM-MRMR, *etc.*) select genes through the weights (*i.e.,* gene ranking criteria) computed from single high-dimensional GE data, which lead to selection of spuriously associated genes [2,3]. Therefore, the permutation tests are used to compute statistical significance values for gene selection [2], which are highly sensitive to small permutations of experimental conditions (*i.e.,* class labels) [2,3], computationally slow [160,161], cannot possibly give any significant *p-values* after multiple testing adjustments [161,162], and large number of permutations are required to get a significant *p-value* [161]. Moreover, the performance of the existing methods were usually assessed through computation of post selection Classification Accuracy (CA) [18,141,142,159,163,164]. Here, it is worthy to note, this traditional criterion

is statistically sound but may not be biologically relevant for performance evaluation of gene selection methods [17,18]. Hence, it is pertinent to evaluate the gene selection methods with respect to biology and traditional classification-based criteria on multiple real crop GE datasets.

In this chapter, we propose an improved statistical approach (BSM=Bootstrap-SVM-MRMR) that combines MRMR filter with SVM wrapper methods to minimize the redundancy among genes and improve the relevancy of genes with the traits/phenotype under a sound statistical setup. Through this, relevant genes are selected from a high-dimensional GE data through the statistical significance values computed using a NP test statistic on bootstrap samples. Further, the comparative performance analysis of the proposed BSM approach is carried out, and compared with nine existing methods (*i.e.* IG [143,153], GR) [143,153] , t-test [6,148], F-score [149,150], MRMR [149,165] , SVM-RFE [142,157], SVM-MRMR [159], PCR [143,153] and Wilcox [151,152]). The comparative performance measures include, CA along with its standard error computed through varying sliding windows size technique, and two biological criteria based on QTL [166], and GO [167] terms. We demonstrate these procedures on six, publicly available, independent crop GE datasets, and find that the BSM approach outperforms in terms of classification and biological relevance criteria compared to the existing methods. Moreover, the developed approach provides an effective statistical framework for combining filter, and wrapper methods for gene selection from high dimensional GE data.

Material and Methods

Motivation for using Crop datasets

The datasets related to expression of genes from various experiments, conducted to understand the behavior of biological mechanisms, are widely available in public domain databases. For example, GE datasets generated for 125,376 series (experiments) over 19,893 Microarray platforms consist of GE data on 3,406,218 samples that are available in NCBI Gene Expression Omnibus (GEO) database till January 2020. Usually, researchers used data from a single experiment to test their methodology or select genes for further study. For instance, Wang *et al.* (2013) used the salinity stress GE samples from GSE14403 to test their methodology, and selected salinity responsive genes to understand salinity tolerance mechanism(s) in rice [2]. Such type of study is important but may not be sufficient to test the hypothesis of salinity tolerance in rice. Hence, the real challenge is to integrate or combine the GE datasets generated for same or cross platforms over different experimental conditions and test the methodology(s) on the meta-data. For instance, we have collected GE datasets related to drought stress from five different experiments and performed meta-analysis to integrate the datasets, and further tested the performance of gene selection methods on the meta-data, shown in Table 3.1. The outlines of meta-analysis are given in Figure 3.1A. Moreover, meta-analysis of data generated by GE experiments for the same or related stress(es) will be essential to enhance the sensitivity of the hypothesis under consideration for drawing valid biological conclusions.

Data source

Rice GE experimental datasets were collected from NCBI GEO database for platforms GPL2025 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2025) [140]. Here, we used the rice data, as it is a model crop plant, and because a huge amount GE and other related biological (QTL and GO) datasets are available publicly, and its genome is well annotated. The selected GE datasets were generated under biotic (bacterial (*Xanthomonas*), fungal (Blast), insect (Brown plant hopper)), and abiotic (salinity, cold and drought) stresses in rice. The summary and details of these datasets are given in Table 3.1. Initially, the raw CEL files of the collected samples were processed using Robust Multichip Average (RMA) algorithm available in *affy* Bioconductor package of R [168]. This procedure involves background correction, quantile normalization, and summarization by median polish approach. Further, the log2 scale transformed expression data for the collected experimental samples were used for meta-analysis to remove the outliers. The GE samples from three, four, five, three, and two independent studies for salinity, cold, drought, bacterial, and fungal stresses respectively, were integrated (Table 3.1) through the meta-analysis to obtain the meta-data. For instance, the salinity stress dataset, originated from three independent studies, are available in GEO database under the accession numbers GSE14403, GSE16108, and GSE6901 and consists of expression measurements over 45 samples. Then these meta-datasets for the respective stresses were further used to remove the control and irrelevant features through the preliminary genes selection to reduce the computational complexity, and dimensions of the datasets. For instance, out

of 57381 genes in drought stress, control (123), and irrelevant (48180) genes are filtered out by setting the FC, and *p-values* (from t-test) parameters as 1, and 0.05 respectively through the preliminary gene selection. Then, the processed datasets (Table 3.1) were used for further data analysis. Further, the QTL datasets for the stresses in rice, *viz.* salinity, drought, cold, insect, fungal and bacterial, were collected from the Gramene QTL database (http://www.gramene.org/qtl/) [169]. The GO annotations data of the rice genome were collected from *AgriGO* database [24].

Table 3.1. Rice gene expression datasets used in the study.

Descriptions	#Series	Series Id	#Genes	#Samples	Type
Salinity stress	3	GSE14403, GSE16108, GSE6901	6637	45 (23, 22)	Abiotic
Cold stress	4	GSE31077, GSE33204 GSE37940, GSE6901	8840	28 (15, 13)	Abiotic
Drought stress	5	GSE6901, GSE26280 GSE21651, GSE23211 GSE24048	9078	70 (35, 35)	Abiotic
Bacterial stress	3	GSE19239, GSE36093 GSE36272	8356	74 (37, 37)	Biotic
Fungal stress	2	GSE41798, GSE7256	7072	26 (13, 13)	Biotic
Insect stress	1	GSE29967	7241	18 (12, 6)	Biotic

#Series: Number of GEO series for each dataset; #Genes: Number of genes; #Samples: Number of GEO samples; (x, y): number samples for case and control respectively

Notations

Let, $X_{N \times M} = [x_{im}]$ be the GE data matrix, where x_{im} represents the expression of i^{th} (i=1, 2, …, N) gene in m^{th} (m = 1, 2, …, M) sample/subject; $\boldsymbol{x}_m$ be the N-dimensional vector of expression values of genes for m^{th} sample; y_m be the outcome variable for target class label of m^{th} sample and takes values, {+1, -1} for case and control conditions respectively; M_1 and M_2 be the number of GE samples in case and control classes respectively ($M_1 + M_2 = M$); $(\bar{x}_{i1}, S_{i1}^2)$ and $(\bar{x}_{i2}, S_{i2}^2)$ be the mean

and variance of i^{th} gene for case and control classes respectively; $\bar{x}_i$ be the mean of i^{th} gene across all *M* samples; S_{ij} be the co-variance between i^{th} and j^{th} genes.

MRMR Filter Method

MRMR method aims at selecting maximally relevant and minimally redundant set of genes for discriminating the tissue samples (*e.g.* case *vs*. control). This method is extensively used for selection of cancer responsible genes from high-dimensional GE data for patient classification (*i.e.* with and without cancer) [149,155,165]. For continuous GE data (*e.g.* Microarrays), the relevance measure for i^{th} gene over the given classes (*i.e.* case and control), is computed through F-statistic [165] and is expressed as:

$$F(i) = \frac{M_1(\bar{x}_{i1}-\bar{x}_i)^2+M_2(\bar{x}_{i2}-\bar{x}_i)^2}{\{(M_1-1)S_{i1}^2+(M_2-1)S_{i2}^2\}/(M-2)} \tag{3.1}$$

Further, the redundancy measure in the MRMR method is computed through Pearson's correlation (ignoring the class information) for continuous GE data [165] and is given as:

$$R(i,j) = Corr(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{S_{i,j}}{S_i S_j} = \frac{\sum_{m=1}^{M}(x_{im}-\bar{x}_i)(x_{jm}-\bar{x}_j)}{\sqrt{\sum_{m=1}^{M}(x_{im}-\bar{x}_i)^2}\sqrt{\sum_{m=1}^{M}(x_{jm}-\bar{x}_j)^2}} \tag{3.2}$$

In MRMR method, genes are ranked by the combination of relevance, and redundancy measures under F-score with Correlation Quotient scheme for continuous GE data [149,155,165]. The weights computed through MRMR method for gene ranking can be expressed in terms of Eq. 3.1 and 3.2, and are given as:

$$w_i = F(i)/\{\frac{1}{N-1}\sum_{\substack{j=1 \\ j\neq i}}^{N}|R(i,j)|\} \qquad \forall \quad i = 1,2,\ldots,N \tag{3.3}$$

where, w_i (≥ 0) is the weight associated with i^{th} gene. The functions $F(i)$ and $R(i, j)$ in Eq. 3.3 represent F-statistic for i^{th} gene and Pearson's correlation co-efficient

between i^{th} and j^{th} genes. In other words, i^{th} gene weight is F-statistic adjusted with average absolute correlation of i^{th} gene with the remaining genes.

SVM Method

SVM method is used for selection of genes (in a two group case) from high dimensional GE data [157]. Let, $\{\boldsymbol{x}_m, y_m\} \epsilon R^N \mathrm{x} \{-1, 1\}$ be the input given to SVM. Here, we wish to find out a hyperplane that divides the GE samples/subjects for case ($y_m = 1$) from that of control class ($y_m = -1$) in such a way that the distance between the hyperplane and the point, $\boldsymbol{x}_m$, is maximum. Then the hyperplane can be written as:

$$\sum_{i=1}^{N} k_i \boldsymbol{x}_{im} + b = 0 \qquad \forall\, m = 1, 2, \dots, M \tag{3.4}$$

where, k_i and b are the weight of i^{th} gene and bias, respectively. Here, we assume that the GE samples for two classes are linearly separable. In other words, we can select two parallel hyperplanes that separate the case, and control classes in such a way that the distance between them is maximum.

For case class, the hyperplane becomes:

$$\sum_{i=1}^{N} k_i x_{ip} + b = 1 \qquad \text{for any } p = 1, 2, \dots, M_1 \tag{3.5}$$

For control class, the hyperplane becomes:

$$\sum_{i=1}^{N} k_i x_{iq} + b = -1 \qquad \text{for any } q = 1, 2, \dots, M_2 \tag{3.6}$$

The expressions in Eq. 3.5, and 3.6 can be combined as:

$$y_m\left(\sum_{i=1}^{N} k_i \boldsymbol{x}_{im} + b\right) = 1 \qquad \forall\, m = 1, 2, \dots, M \tag{3.7}$$

Here, we wish to maximize the distance between the case, and control hyperplanes in Eq. 3.5 and 3.6 respectively under the constraint that there will be

no GE samples between these two hyperplanes given in Eq. 3.7. Mathematically, it can be written as:

$$\sum_{i=1}^{N} \frac{k_i}{\sum {k_i}^2} \left| x_{ip} - x_{iq} \right| = \frac{\mathbf{2}}{\sum {k_i}^2} \quad (3.8)$$

So, to maximize the distance between the planes in Eq. 3.8, we need to minimize $\frac{\sum {k_i}^2}{2}$ under the constraint of Eq. 3.7. Mathematically, it can be written as:

$$L_p = \min_{k_i} \frac{\sum {k_i}^2}{2} + \sum_{m=1}^{M} \varphi_m \{1 - y_m (\sum_{i=1}^{N} k_i x_{im} + b)\} \quad \forall\, m = 1, 2, \ldots, M \quad (3.9)$$

where, $\varphi_m \ (\geq 0)$: Lagrange multiplier. Here, k_i's are obtained by minimizing the objective function in Eq. 3.9. Through the principle of maxima-minima, we have:

$$\frac{\partial L_p}{\partial k_i} = \sum_i k_i - (\sum_{m=1}^{M} \varphi_m y_m x_{im}) = 0 \ \text{ and } \ \frac{\partial L_p}{\partial b} = \sum_{m=1}^{M} \varphi_m \, y_m = 0 \quad (3.10)$$

The value of k_i can be obtained through solving the system of linear equations given in Eq. 3.10 and is expressed as:

$$k_i = \sum_{m=1}^{M} \varphi_m y_m x_{im} \quad \text{with } \sum_{m=1}^{M} \varphi_m \, y_m = 0 \text{ and } \varphi_m \geq 0 \quad (3.11)$$

Here, $|k_i| \ (\geq 0)$ in Eq. 3.11 is used as a metric for ranking of genes in the GE data [157]. Alternatively, k_i^2 as a gene ranking metric can also be derived by using Taylor series approximation [170], which is given in Appendix II.

Proposed BSM approach of gene selection

MRMR method may not yield optimal CA because it performs independent of the classifier and is only involved in selection of genes [159]. On the other hand, the SVM method of gene selection does not consider the redundancy among genes (*i.e.,* gene-gene correlations) while selecting genes [159]. Hence, Mundra and Rajapakse (2010) have developed a gene selection method by taking linear

combination of weights computed through MRMR, and SVM methods [159], and this is given as:

$$SL_i = \delta w_i + (1-\delta)|k_i| \tag{3.12}$$

where, parameter $\delta \in [0,1]$ decides the trade-off between SVM and MRMR weights. The SL_i in Eq. 3.12 is highly dependent on the value of δ. In other words, the choice of δ may alter the order of genes by MRMR (w_i) or by SVM (k_i); especially when w_i and k_i are negatively correlated. Hence, we propose a statistical approach by combining SVM and MRMR weights under sound statistical framework, where genes are selected through *p-values* computed using the NP test statistic, which is described as follows.

First, we normalized the w_i, and k_i's through minimax normalization. Then w_i and k_i are ranked based on the ascending order of their magnitudes and assign ranks $\gamma_i^{MR}$and γ_i^{SV} for i^{th} gene, respectively. Then, we developed a technique, *i.e.,* quadratic integration, for integrating the gene scores based on ranks, which automatically assigned more weights to the higher value of w_i and k_i. Now, the quadratic integration score can be expressed as:

$$SD_i = \frac{\beta\gamma_i^{MR} {w_i}^{norm} + (1-\beta)\gamma_i^{SV}|k_i|^{norm}}{\beta\gamma_i^{MR} + (1-\beta)\gamma_i^{SV}} \tag{3.13}$$

where, ${w_i}^{norm}$ and $|k_i|^{norm}$ are the normalized values, expressed in Eq. 3.14 and 3.15, respectively.

$${w_i}^{norm} = \left(w_i - \min_i w_i\right)/\left(\max_i w_i - \min_i w_i\right) \tag{3.14}$$

$$|k_i|^{norm} = \left(|k_i| - \min_i |k_i|\right)/\left(\max_i |k_i| - \min_i |k_i|\right) \tag{3.15}$$

Further, $\beta(\in (0,1))$ in Eq. 3.13 is determined empirically from the data through five-fold cross validation technique. If SD_i in Eq. 3.13 is alone used for ranking of genes, it will become a filter approach and lead to selection of spuriously associated genes. Hence, we used bootstrap procedure under a subject sampling model setup to obtain the empirical distribution of SD_i for computation of statistical significance value for i^{th} (i=1, 2, …, N) gene. Here, the bootstrap procedure is described below.

The M samples (as columns) in the GE data matrix either belong to case or control class, and can be considered as subjects/units in a population model, as shown in Eq. 3.16.

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m), \ldots, (\boldsymbol{x}_{M-1}, y_{M-1}), (\boldsymbol{x}_M, y_M) \tag{3.16}$$

Here, we assume that the subjects are independent and identically distributed, but the genes within each subject may be correlated. In the bootstrap procedure, M units are randomly drawn from M population units in Eq. 3.16 with replacement to constitute a bootstrap GE data matrix, *i.e.* $X_{NXM}^{(b)}$ (M units serve as M columns of X). This process is repeated B times to get B bootstrap GE data matrices, *i.e.* $X_{NXM}^{(1)}, X_{NXM}^{(2)}, \ldots, X_{NXM}^{(b)}, \ldots, X_{NXM}^{(B)}$. Here, B (*i.e.* number of bootstrap samples) depends on several factors such as, number of units in the population model in Eq. 3.16 and must be sufficiently large. So, we set B=200 as several empirical studies showed that the number of bootstrap samples required for an estimation procedure is ~ 200 [2,171].

Now, the *B* bootstrap GE data matrices are given as input to Eq. 3.3, 3.11, and 3.13 to compute the *SD* scores, and subsequently gene ranking is performed on each of the *B* bootstrap GE data matrices.

Let, P_{ib}, be a random variable (*rv*) shows the position of i^{th} gene in b^{th} bootstrap GE matrix. Then, another *rv* can be defined based on P_{ib} (without loss of generality), given as:

$$R_{ib} = \frac{N+1-P_{ib}}{N} \; ; \; 0 \leq R_{ib} \leq 1 \tag{3.17}$$

where, R_{ib} in Eq. 3.17 is the rank score of i^{th} (i=1, 2, …, N) gene in b^{th} (b=1, 2, …, B) bootstrap GE matrix. Here, it may be noted that the distribution of the rank scores of genes, computed from a bootstrap GE data matrix, is symmetric around the median value (as rank scores are function of ranks). The values of median and the third quartile (Q_3) are given as 0.5 and 0.75, respectively.

To decide, whether i^{th} gene is biologically relevant or not to the condition/trait under study, the following null hypothesis can be tested.

$$H_0 : R_i \leq Q_3 \; (i-th \text{ gene is not so relevant to the trait})$$

$$H_1 : R_i > Q_3 \; (i-th \text{ gene is relevant to the trait})$$

where, R_i is the rank score for i^{th} gene over all possible bootstrap samples.

To obtain the distribution of test statistic under H_0, we define another *rv* Z_{ib}, as:

$$Z_{ib} = \begin{cases} 1 & (R_{ib} - Q_3) > 0 \\ 0 & (R_{ib} - Q_3) < 0 \end{cases} \tag{3.18}$$

Let, r_{ib} be another *rv* represents the rank assigned to $(R_{ib} - Q_3)$ (after arranging in ascending order of their magnitudes). To test H_0 *vs*. H_1 the test statistic for i^{th} gene, W_i, is developed, and is given as:

$$W_i = \sum_{b=1}^{B} Z_{ib} r_{ib} = \sum_{b=1}^{B} U_{ib} \; (say) \tag{3.19}$$

In other words, W_i in Eq. 3.19 is the sum of the ranks of positive signed scores for i^{th} gene over B bootstrap samples. Further, U_{ib} in Eq. 3.19 is a Bernoulli *rv,* and its Probability Mass Function (PMF) can be given as:

$$P[U_{ib} = u_{ib}] = \begin{cases} \frac{3}{4} & if\ u_{ib} = 0 \\ \frac{1}{4} & if\ u_{ib} = 1 \end{cases} \tag{3.20}$$

Here, the expected value and variance of W_i, in Eq. 3.19, under H_0 can be obtained as:

$$E(W_i) = \sum_{b=1}^{B} E(U_{ib}) = \sum_{b=1}^{B} (0.\frac{3}{4} + b.\frac{1}{4}) = \frac{1}{4}\sum_{b=1}^{B} b = \frac{B(B+1)}{8} \tag{3.21}$$

The variance becomes:

$$V(W_i) = E(W_i^2) - [E(W_i)]^2 = \sum_{b=1}^{B} \left(\frac{b^2}{4} - \frac{b}{16}^2\right) = \frac{B(B+1)(2B+1)}{32} \tag{3.22}$$

As B is sufficiently large, then under the central limit theorem, the distribution of W_i in Eq. 3.19 becomes:

$$Z_i = \frac{W_i - E(W_i)}{\sqrt{V(W_i)}} \xrightarrow{d} N(0,1) \tag{3.23}$$

Through the Eq. 3.23, the *p-value* for i^{th} (i=1, 2, …, N) gene is computed and this testing procedure is similarly repeated for the remaining N-1 genes. Let, $p_1, p_2, \ldots, p_N$ be the corresponding *p-values* for all the genes in GE data, and α be the level of significance. Here, we assume that all genes in the GE data are equally important for the trait development; hence, we employed the Hochberg procedure [172] for correcting the multiple testing, and to compute the adjusted (*adj.*) *p-values* for genes. It is worthy to note that Hochberg's procedure is computationally simple, quite popular in genomic data analysis [173] and more powerful than Holm's procedure [174]. The algorithm for Hochberg's procedure [172] is as follows.

Step 1. If $p_{(l)} > \alpha$, then retain corresponding null hypothesis ($H_{(l)}$) and go to the next step. Else reject it and stop.

Step $i = 2, 3, \ldots, N-1$. If $p_{(N-i+1)} > \alpha/i$, then retain $H_{(N-i+1)}$ and go to the next step. Else reject all remaining hypotheses and stop.

Step *N*. If $p_{(1)} > \alpha/N$, then retain ($H_{(1)}$). Else reject it.

Now, the *adj. p-values* are given recursively beginning with the largest *p-value* [172]:

$$\widetilde{p_{(i)}} = \begin{cases} p_{(i)} & if\ i=N \\ \min\left(\widetilde{p}\ (i+1), (N-i+1)p_{(i+1)}\right) & if\ i = N-1, \ldots, 1 \end{cases} \quad (3.24)$$

Further, based on the computed *adj. p-values,* the relevant genes are selected from the GE data. In other words, a lesser value of *adj. p-value* indicates more relevance of the gene for the target trait, and *vice-versa*. The outlines and key analytical steps of the proposed BSM approach are shown in Figure 3.1B.

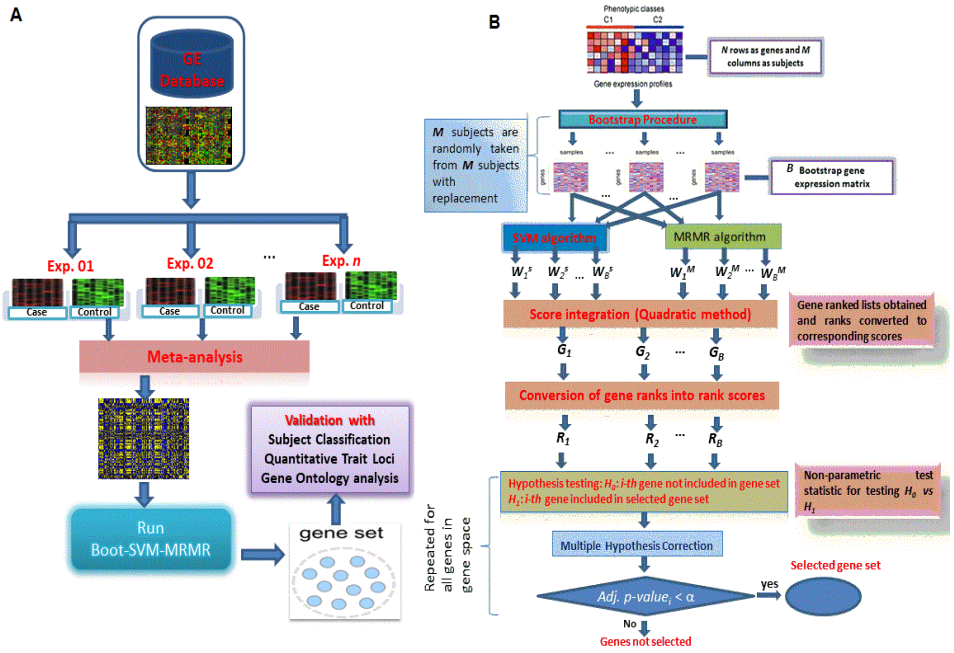


Figure 3.1. Operational procedure for data integration and the use of proposed BSM approach. (A) Outlines for the data integration used in this study for the application of BSM approach. The first step indicates the integration and meta-analysis of GE datasets obtained from

various GE studies. Then gene selection methods are applied on the meta GE data. (B) Flowchart depicting the implemented algorithm of BSM approach. $W_i^{(S)}$'s and $W_i^{(M)}$'s are the N-dimensional vectors of weights computed through SVM and MRMR approach, respectively. G_i's and R_i's are the N-dimensional vectors of gene lists and corresponding gene rank scores. SVM, MRMR stand for Maximum Relevance and Minimum Redundancy and Support Vector Machine algorithms. p_i-*value* is the statistical significance value for i^{th} gene. α is the desired level of statistical significance.

Comparative performance analysis of the proposed approach

The comparative performance analysis of the proposed BSM approach with respect to 9 competitive gene selection methods was carried out on 6 different rice GE datasets (Table 3.1). For this purpose, different gene sets (***G***) of sizes 100, 200, …, 2000, were selected through the 10 gene selection methods including proposed BSM approach. Then, these gene sets were validated with respect to subject classification, QTL testing, and GO analysis.

Performance analysis with subject classification

Under this comparison setting, the performance of the gene selection methods including the proposed approach, were assessed in terms of subject classification using mean CA, and Standard Error (SE) in CA as computed through a varying sliding window size technique [3,18]. Here, we used the varying window size technique to study the impact of gene ranking on classification of subjects. In other words, genes in ***G*** were validated with their ability to discriminate the class labels of subjects/samples between case (+1), and control (-1). Further, the gene set selected through a method which provides maximum discrimination between the subjects of two groups (*i.e.,* case *vs*. control) through CA will be considered as highly relevant gene sets. The expressions for mean CA, and SE in CA computed through varying window size technique are given in Eq. 3.25 and 3.26.

Let, n be the size of ***G***, S be the size of the windows (*i.e.,* size refers to number of ranked genes), and L be the sliding length. Then, the total number of

windows becomes, $K = (n - S)/L$. The genes in ***G****,* arranged in different windows along with their expression values, were then used in SVM classifiers with four basis-functions, *i.e.,* linear (SVM-LBF), radial (SVM-RBF), polynomial (SVM-PBF), and Sigmoidal (SVM-SBF) to compute CA over a five-fold cross validation. Let, CA_1, CA_2, …, CA_K be the CA's for each sliding windows, then the mean CA and SE in CA can be defined as:

$$\mu_{CA}^{G} = {\left(\sum_{k=1}^{K} CA_k\right)}/{K} \tag{3.25}$$

$$SE_{CA}^{G} = \sqrt{{\sum_{k=1}^{K}(CA_k - \mu_{CA}^{G})^2}/{K}} \tag{3.26}$$

Here, we took different combinations of *n*, *S,* and *L* to compute μ_{CA}^{G} and SE_{CA}^{G} for the comparative performance analysis of the gene selection methods. The higher value of μ_{CA}^{G}, and a lower value of SE_{CA}^{G} indicate the better performance of the gene selection method, and *vice-versa.*

Performance analysis with QTL testing

The comparative criteria based on subject classification are popularly used for assessing the performance of gene selection methods [18,141,142,159,163–165]. However, these criteria fails to tell the biological relevancy of the genes selected through the gene selection methods [17]. Hence, under this comparative setting we assessed the performance of the proposed and existing methods through their ability to select genes which are associated with QTL regions. For this purpose, the criteria given in Eq. 3.27 and 3.28 were developed as:

$$Qstat = \sum_{t=1}^{|Q|} \sum_{i=1}^{n} I_{q_t}(g_i) \tag{3.27}$$

where, ***G***: gene set selected by a method, *Qstat: rv* whose values represent the

number of genes covered by QTLs, *Q*: set of associated QTLs, and the indicator function present in Eq. 3.27 is represented in Eq. 3.28.

$$I_{q_t}(g_i) = \begin{cases} 1 & if\ g_i^c[a,] \geq q_t^c[d,]\ and\ g_i^c[,b] \leq q_t^c[,e] \\ 0 & else \end{cases} \quad (3.28)$$

where, g_i^c [*a*, *b*] ϵ ***G*** (*a* and *b* represent start and stop positions in terms of bp of the gene g_i on chromosome *c*) and q_t^c[*d*, *e*] ϵ *Q* (*d* and *e* represents the start and stop positions of the QTL q_t on chromosome c). Here, the $Qstat$ *rv* follows a hyper-geometric distribution and its distribution function is given in Eq. 3.29.

$$\mathrm{P}[Qstat = v] = 1 - \binom{V}{v}\binom{N-V}{n-v}/\binom{N}{n} \quad (3.29)$$

where, *V*: total number of genes covered by the QTLs in the whole GE data and *v*: number of genes in ***G*** that are covered by QTLs. Further, using the Eq. 3.29, the statistical significance value (*p-value*) associated with the ***G*** can be computed. In other words, this *p-value* reveals the enrichment significance of ***G*** with trait specific QTLs. Here, the higher values of $Qstat$ and $-log_{10}(p-value)$ indicate the better performance of the gene selection method, and *vice-versa*.

Performance analysis with GO enrichment

GO analysis involves with annotation of gene functions under three taxonomic categories, *i.e.* Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) [167]. This analysis helps in evaluating the functional similarities among the genes in ***G*** [175], as there exists a direct relationship between semantic similarity of gene pairs with their structural (sequence) similarity [176,177]. Under this comparison setting, we assessed the performance of 10 gene selection methods including the proposed method using GO based biologically relevant criterion. In other words, first different gene sets are selected through these

methods, then GO based criterion is computed for each selected gene set. For this purpose, we developed a GO based semantic distance measure to assess the GO based biologically relevancy of ***G*** selected thorough the proposed and existing gene selection methods. The GO based semantic distance measure (d_{ij}) between i^{th} and j^{th} genes can be expressed in Eq. 3.30, as:

$$d_{ij\,(i\neq j)}^{GO} = 1 - \frac{|GO_i \cap GO_j|}{|GO_i \cup GO_j|} \qquad \forall\, i, j = 1, 2, \ldots, n \tag{3.30}$$

where, $GO_i = \{go_{i1}, go_{i2}, \ldots, go_{il}\}$ and $GO_j = \{go_{j1}, go_{j2}, \ldots, go_{jJ}\}$ are the two sets of GO terms that annotate i^{th} and j^{th} genes in ***G***, respectively. Further, the GO based average biologically relevant score for ***G*** (for a gene selection method) can be developed based on Eq. 3.30 and is shown in Eq. 3.31.

$$D_G^{avg} = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} d_{ij}^{GO} \tag{3.31}$$

where, D_G^{avg} in Eq. 3.31 represents is the average biologically relevant score for ***G*** based on GO annotations. Using Eq. 3.31, the D_G^{avg} scores under MF, BP, and CC taxonomies were computed for each of the gene sets selected through different methods. A lower value of D_G^{avg} indicates better performance of the gene selection method and *vice-versa*.

Results and Discussion

Computation of genes selection criteria through proposed approach

The distributions of weights computed from SVM-MRMR method [159] and *adj. p-values* for genes computed from the proposed BSM approach for abiotic and biotic stresses in rice are shown in Figure 3.2. The distributions of SVM-MRMR weights of genes for salinity stress data contained values, which are not so clearly

separated (*i.e.,* higher values from lower values) (Figure 3.2A). In other words, the genes relevant to the salinity stress condition were not well visualized from Figure 3.2A. However, from the distribution of *adj. p-values* computed through the proposed approach, it was observed that the relevant genes were found to be well separated from the irrelevant genes, and a small number of genes were found to be statistically significant (*i.e.,* relevant to salinity stress) (Figure3.2A1). Thus, for the salinity stress data, the separation between relevant, and irrelevant genes can be well visualized from Figure 3.2A1 as compared to Figure 3.2A. Similar interpretations can be observed for other stress datasets, *viz*. cold, drought, bacterial, fungal, and insect (Figure 3.2). Hence, the comparative graphical analysis showed a clear distinction between relevant, and irrelevant genes through the proposed BSM approach as compared to the existing SVM-MRMR approach. In summary, this comparative analysis showed the improvement of BSM approach over SVM-MRMR method (Figure 3.2), at least in terms of visualization. Further, the relevant genes selection using *adj. p-values* computed through the NP test statistic is more statistically sound as it is independent from the distribution of GE data, and corrected over multiple hypothesis testing. These metrics (values between 0 and 1) are scientifically well defined, and statistically calculated biologically interpretable values to genome researchers, and experimental biologists as compared to gene ranks/weights. In the BSM approach, a significant *p-value* gives confidence that the given gene is relevant for the target condition/trait.

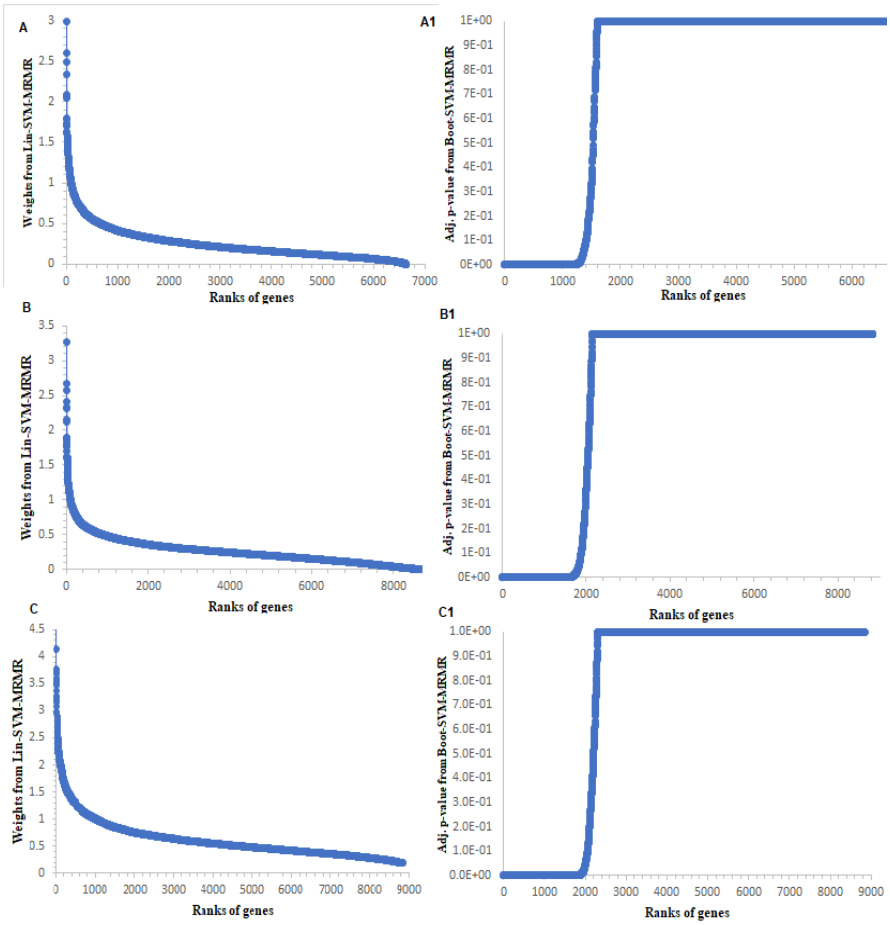


Figure 3.2. Graphical analysis of the proposed BSM approach with SVM-MRMR approach for abiotic stress datasets. Distribution of gene weights computed from SVM-MRMR approach for the abiotic stresses. The distributions of gene weights from the SVM-MRMR are shown for (A) Salinity; (B) Cold; and (C) Drought stress datasets in rice. Distribution of adjusted *p-values* computed from the proposed BSM approach for the abiotic stresses. The distributions of the adjusted *p-values* are shown for (A1) Salinity; (B1) Cold; and (C1) Drought stress datasets.

Comparative performance analysis based on subject classification

We used μ_{CA}^{G} and SE_{CA}^{G} computed through the varying sliding window size technique as statistically necessary criteria for performance analysis of the proposed BSM approach on 6 different GE datasets. Here, these measures were computed over five-fold cross validations through training the SVM-LBF, SVM-PBF, SVM-RBF, and SVM-SBF classifiers. The results are shown in Figures 3.3 and 3.4 for abiotic stresses and in Figure 3.5 for biotic stresses. For cold stress

data in rice, the μ_{CA}^{G} computed through SVM-LBF classifier for the proposed BSM approach was observed to be higher than other gene selection methods followed by SVM-RFE and SVM-MRMR over all selected gene sets (Figure 3.3). This indicated a better performance of the BSM approach in terms of its ability to classify the subjects/samples through selecting relevant genes from cold stress GE data. Further, the values of SE_{CA}^{G} from SVM-LBF classifier for the BSM approach was found to be much lower (over all the gene sets) than those of the 9 existing gene selection methods considered in this study. This shows that the genes selected through the proposed BSM approach are much more relevant (informative) and robust than other methods. For instance, the gene set of size 50 (*i.e.,* optimum gene set) provided satisfactory results in terms of higher μ_{CA}^{G} and lower SE_{CA}^{G}, irrespective of the gene selection method used. For cold stress data, similar interpretations can be made for SVM-PBF, SVM-RBF, and SVM-SBF classifiers from Figures 3.3 and 3.4.

For salinity stress data, the μ_{CA}^{G} (except gene sets of sizes 500, 1000 and 1500) computed for the proposed BSM approach through the SVM-LBF classifier were found to be higher than other methods followed by SVM-RFE, and SVM-MRMR (Figure 3.3). This indicated the proposed approach was better and was competitive with the popular methods, *i.e.,* SVM-RFE, SVM-MRMR. Moreover, for SVM-PBF classifier, the μ_{CA}^{G} over all gene sets for the BSM approach was higher than all other considered gene selection methods followed by SVM-RFE (Figure 3.3). Furthermore, the SE_{CA}^{G} computed through SVM-LBF, and SVM-PBF classifiers for the BSM approach was found to be the least followed by SVM-RFE,

indicating the selection of robust, and relevant gene sets. Similar interpretation can be made for SVM-RBF and SVM-SBF classifiers from Figure 3.4. It was observed that the μ_{CA}^{G} from the SVM-SBF classifier was found to be the least (with high SE_{CA}^{G}) among the four classifiers for all the datasets (Figure 3.4). Here, it is pertinent to note that the sigmoid basis function may not be recommended to use in SVM training for real crop GE datasets. Furthermore, similar interpretations can be made for other abiotic (*i.e.* drought) and biotic (*i.e.* bacterial, fungal and insect) stress GE datasets (Figures 3.3-3.5).

The classification-based performance metrics can be considered as statistically necessary to check the informativeness, and robustness of the selected genes.

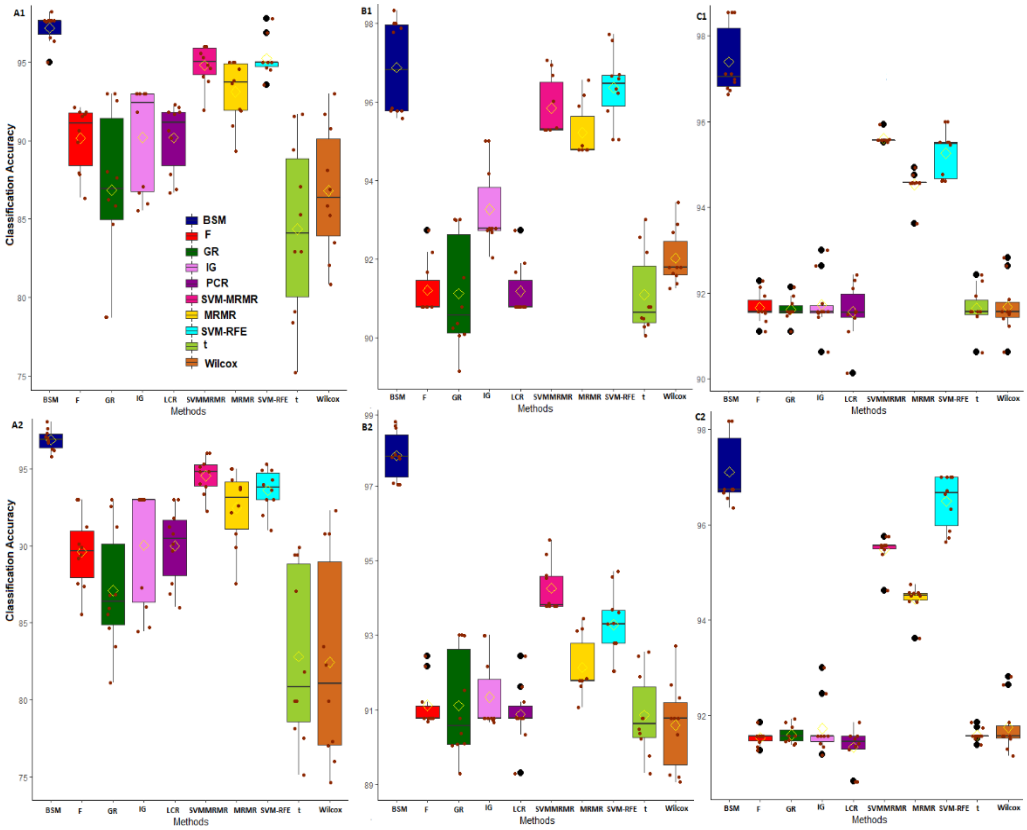


Figure 3.3. Classification based comparative performance analysis of gene selection methods through SVM-LBF and SVM-PBF Classifiers for abiotic stress

datasets. The horizontal axis represents the gene selection methods. The vertical axis represents post selection classification accuracy obtained by using varying sliding window size technique. The classification accuracies over the window sizes are presented as boxes. The bars on the boxes represent the standard errors. The distributions of classification accuracies are shown for Cold stress with SVM-LBF (A1), and SVM-PBF (A2) classifiers. The distributions of classification accuracies are shown for Salinity stress with SVM-LBF (B1), and SVM-PBF (B2) classifiers. The distributions of classification accuracies are shown for Drought stress with SVM-LBF (C1), and SVM-PBF (C2) classifiers.

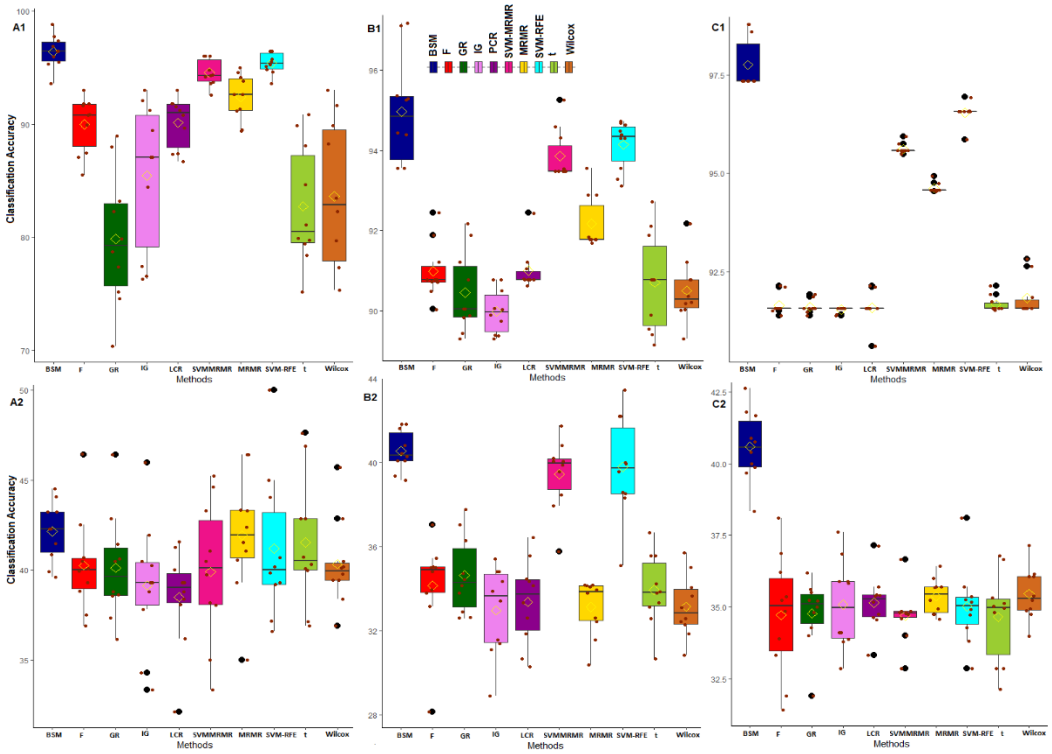
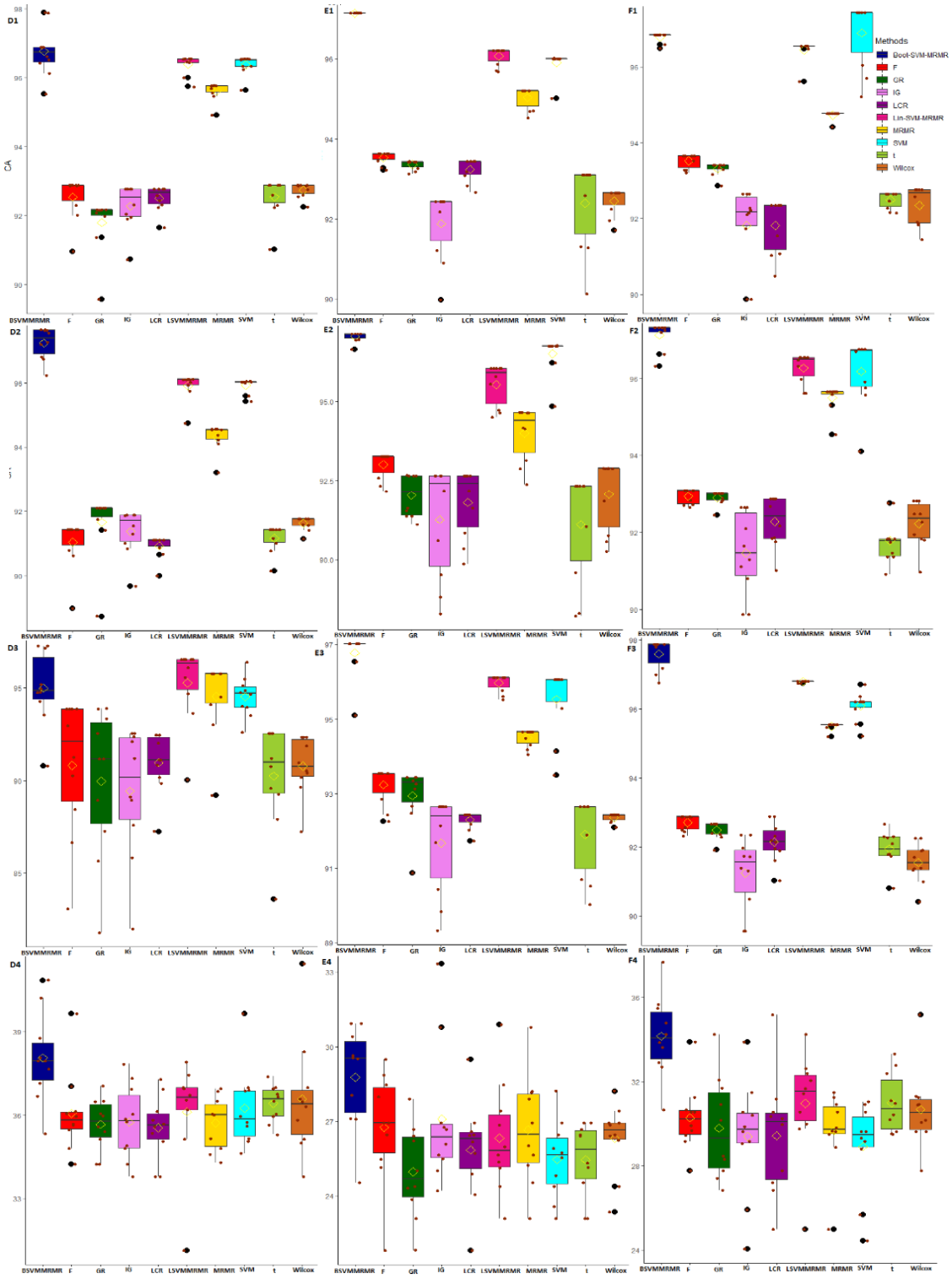


Figure 3.4. Classification based comparative performance analysis of gene selection methods through SVM-RBF and SVM-SBF Classifiers for abiotic stress datasets. The horizontal axis represents the gene selection methods. The vertical axis represents post selection classification accuracy obtained by using varying sliding window size technique. The classification accuracies over the window sizes are presented as boxes. The distributions of classification accuracies are shown for Cold stress with SVM-RBF (A1), and SVM-SBF (A2) classifiers. The distributions of classification accuracies are shown for Salinity stress with SVM-RBF (B1), and SVM-SBF (B2) classifiers. The distributions of classification accuracies are shown for Drought stress with SVM-RBF (C1), and SVM-SBF (C2) classifiers.

Figure 3.5. Classification based comparative performance analysis of gene selection methods in biotic stresses. The distributions of classification accuracies are shown for Bacterial stress dataset with SVM-LB (D1), SVM-PBF (D2) SVM-RBF (D3), and SVM-SBF classifiers (D4); The distributions of classification accuracies are shown for rice Fungal stress dataset with SVM-LBF (E1), SVM-PBF (E2) SVM-RBF (E3), and SVM-SBF (E4) classifiers; The distributions of classification accuracies are shown for rice Insect stress dataset with SVM-LBF (F1), SVM-PBF (F2), SVM-RBF (F3), and SVM-SBF (F4) classifiers.



Through such analysis, it was found that the BSM approach performed better in terms of selecting informative, and robust genes from the high-

dimensional GE data as compared to other competitive methods such as SVM-RFE, MRMR, SVM-MRMR, and the information theoretic measures. The reason may be attributed to the inclusion of bootstrap based subject sampling model along with the self-contained statistical tests, which reduces the spurious association of genes with the target trait as well as with other genes. Further, the performance of BSM approach, in terms of the ability to classify the GE samples, found to be better as compared to multivariate approaches, *i.e.,* MRMR, SVM-MRMR, univariate approaches, (t-test, F-score, Wilcox, and informative theoretic measures, *i.e.,* IG and GR). Here, it is worthy to note that the multivariate approaches performed better compared to the univariate approaches when assessed under classification-based criteria, as the former considers gene-gene associations.

Comparative performance analysis based on QTL testing

We used publicly available QTL data to statistically measure the biological relevancy of the genes selected through the proposed and existing gene selection method(s). The main rationale behind such analysis is that the genes selected for a stress condition (through a gene selection method) are expected to overlap with that stress specific QTL regions. Therefore, the QTLs and genes selected through these 10 gene selection methods, including the proposed BSM, are mapped to the whole rice genome using MSU rice genome browser [178].

The biological relevance of the selected genes through both proposed and existing gene selection methods were measured through two criteria, *i.e., Qstat* and *-log*10(*p-value*). The distributions of *Qstat* and *-log*10(*p-value*) over the

selected genes for the 6 different datasets in rice are given in Figures 3.6 and 3.7, respectively.

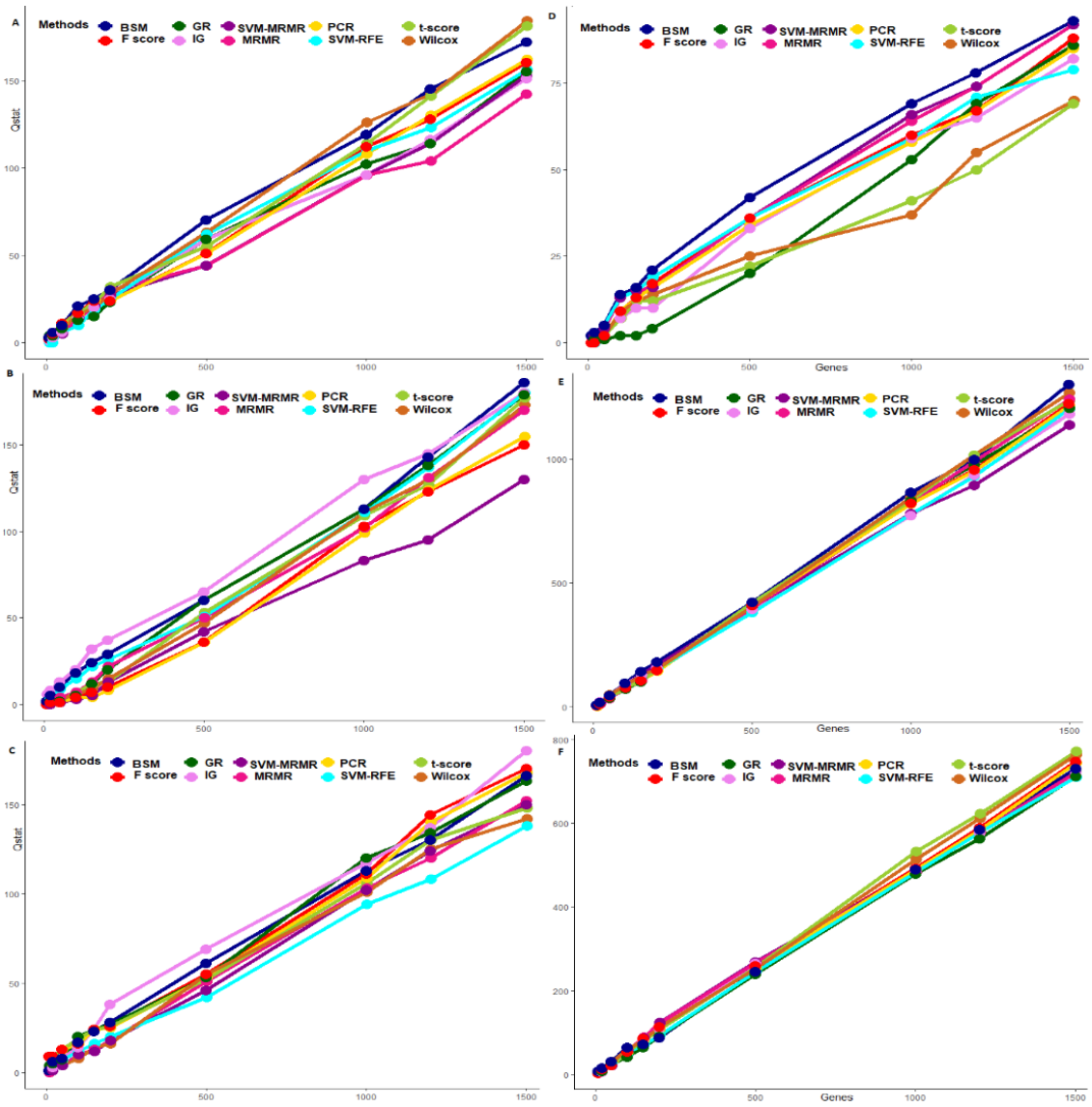


Figure 3.6. Comparative performance analysis of gene selection methods through distribution of *Qstat* statistic. The horizontal axis represents the informative gene sets obtained through gene selection methods. The vertical axis represents the value of *Qstat* statistic. The distribution of *Qstat* statistic are shown for (A) Salinity; (B) Cold; (C) Drought; (D) Bacterial; (E) Fungal, and (F) Insect stress datasets in rice. The lines in different colors represent different gene selection methods.

For salinity stress data, the values of *Qstat* over all the gene sets of sizes (<1000) selected through the proposed BSM approach were found to be higher than those of SVM-MRMR, SVM-RFE, MRMR, IG, F, Wilcox, and PCR (Figure

3.6A). Further, it may be noted that the proposed approach was equally competitive with the univariate gene selection method such as t-test, while they are assessed in terms of *Qstat* (Figure 3.6A). For higher gene set sizes (>1000), the values of *Qstat* for the Wilcox method was found to be higher than those of the proposed, and existing approaches (Figure 3.6A) in the same data. This may be attributed to the fact that the Wilcox method is NP and does not depend on the distribution of GE data.

For cold stress data, the values of *Qstat* statistic for all the selected gene sets through the BSM approach were higher than those of other existing methods (Figure 3.6B). This indicates that the performance of the proposed BSM approach is better in terms selecting cold stress related biologically relevant genes that are mostly overlapped with cold stress QTL regions in rice. Similar interpretations can be made for other abiotic (drought), and biotic (bacterial, fungal and insect) stress datasets in rice (Figure 3.6). Here, it is worthy to note that the *Qstat* is a linear function of the number of genes selected (through a selection approach), number of QTLs reported for that stress, and length of the QTL regions (Figure 3.6).

Further, for salinity stress data, the -log10(*p-value*) from hypergeometric test over all the selected gene sets for the proposed BSM approach was observed to be higher than other existing gene selection methods (except t and GR) (Figure 3.7). In other words, genes selected by the BSM approach were enriched with the underlying salinity responsive QTLs as compared to other existing methods. Similar interpretations can be made for other abiotic (*i.e.,* cold and drought), and biotic (*i.e.,* bacterial, fungal and insect) stress datasets in rice (Figure 3.7).

Moreover, it is interesting to note that the values of *Qstat* and *-log*10(*p-value*) for (univariate) methods, such as t, F, PCR, Wilcox, IG, and GR were found to be higher than those of the existing (multivariate) methods (*i.e*., MRMR, SVM-MRMR) (Figures 3.6, 3.7). This indicates the better, and equally competitive performance of univariate over multivariate methods of gene selection, when evaluated through QTL based biological relevancy criteria. Such observations are not expected in statistics, but well established in biology through previous studies [33].

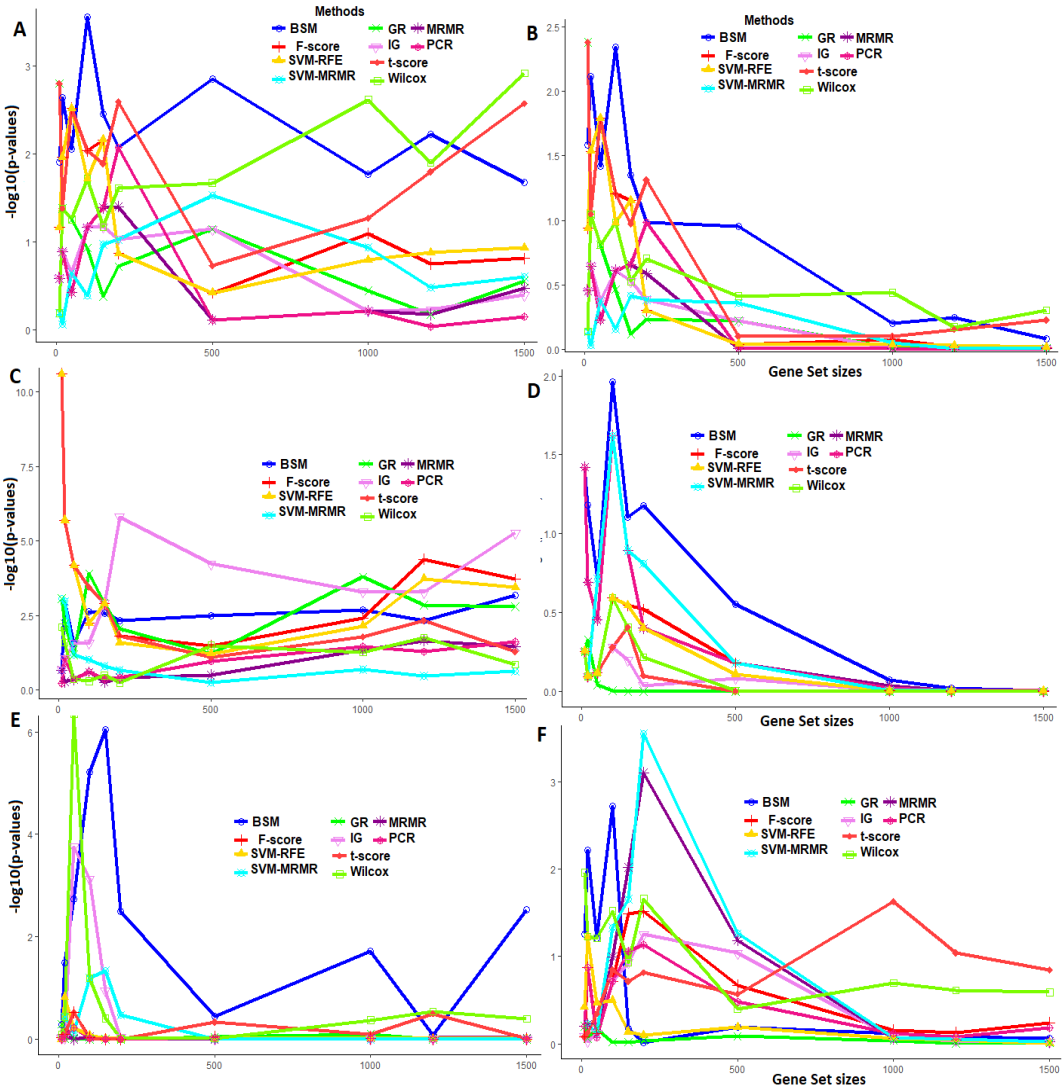


Figure 3.7. Comparative performance analysis of gene selection methods through distribution of *p-values* from QTL-hypergeometric test. The horizontal axis represents the gene sets obtained through gene selection methods. The vertical axis represents the value of

-log10(*p-value*) from QTL-hypergeometric test. The distribution of -log10(*p-value*) are shown for (A) Salinity; (B) Cold; (C) Drought; (D) Bacterial; (E) Fungal, and (F) Insect stress datasets in rice. The lines in different colors represent different gene selection methods.

Judging the performance of gene selection methods based on only classification, might lead to the selection of biologically irrelevant genes. Therefore, we used criteria based on QTLs to test the biological relevancy of the selected genes through proposed, and existing gene selection methods. Through this performance analysis, it was found that BSM approach selects more biological relevant genes measured in terms of overlapping of the selected genes with given QTL regions compared to multivariate approaches, *i.e.* MRMR, SVM-MRMR and machine learning approach such as SVM-RFE. The proposed BSM approach was equally competitive (and better) using the univariate approaches, *i.e.,* t-test, F-score, and Wilcox, and information theoretic measures, *i.e.,* IG and GR, when QTL based criteria were considered. Through the QTLs-hypergeometric test analysis, it was evident that genes selected through the proposed BSM approach were more statistically enriched with the underlying QTL regions.

Comparative performance analysis based on GO analysis

The comparative performance analysis of the proposed BSM approach with 9 competitive gene selection methods was carried out through GO based distance analysis on 6 different rice GE datasets. Here, we set *n* (*i.e.* number of selected genes) as 10, 20, 50, 100, 150, 200 and 500. Then, the selected genes, through the 10 gene selection methods including the proposed BSM, were annotated with the GO terms under MF, BP, and CC categories using *AgriGO* database [24]. The results from this analysis for abiotic stresses under MF, BP, and CC GO categories are given in Tables 3.2-3.4 respectively.

Table 3.2. Comparative Performance analysis of the gene selection methods through GO (MF) based biologically relevant score for abiotic stresses in rice.

Methods	MRMR	SVM	SVM-MRMR	IG	GR	Wilcox	t	PCR	F	BSM
Salt stress in rice										
10	0.98	0.95	0.97	0.92	0.89	0.93	0.93	0.96	0.96	**0.88**
20	0.97	0.89	0.93	0.92	0.86	0.89	0.89	0.91	0.91	**0.86**
50	0.92	0.91	0.92	0.90	0.90	0.87	0.87	0.92	0.92	**0.85**
100	0.92	0.90	0.89	0.90	0.88	0.87	0.88	0.92	0.91	**0.83**
150	0.90	0.89	0.90	0.89	0.88	0.87	0.87	0.90	0.91	**0.83**
200	0.90	0.89	0.88	0.89	0.87	0.88	0.88	0.90	0.90	**0.84**
500	0.90	0.90	0.89	0.90	0.90	0.89	0.90	0.89	0.89	**0.83**
Cold stress in rice										
10	0.82	0.84	0.82	0.92	0.99	0.92	0.87	0.77	0.77	**0.75**
20	0.93	0.88	0.93	0.95	0.93	0.88	0.90	0.91	0.88	**0.71**
50	0.91	0.88	0.91	0.93	0.90	0.91	0.91	0.92	0.92	**0.73**
100	0.91	0.90	0.91	0.90	0.88	0.91	0.91	0.91	0.91	**0.74**
150	0.90	0.89	0.90	0.89	0.89	0.89	0.90	0.91	0.91	**0.72**
200	0.90	0.89	0.90	0.89	0.88	0.89	0.90	0.90	0.90	**0.73**
500	0.90	0.88	0.90	0.90	0.89	0.88	0.89	0.88	0.89	**0.73**
Drought stress in rice										
10	0.82	0.86	0.81	0.90	0.93	0.65	0.76	0.76	0.76	**0.71**
20	0.79	0.86	0.78	0.91	0.90	0.80	0.81	0.81	0.81	**0.75**
50	0.88	0.84	0.87	0.88	0.90	0.84	0.88	0.89	0.89	**0.75**
100	0.89	0.89	0.88	0.89	0.89	0.88	0.88	0.88	0.88	**0.76**
150	0.88	0.88	0.87	0.89	0.88	0.88	0.88	0.88	0.88	**0.76**
200	0.88	0.88	0.87	0.88	0.89	0.89	0.88	0.88	0.88	**0.74**
500	0.88	0.88	0.87	0.88	0.88	0.89	0.88	0.87	0.87	**0.73**

Values marked as bolds represent dissimilarity scores obtained from proposed BSM approach

Table 3.3. Comparative Performance analysis of the gene selection methods through GO (BP) based biologically relevant score for abiotic stresses in rice.

Methods	MRMR	SVM	SVM-MRMR	IG	GR	Wilcox	t	PCR	F	BSM
Salt stress in rice										
10	0.86	0.94	0.86	0.92	0.97	0.90	0.90	0.88	0.88	**0.83**
20	0.90	0.91	0.90	0.89	0.91	0.92	0.92	0.84	0.85	**0.84**
50	0.89	0.90	0.88	0.88	0.90	0.88	0.89	0.88	0.88	**0.82**
100	0.88	0.89	0.86	0.89	0.89	0.85	0.86	0.89	0.87	**0.82**
150	0.87	0.89	0.90	0.88	0.89	0.85	0.85	0.89	0.89	**0.83**
200	0.87	0.89	0.86	0.88	0.89	0.84	0.85	0.89	0.88	**0.82**
500	0.87	0.89	0.87	0.87	0.89	0.86	0.86	0.86	0.86	**0.82**
Cold stress in rice										
10	0.79	0.82	0.79	0.86	0.94	0.91	0.90	0.79	0.79	**0.79**
20	0.93	0.89	0.93	0.90	0.88	0.86	0.88	0.90	0.86	**0.82**
50	0.88	0.89	0.88	0.90	0.88	0.88	0.87	0.89	0.90	**0.71**

100	0.88	0.89	0.88	0.89	0.87	0.90	0.88	0.89	0.89	**0.74**
150	0.89	0.88	0.89	0.88	0.88	0.88	0.87	0.88	0.88	**0.73**
200	0.89	0.87	0.89	0.87	0.87	0.87	0.87	0.88	0.84	**0.73**
500	0.88	0.86	0.88	0.86	0.86	0.84	0.86	0.87	0.83	**0.71**
Drought stress in rice										
10	0.86	0.79	0.85	0.81	0.89	0.62	0.83	0.83	0.83	**0.73**
20	0.84	0.79	0.83	0.89	0.90	0.80	0.84	0.84	0.84	**0.72**
50	0.88	0.81	0.87	0.88	0.88	0.81	0.88	0.88	0.88	**0.72**
100	0.87	0.84	0.86	0.88	0.88	0.84	0.86	0.87	0.87	**0.72**
150	0.86	0.84	0.85	0.88	0.88	0.84	0.87	0.87	0.87	**0.71**
200	0.86	0.84	0.85	0.87	0.87	0.85	0.86	0.86	0.86	**0.72**
500	0.87	0.85	0.86	0.86	0.87	0.87	0.86	0.85	0.83	**0.72**

Values marked as bolds represent dissimilarity scores obtained from proposed BSM approach

Table 3.4. Comparative Performance analysis of the gene selection methods through GO (CC) based biologically relevant score for abiotic stresses in rice.

	MRMR	**SVM**	**SVM-MRMR**	**IG**	**GR**	**Wilcox**	**t**	**PCR**	**F**	**BSM**
Salt stress in rice										
10	0.77	0.71	0.70	0.94	0.97	0.93	0.93	0.95	0.95	**0.78**
20	0.88	0.87	0.85	0.92	0.90	0.91	0.91	0.88	0.88	**0.81**
50	0.88	0.89	0.86	0.92	0.92	0.90	0.90	0.89	0.89	**0.84**
100	0.88	0.90	0.8	0.91	0.89	0.86	0.86	0.88	0.88	**0.83**
150	0.87	0.90	0.87	0.90	0.89	0.86	0.87	0.88	0.88	**0.83**
200	0.87	0.89	0.85	0.90	0.90	0.88	0.89	0.88	0.88	**0.83**
500	0.88	0.90	0.88	0.89	0.90	0.88	0.89	0.87	0.87	**0.82**
Cold stress in rice										
10	0.78	0.80	0.78	0.96	0.81	0.87	0.86	0.70	0.70	**0.70**
20	0.88	0.86	0.88	0.96	0.87	0.87	0.89	0.81	0.83	**0.71**
50	0.86	0.89	0.86	0.90	0.85	0.84	0.85	0.89	0.90	**0.73**
100	0.88	0.90	0.88	0.90	0.81	0.83	0.84	0.87	0.87	**0.74**
150	0.88	0.89	0.88	0.90	0.82	0.82	0.86	0.87	0.88	**0.74**
200	0.87	0.90	0.87	0.90	0.84	0.85	0.86	0.87	0.85	**0.73**
500	0.88	0.89	0.88	0.89	0.86	0.97	0.86	0.88	0.87	**0.73**
Drought stress in rice										
10	0.82	0.86	0.81	0.91	0.89	0.83	0.87	0.87	0.87	**0.74**
20	0.89	0.85	0.88	0.93	0.90	0.87	0.89	0.89	0.89	**0.74**
50	0.86	0.88	0.85	0.91	0.87	0.87	0.88	0.88	0.88	**0.73**
100	0.87	0.87	0.86	0.89	0.86	0.87	0.88	0.88	0.88	**0.74**
150	0.87	0.87	0.86	0.90	0.85	0.85	0.87	0.87	0.87	**0.74**
200	0.87	0.87	0.86	0.89	0.86	0.86	0.87	0.87	0.87	**0.73**
500	0.87	0.86	0.86	0.89	0.87	0.88	0.87	0.86	0.85	**0.72**

Values marked as bolds represent dissimilarity scores obtained from proposed BSM approach

For salinity stress data, under the MF category, the values of GO based average distance scores for the proposed BSM approach were found to be the

least compared to the 9 existing methods over all the selected gene sets (Table 3.2). This indicated that the proposed approach selected more (molecular) functionally similar genes which are responsible salinity tolerance in rice. Similar results can be found for BP, and CC GO based distance analysis for the same stress data (Table 3.2). In other words, the proposed BSM approach selects more biologically relevant genes attributed to each GO category for salinity stress as compared to the other 9 competitive methods (Table 3.2). For bacterial stress, the values of GO based average distance score under MF, BP, and CC GO categories for the proposed BSM approach were found to be least among other gene selection methods. Similar interpretations can be made for other abiotic (*i.e.,* cold and drought) and biotic (*i.e.,* fungal and insect) datasets in rice (Tables 3.2-3.4). Through this analysis, it was found that the proposed BSM approach performed better in terms of selecting more functionally relevant genes, which conferred biotic and abiotic stresses tolerance in rice.

The GO based distance analysis showed that higher functional similarities (which may have biological functions important to stress tolerance) exist among the genes selected by the BSM, as compared to existing methods. The performance of the BSM was found to be better and equally competitive with the univariate approaches, *viz.* t-score, F-score, Wilcox, and correlation-based approach in terms of selecting genes which are biologically relevant (in terms of GO annotations) for the target trait/condition. It is worthy to note that the univariate approaches performed better compared to the multivariate approaches under the biology-based criteria, but the former performed worse than latter under

classification-based criteria. This indicates the real biological complexity for assessing the performance of gene selection approaches on real data. Therefore, we used the comprehensive framework of performance analysis of the gene selection methods under both statistical necessary, and biological relevant criteria. The comparative performance analysis revealed that the proposed BSM approach is better as well as competitive under the classification, QTL and GO based criteria.

Comparative performance analysis based on runtime

The recursive feature elimination algorithms-based gene selection methods such as SVM-RFE, and SVM-MRMR are computationally intensive and time consuming. Thus, we used the runtime criterion to evaluate the performance of gene selection methods. Here, the runtime refers to the amount of computational time required to analyze the GE data through running the codes of the respective methods in R software (*v*. 4.0.1). The detail results from the runtime-based evaluation of gene selection methods is given in Table 3.5.

Table 3.5. Runtime based analysis of gene selection methods in Salinity dataset.

SN.	Methods	Symbol	Tools	Run time	Ranks	Score
01	BSM	BSM	BSM	15 Min	8	0.3
02	SVM (RFE)-MRMR	SVM-MRMR	BSM, e1071	75 Min	10	0.1
03	MRMR	MRMR	BSM	10 Min	7	0.4
04	SVM with Recursive Feature Elimination	SVM-RFE	e1071	70 Min	9	0.2
05	t-score	t	stats	1.5 Min	1	1
06	F-score	F	stats	2 Min	2	0.9
07	Pearson's Linear Correlation	PCR	FSelector	4 Min	4	0.7
08	Information Gain	IG	FSelector	4.5 Min	5	0.6
09	Gain Ratio	GR	FSelector	5 Min	6	0.5
10	Wilcoxon Statistic	Wilcox	stats	2.5 Min	3	0.8

Tool: R packages used for each method

For stress salinity data (with 6636 genes over 45 samples), SVM-RFE and SVM-MRMR took ~75 and 70 Minutes respectively to analyze on a 2-core DELL PC with 8 GB RAM with Intel(R) Core (TM) i3-6100U CPU @ 2.30GHz (Table 3.5). In contrast, the BSM approach took ~15 minutes to analyze the same GE data to obtain biologically informative genes. The BSM method required less computational time than popular methods of gene selection such as SVM-RFE and with much better performance in terms of obtaining biologically informative gene sets. Similar interpretations can be made for the gene selection methods based on the runtime criterion to analyze the remaining 5 datasets.

The BSM approach is based on the NP test statistic(s) and does not depend on the distribution of the GE data, unlike the t-test. Further, the bootstrap procedure in the BSM can minimize the redundancy among genes as well as reduce the spurious association of genes with traits during gene selection. The proposed BSM approach is also less computationally expensive compared to SVM-RFE, and SVM-MRMR and can be implemented on a personal or workstation computer for analyzing large GE datasets The comparative analysis revealed the BSM approach has the features of an ideal technique of gene selection, as it performed better under both statistically necessary, and biologically relevant criteria.

“Data is the new oil of the twenty-first century…”

Clive Humby

CHAPTER 4

STATISTICAL APPROACH FOR GENE SET ANALYSIS WITH TRAIT SPECIFIC QUANTITATIVE TRAIT LOCI

Introduction

The recent advancement in genome sequencing technologies leads to generation of tremendous volume of high-throughput biological data [137]. Meanwhile, exploiting these data and drawing valid biological knowledge has posed a great challenge to scientists across the globe. For instance, in genome wide expression study, the traditional objectives are: (a) obtaining the expression levels of several thousand(s) of genes for the samples belonging to at least two different contrasting phenotypic/ environmental conditions; and (b) identifying the genes which are relevant to these conditions under study among the large pool of genes. Moreover, for the latter objective, several statistical and machine learning approaches have been developed [3,143]. Further, the selected genes are expected to have major causal role for the phenotypic trait under study [2].

The focus in GE data analysis has been shifted from single gene to the gene set level, as a gene does not work alone; rather, it works as an intricate network of a set of genes [8]. Analysis of GE data in terms of gene sets has numerous computational advantages over single gene studies [26]. Keeping in view this fact, a variety of methods for GSA have been developed and used in GE analysis.

The popular GSA methods include GSEA [8,179,180], SAFE [32] and Random set methods [52]. These competitive methods compare the gene set with its complement in terms of association with previous biological knowledge base, *i.e.,* pathways, GO terms, differential expression, *etc.,* under the framework of the statistical hypothesis [5,124].

Along with the development of GSA methods and expression measurement technology, the availability of other biological data such as QTLs is also growing rapidly in public domain databases. QTLs are segments of genomic regions either containing or linked to genes that correlate with variation in a phenotype (quantitative trait) [166]. Moreover, it is a classical and widely used molecular breeding method and can be a potential source for understanding the genotype-phenotype relationships in plant biology. Further, the causal relation between variations in a specific trait and differences in the underlying genotypic level is of paramount importance for understanding genome function and evolution [181], which is the basis for targeted molecular breeding.

Therefore, performing analysis of gene sets based on trait specific QTLs through a computational approach instead of traditional GO or pathways information will be very helpful in unraveling genotype-phenotype relationships. The enrichment analysis of gene sets is well developed in human disease genetics, where, GO terms and known biological pathways are taken into account [8]. These approaches may not be useful to establish any formal relation between genotype and trait specific phenotype in plants. Thus, in plant biology and breeding, analysis of gene sets with trait specific QTLs breeding, analysis of gene

sets with trait specific QTLs requires innovative and advanced statistical techniques.

In this chapter, we propose an innovative statistical approach for analysis of gene sets with trait specific QTLs (GSAQ) under a sound computing framework. Further, its utility was evaluated on five complex abiotic and biotic stresses in rice (*Oryza sativa* L.), as the rice genome is well annotated. The performance analysis of the GSAQ approach indicated its effectiveness and efficiency in performing the trait specific enrichment analysis of gene sets through incorporating background QTL information. This proposed technique integrated the GE data with QTL data to provide effective gene sets enriched with the QTL information. Further, we also illustrated the application of the developed GSAQ approach as a biological relevant criterion to evaluate the performance of gene selection methods based on high dimensional GE data. For this purpose, we used ten different gene selection methods, *viz.* SVM-RFE [142,157], t-score[6,148], F-score[3], MRMR [149,165], Random Forest (RF) [144], IG [143,153], GR [143,153], Symmetrical Uncertainty (SU) [143,153], PCR [143,153], and Spearman's Rank Correlation (SRC) [143,153]. Our results showed that the GSAQ approach provided two biologically relevant criteria for evaluating the performance of gene selection methods on GE data.

Material and Methods

The performance analysis of the proposed GSAQ approach was carried out on rice, as it is a model crop plant and a huge amount of GE and QTL datasets are publicly available. Therefore, five different GE datasets related to two biotic

stresses (blast (fungal) and brown plant hopper (insect)) and three abiotic stresses (salinity, cold and drought) for rice were collected. These GE datasets were obtained from GEO database of NCBI (http://www.ncbi.nlm.nih.gov) with platform GPL2025, as this platform contains 191 GE experiments (series) comprising 3096 samples/subjects of rice. Among these samples, 304 experimental samples related to these biotic and abiotic stresses for rice were taken in this study through performing meta-analysis individually for each of the stresses. The summary and detail descriptions of the GE datasets are given in Table 4.1. Further, the trait specific QTLs for the stresses, *viz*. fungal, insect, salinity, drought and cold (for each of the GE datasets) for rice were collected from the Gramene QTL database (http://www.gramene.org/qtl) [169].

Table 4.1. Summary of datasets used in this study.

SN	Descriptions	#Series	#Genes	#Sample	#Class	#QTL	# UQTL
A	Salinity Stress	6	6637	70	2	17	13
B	Cold Stress	5	8840	100	2	37	21
C	Drought Stress	5	9078	90	2	77	20
D	Blast Stress	2	7071	26	2	183	77
E	Brown Plant Hopper	1	7240	18	2	93	57

#Series: Number of GEO series for each dataset; #Genes: Number of genes; #Sample: Number of GEO samples; #S: Number of GE samples belonging to 2 classes (control vs. stress); #QTL: Number of QTLs found for each stress; #UQTL: Number of unique QTLs found in rice for each stress.

Data preprocessing

The preprocessing of the data is intended to remove noises, including missing probes and mislabeled probes [2]. For this purpose, the analysis was conducted by using Bioconductor platform of R. Initially, the raw CEL files of the collected samples were processed using the RMA algorithm available in the *affy* Bioconductor package of R [168,182]. This includes background correction, quantile normalization and summarization by the median polish approach [183].

Further, the log2 scale transformed expression data from RMA for the selected experimental samples were used for further selection of relevant gene sets.

Preliminary gene selection for dimension reduction

For tens of thousands of genes in GE data, it would be of high computational complexity to use the gene set selection methods directly [3]. Hence, we first employed t-test and FC criteria to filter out unlikely genes to reduce the dimension of the GE datasets. In our preliminary selection, we assigned 1 and 0.05 as the |FC| and *p-value* thresholds respectively, resulting in selection of several thousands of genes. Further, GE data on these selected genes (Table 4.1) (at the preliminary stage) were further used for final gene set selection using different gene set selection methods.

Selection of gene sets

Among the thousand(s) of genes in GE datasets, it is challenging from a systems biology point of view to choose those genes that are most relevant to the specific trait [143]. Here, we have taken eight statistical methods, *viz*. t-score, F-score, MRMR, IG, GR, SU, PCF, SRC and two machine-learning methods, *viz*. RF and SVM-RFE to select relevant gene sets. These ten gene selection methods were applied to high dimensional GE datasets related to five different stresses for selection of pertinent gene sets of rice. For all gene selection methods, the gene lists were prepared by arranging the genes based on the descending order of the respective computed metrics. The gene sets of different sizes were selected from the prepared gene lists through each gene selection method for each stress.

Proposed approach for Gene Set Analysis with QTL (GSAQ)

Let **Ω** be the whole gene space (set of genes in a genome), *G* be a selected gene set obtained by using a gene selection method for a particular condition/ trait, G' (*i.e.,* **Ω -** *G*) be the set of not selected genes *i.e.,* complement of *G*. Let, *N* and *n* be the number of elements in **Ω** and *G,* respectively. Let *Q* be the set of associated QTLs for the same trait. Suppose for a member gene (i^{th} gene) in *G*, *i.e.,* g_i^c [*a*, *b*] ϵ *G*, where *a* and *b* represent start and stop positions (in terms of base pairs) of the gene g_i in chromosome *c*. Similarly, a member QTL (t^{th} QTL) in *Q*, *i.e.,* q_t^c[*d*, *e*] ϵ *Q*, where, *d* and *e* represent the start and stop positions of the QTL q_t on chromosome *c*. The complete overlap of the genomic regions of the gene g_i^c with that of a QTL q_t^c can be expressed by using an indicator function, which is shown as:

$$I_{q_t}(g_i) = \begin{cases} 1 & if\ g_i^c[a,b] \in q_t^c[d,e] \\ 0 & if\ g_i^c[a,b] \notin q_t^c[d,e] \end{cases} \quad (4.1)$$

In other words, the selected gene would have a QTL hit, if its genomic regions completely overlapped with that of a QTL for a particular trait (both belong to the same chromosome). Further, the total number of genes in *G* overlapped with QTL regions can be defined by a statistic called as total number of QTL hits (*NQhits*) in *G* and given as:

$$NQhits = \sum_{t=1}^{|Q|} \sum_{i=1}^{n} I_{q_t}(g_i) \quad (4.2)$$

In addition, the proportions of genes those got QTL hits (Pr_{GQ}) in *G* can also be computed through Eq. 4.3.

$$Pr_{GQ} = \frac{NQhits}{n} \quad (4.3)$$

Similarly, proportions of genes with QTL hits in G' ($Pr_{G'Q}$) can be expressed as:

$$Pr_{G'Q} = \frac{NQhits'}{N-n} \quad (4.4)$$

where, $NQhits'$ is the total number of QTL hits in G'.

The expressions in Eq. 4.1 and 4.2 can be used to show whether a gene had a QTL hit or not, and to compute the *NQhits* statistic for all genes in *G* respectively. The developed statistic may not be sufficient to evaluate the statistical significance of selected gene set related with the specific trait. To this end, Wang *et al*. (2013) proposed the Gene Set Validation with QTLs (GSVQ) (or Microarray-QTL) test using Hypergeometric distribution to validate the selected salinity responsive genes in rice with salinity QTLs [2]. However, the GSVQ test is unique, but it is not statistically sound as it violates the basic assumptions of Hypergeometric distribution (*i.e.* sampling without replacement) and fails to state the underlying null hypothesis.

Therefore, to perform the gene set analysis with the underlying trait specific QTLs under a sound computing framework, we developed the GSAQ approach. This approach can be used to evaluate the statistical significance of selected gene sets related to specific trait based on available QTL information. Under this approach, the following hypothesises can be constructed for testing purpose.

H_0: Genes in *G* are at most as often overlapped with the QTL regions as the genes in $G'(Pr_{GQ} = 0)$

H_1: Genes in *G* are more often overlapped with the QTL regions as compared to genes in G' $(Pr_{G'Q} > 0)$

In other words, the above constructed null hypothesis is a competitive one as it considers the genes from both *G* and G' [5].

The proposed GSAQ approach is based on formation of 2×2 contingency tables and a Hypergeometric distribution. Further, the 2×2 tables have been extensively used in differential expression analysis, GO and pathways enrichment analysis [5,28,32]. The basic concept behind this 2×2 table method is a gene sampling model. Moreover, each cell of such table is filled with a sample of genes, each of which is drawn at random from the gene space. Here, in this sampling model, each sampling unit (*i.e.,* gene) can be subjected to two fixed set of indicator measurements, *i.e.,* (*A*, *B*), where: (i) *A* (1 or 0) indicates whether the gene is a part of the selected gene set or not; and (ii) *B* (1 or 0) indicates whether that gene had the QTL hit or not. Further, the gene space can be formalized into a population having *N* units (for *N* genes) and shown as:

$$(A_1, B_1), (A_2, B_2), \ldots, (A_i, B_i), \ldots, (A_N, B_N) \quad (4.5)$$

where, i^{th} unit in Eq. 4.5, *i.e.,* (A_i, B_i), shows that whether i^{th} gene is a part of the gene set or not and whether it also got a QTL hit or not.

Here, the gene sampling model (where genes are taken as sampling units) is quite different from the usual classical subject sampling model (where the GE profiles are considered as sampling units) [5]. Through this gene sampling model (by fixing A_i=1), *K* gene samples, *i.e.* G_1, G_2, …, G_K, each of size *m* (≤ *n*) are randomly drawn from the population with equal probability by using simple random sampling without replacement procedure. For each G_k (*k*=1, 2, …, *K*), a 2×2 table, as shown in Table 4.2, was constructed. Similarly, using this procedure, *K*, 2×2

contingency tables were obtained for *K* gene samples. The *NQhits* statistic computed through Eq. 4.2 from 2×2 table (Table 4.2) constructed for k^{th} gene sample follows a Hypergeometric distribution [2] and given in Eq. 4.6.

$$P[X = N_{G_kQ}] = \frac{\binom{N_Q}{N_{G_kQ}}\binom{N - N_Q}{m - N_{G_kQ}}}{\binom{N}{m}} \tag{4.6}$$

where, *X* is a random variable representing the value of *NQhits* (N_{G_kQ}) for *k-th* gene sample (*k*=1, 2, …, *K*), N_Q is total number of QTL hits in Ω and *m* is the size of k^{th} gene sample.

Table 4.2. 2 × 2 contingency table for gene set enrichment test with QTL.

	Overlapped with QTL regions	**Not overlapped with QTL regions**	**Total**
Selected gene set	$n_{G^{(k)}Q}$	$n_{G^{(k)}Q^C}$	$n_{G^{(k)}}$
Not selected gene set	$n_{G^{(k)C}Q}$	$n_{G^{(k)c}Q^c}$	$n_{G^{(k)C}}$
Total	n_Q	n_{Q^C}	N

Ω: gene space; $n_{G^{(k)}}$: number of genes in $G^{(k)}$; $n_{G^{(k)C}}$: number of genes in $G^{(k)C}$; n_Q: number of QTL hit genes in gene space: n_{Q^C}: number of non-QTL hit genes in gene space. $G^{(k)}$: *k*-th gene sample from selected gene set; $G^{(k)C}$: **Ω** -$G^{(k)}$

Through the Hypergeometric distribution (Eq. 4.6), the statistical significance value or *p-value* (p_k) for *k-th* gene sample can be computed by using Eq. 4.7.

$$p_k = P[X_k \geq x \mid H_0] = 1 - P[X_k \leq x \mid H_0] \tag{4.7}$$

For assessing the final statistical significance of the test, the individual *p-values* needs to be combined.

Methods for combining p-values

Suppose there are *K* independent tests (for *K* random gene samples) and their associated *p-values* are p_1, p_2, …,p_K. Under H_0, the *p-values* from individual gene samples are uniformly distributed between 0 and 1 (*i.e.,* p_k ~ U [0, 1]) [184]. To

obtain the overall statistical significance value for the test (H_0 *vs.* H_1), the individual *p-values* for each gene samples can be combined. For this purpose, the methods described in Table 4.2 can be used.

Table 4.2. List of methods used for combining *p-values* to assess final enrichment significance.

Methods	Transformed variable	Test statistic	Dist. under H_0	Reference
Inverse Normal	$Z_k = \Phi^{-1}(p_k)$	$T = \sum_{k=1}^{K} w_k Z_k$	N (0, 1)	[185]
Meanp	$\bar{p} = \sum_{k=1}^{K} p_k \Big/ K$	$W = (0.5 - \bar{p})\sqrt{12K}$	N (0, 1)	[186]
Inverse Chi-Square	$Z_k = -2\log p_k$	$L = 2\sum_{k=1}^{K} Z_k$	$\chi^2_{2k\ df}$	[186,187]
Logit	$S_k = \log\left[p_k/(1-p_{k)}\right]$	$S = \sum_{k=1}^{K} S_k$	$t_{5K+4\ df}$	[188]

p_k: statistical significance value of *k-th* gene sample; Φ, Φ^{-1}: standard normal cumulative distribution function and its inverse respectively; *K*: Number of random gene samples; df: degrees of freedom; H_0: Competitive null hypothesis; *N*(): Normal distribution; t: Central t-distribution; χ^2: Central Chi-square-distribution.

Using the above approach, the final statistical significance values (*p-values*) and FDR values for the selected gene sets were computed. In this case, the gene sets were selected using ten existing gene selected methods. For the computation of *p-values,* we took different combinations of *m* and *K* for selected gene sets. The performance analysis of the proposed GSAQ approach and gene selection methods was carried out on complex abiotic and biotic stresses, *viz.* salinity, cold, drought, fungal and insect, in rice. Moreover, for the computation of FDR for each selected gene set, we executed the *fdrtool* function implemented in *fdrtool* R package [189] which is based on the approach developed by Strimmer (2008) [190]. The operational procedure of the GSAQ approach and its implemented algorithm are shown in Figure 4.1.

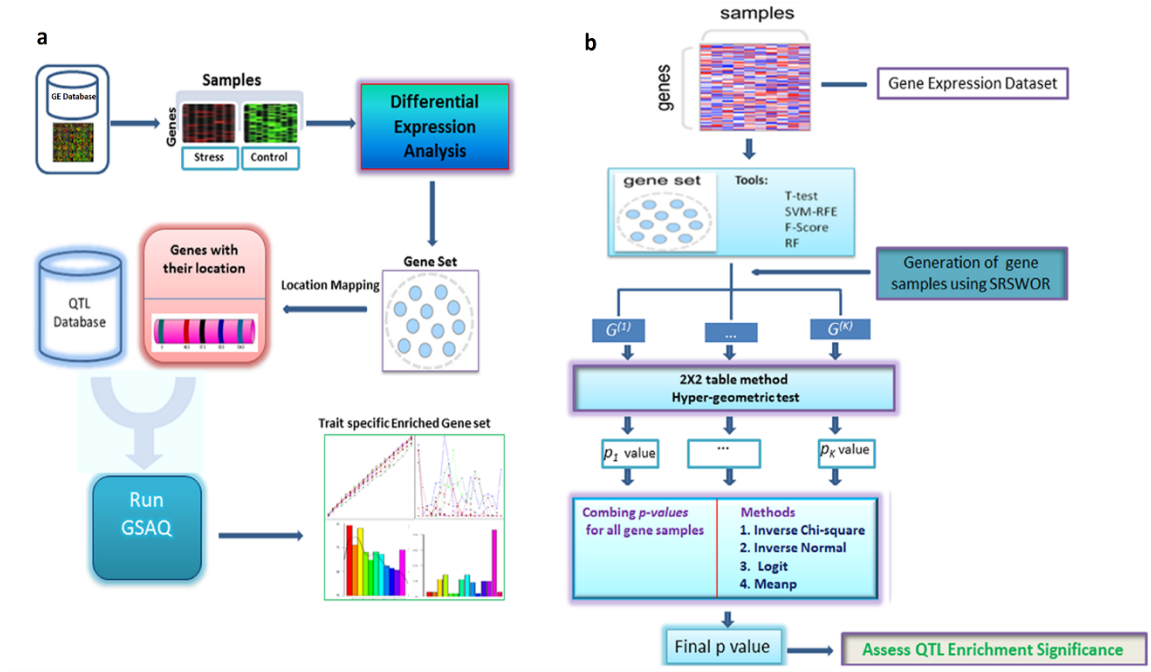


Figure 4.1. Operational procedure and algorithm of GSAQ approach. (a) Operational procedures involved in GSAQ are shown in pictorial form. (b) Flowchart of the computational algorithm implemented in GSAQ approach. $G^{(k)}$'s represents random gene samples and p_k-values represent corresponding statistical significance for each $G^{(k)}$. SRSWOR represents simple random sampling without replacement.

Results

Selection of gene sets

Using high dimensional GE datasets pertaining to various biotic and abiotic stresses, we selected different gene sets of sizes, *viz*. 100, 200, 300, …, 2000 through a two-stage approach of preliminary gene selection and ten different gene selection methods, which are relevant to individual traits/stresses in rice. Further, we mapped the QTLs and genes in each selected gene set (for each gene selection methods) in the whole genome using MSU rice genome browser [178].

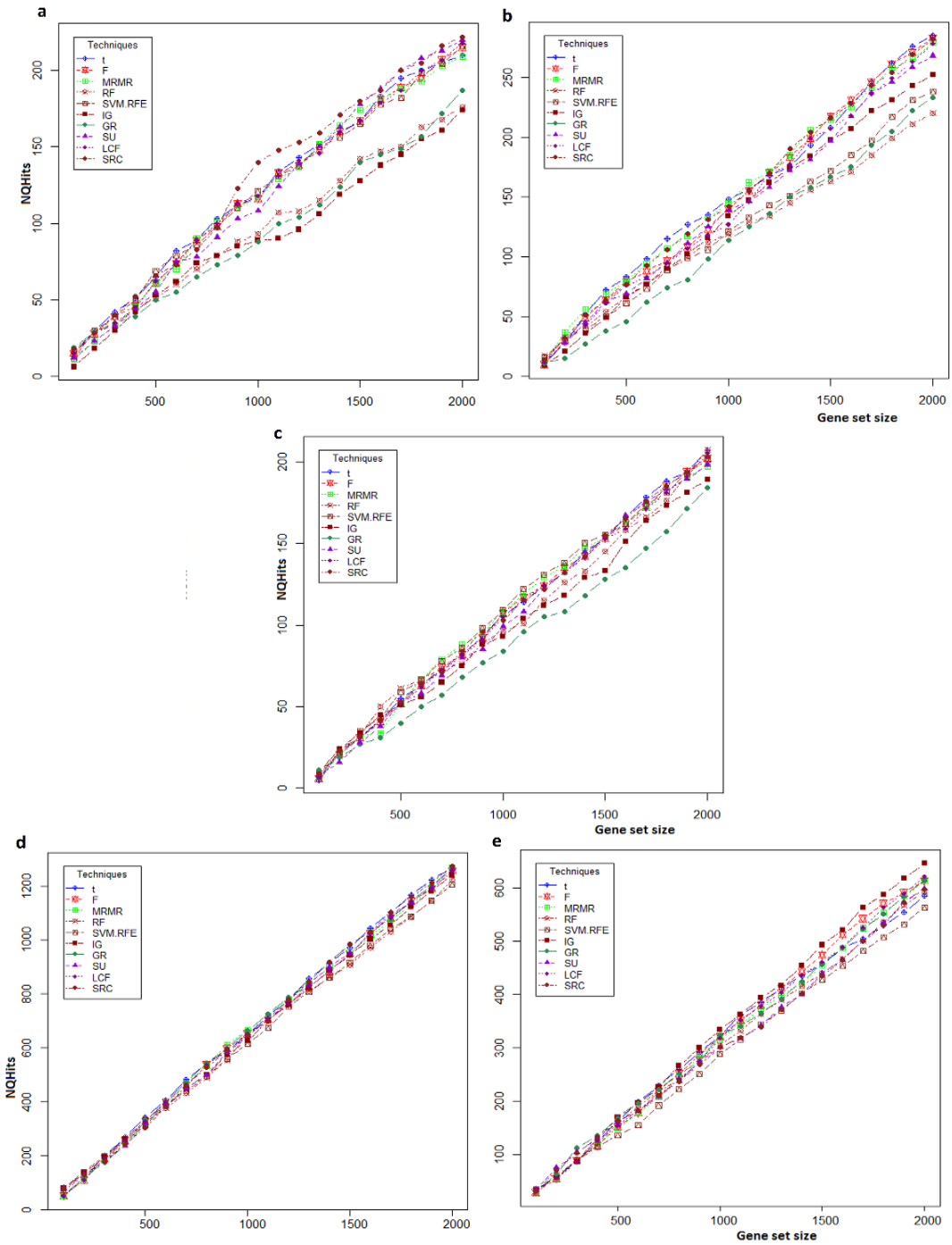


Figure 4.2. Distribution of *NQhits* statistic over the selected gene sets. The horizontal axis represents the gene sets obtained by each of the ten gene selection methods. The vertical axis represents *NQhits* statistic obtained through GSAQ approach. Distribution of *NQhits* are shown for (a) salinity, (b) cold, (c) drought, (d) fungal and (e) insect stress datasets in rice.

Distribution of NQhits statistic

The distributions of *NQhits* statistic over gene sets obtained by ten different gene selection methods for each of these five stresses are shown in Figure 4.2. For all these stresses, the value of *NQhits* statistic is found to be directly proportional to size of the gene set (Figure 4.2). In other words, the value of the *NQhits* statistic depends on the factors such as length of QTLs and number of genes in a gene set linked to QTLs for a given stress. This observation is true for all gene selection methods irrespective of stresses. Moreover, the developed *NQhits* statistic can also be used as a metric for evaluating the performance of gene selection methods. The performance of different gene selection methods based on *NQhits* statistic for the abiotic stresses, *viz.* salinity, cold and drought are at par for selection of smaller gene sets, as the value of *NQhits* for relatively smaller gene set sizes (*e.g.* 100-500) are almost equal (Figure 4.2). But, in the case of larger gene sets, the performance based on *NQhits* statistic is found to be better for t-score, F-score, MRMR, SU, LCF, SRC and SVM-RFE as compared to IG, GR and RF. However, for the biotic stresses (fungal and insect) most of the gene selection methods performed equally well over various gene sets in terms of *NQhits* statistic (Figure 4.2). This variation in performance of gene selection methods under abiotic stresses may be due to the complex/polygenic nature of abiotic stresses (due to non-living climatic factors) as compared to biotic stresses (living factors).

Gene sets analysis with QTLs

Although the *NQhits* statistic can be used as a performance evaluation metric but, it failed to tell the trait-specific enrichment of gene sets or association of genotype-

phenotype relation. Therefore, we proposed the GSAQ approach to test the trait specific enrichments of the gene sets with underlying QTLs. For this purpose, gene sets were selected from the high dimensional GE data by using ten different methods. Further, we explored the ability of the proposed GSAQ approach along with existing GSVQ approach to provide biologically meaningful insights (*e.g.* establishing genotype-trait specific phenotype associations) in five complex abiotic and biotic stresses in rice. For both the approaches, we searched significantly associated gene sets enriched with underlying QTLs, which were selected by a particular gene selection method in each of the stresses.

The distribution of *p-values* computed from both existing GSVQ and proposed GSAQ approaches are shown in Figures 4.3, 4.4. For salinity stress, the distribution of *p-values* computed from GSAQ using Inverse normal method for all gene sets (through all gene selection methods) are shown in Figure 4.3A1. It has been observed that except IG, GR and RF, all gene selection methods provided gene sets which were highly statistically significant at 0.001% level of significance (as *p-values* < 10E-5) (Figure 4.3A1). These findings clearly indicate that the gene sets obtained by most of the methods are enriched with underlying trait specific QTLs through our GSAQ approach. Similar interpretations can be made for all other methods as given in Table 4.2 for GSAQ test for same stress considered in this study.

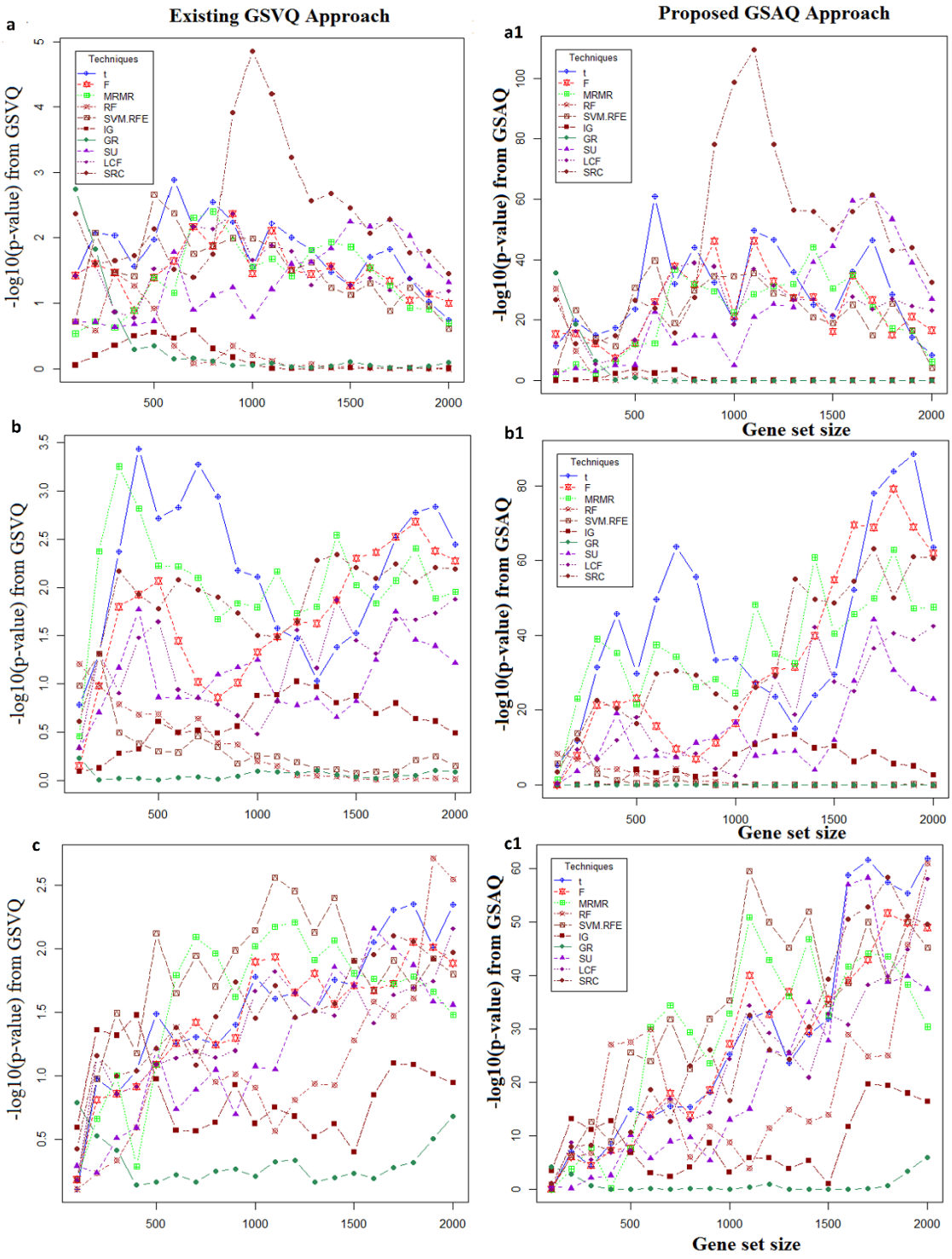


Figure 4.3. Performance analysis of GSAQ approach with GSVQ for abiotic stresses. The horizontal axis represents the gene sets obtained by each of the ten gene selection methods. The vertical axis shows the *negative logarithm of statistical significance values* computed from existing GSVQ approach for (a) salinity, (b) drought, (c) cold stresses and proposed GSAQ approach (with Inverse normal method) for (a1) salinity, (b1) drought, (c1) cold stresses.

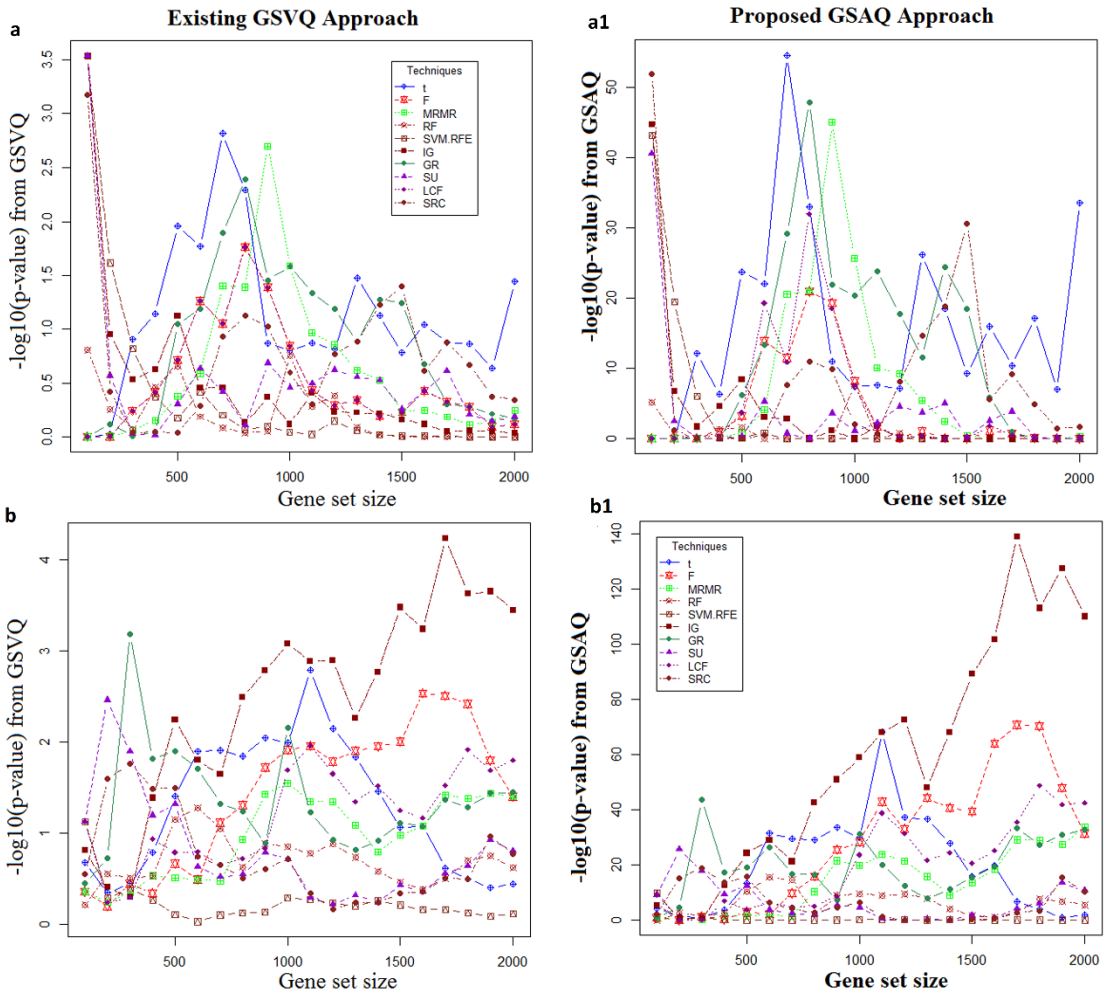


Figure 4.4. Performance analysis of GSAQ approach with GSVQ for biotic stresses. The horizontal axis represents the gene sets obtained by each of the ten gene selection methods. The vertical axis shows the *negative logarithm of statistical significance values* computed from existing GSVQ approach for (a) fungal, (b) insect stresses and proposed GSAQ approach (with Inverse normal method) for (a1) fungal, (b1) insect stresses in rice.

On the contrary, when the existing GSVQ approach was used for performing salinity trait specific enrichment analysis, none of the gene sets selected by any method (except gene sets of sizes 900-1200 from SRC) was found to be significant at the same level of significance (Figure 4.3A). Such findings may not be valid as per our expectation, as these gene sets are selected by the most

powerful contemporary gene selection methods *like* SVM-RFE, RF, GR, SU, t-score, *etc.*

Further, the magnitude of –log10 (*p-values*) from GSAQ enrichment analysis for salinity stress (Figure 4.3A1) is found to be much higher than that of existing GSVQ test (Figure 4.3A). In other words, GSAQ approach more often rejects H_0 (*i.e.* equal salinity QTL enrichment of both selected and not selected gene sets) as compared to GSVQ test. Therefore, it is found that the salinity trait-specific gene set enrichment analysis was better through GSAQ as compared to GSVQ. In order to cross validate these findings on the same datasets related to salinity stress, we computed FDR for both GSAQ and GSVQ for all gene sets. The results are given in Table 4.3. It has been observed that the FDR values from the proposed GSAQ approach for all these gene sets irrespective of gene selection methods are far below than that of existing GSVQ test (Table 4.3). Therefore, it can be concluded that the proposed GSAQ is more efficient than the GSVQ for performing gene set enrichment testing with salinity trait specific QTLs.

Table 4.3. Performance analysis of GSAQ and GSVQ approaches.

Methods	100	200	300	400	500	1000	2000
t	< 0.5 (> 0.5)	< 0.01 (> 0.01)	< 0.001 (> 0.01)	< 0.001 (> 0.01)	< 0.001 (> 0.01)	< 0.0001 (> 0.1)	< 0.0001 (> 0.01)
F	< 0.5 (> 0.5)	< 0.01 (> 0.01)	< 0.001 (> 0.01)	< 0.001 (> 0.01)	< 0.001 (> 0.01)	< 0.0001 (> 0.01)	< 0.0001 (> 0.01)
MRMR	< 0.01 (> 0.1)	< 0.01 (> 0.1)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.0001 (> 0.01)	< 0.0001 (> 0.1)	< 0.0001 (> 0.01)
SU	< 0.1 (> 0.1)	< 0.1 (> 0.1)	< 0.0001 (> 0.1)	< 0.0001 (> 0.2)	< 0.0001 (> 0.01)	< 0.0001 (> 0.01)	< 0.0001 (> 0.5)
PCF	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.0001 (> 0.5)	< 0.0001 (> 0.5)
SRC	< 0.01 (> 0. 01)	< 0.01 (> 0.1)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.0001 (> 0.001)	< 0.0001 (> 0.001)
SVM	< 0.01 (> 0.1)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.01 (> 0.01)	< 0.0001 (> 0.1)	< 0.0001 (> 0.1)	< 0.01 (> 0.1)

FDR: False discovery rate; Gene sets: gene sets obtained from each method; (.): the values in parentheses indicate the FDR value computed through GSVQ approach; t: t-score; F: F-score; MRMR: Maximum Relevance Minimum Redundancy; SU: Symmetrical Uncertainty; PCF: Pearson's Correlation Filter; SRC: Spearman's Rank Correlation filter; SVM: Support Vector Machine with recursive Feature Elimination.

For drought and cold stresses, none of the gene sets selected by any of the ten gene selection methods considered in this study, was found to be enriched with the respective stress specific QTLs, when enrichment analysis was performed through the GSVQ approach (Figures 4.3B and 4.3C). However, all selected gene sets (for drought and cold stresses), irrespective of the gene selection methods (except GR), were found to be more enriched with underlying QTLs through the proposed GSAQ approach using Inverse normal method (Figures 4.3B1 and 4.3C1). Further, the –log10 (*p-values*) computed through GSAQ approach (Figures 4.3B1 and 4.3C1) were found to be much higher than those of the GSVQ test for drought and cold stresses (Figures 4.3B and 4.3C). Subsequently, it was also verified that the FDR values for all the gene sets from the GSAQ approach was found to be less than those from the GSVQ for these stresses (Table 4.3). Similar interpretations can be made from the results obtained for other methods used in the GSAQ approach.

Therefore, it can also be concluded that as with salinity stress, the proposed GSAQ approach was found to be better and more efficient than GSVQ for performing QTLs-specific gene set enrichment testing for drought and cold stress. Further, similar interpretations for the GSVQ and GSAQ approaches can be made for the biotic stresses (insect and fungal) (Figure 4.4). Therefore, we observed that the GSAQ approach performs QTL enrichment analysis of gene sets more successfully and efficiently as compared to existing GSVQ test when there is sufficient background QTL information available. Our analysis showed that we find

much greater consistency in QTL specific gene set enrichment analysis across five different stress scenarios, *viz.* salinity, cold, drought, fungal and insect, by using GSAQ than GSVQ (Figures 4.3, 4.4).

Performance analysis of gene set selection methods based on GSAQ

Apart from assessing the significance of the genotype (gene set) to phenotype (trait) enrichment test, GSAQ can also be used as a performance evaluation metric of gene selection methods for high dimensional GE data. For instance, in salinity stress, 7 different methods, *viz*. t-score, F-score, MRMR, SU, LCF, SRC and SVM-RFE, performed equally well in terms of statistical significance of the GSAQ enrichment testing using Inverse normal method (Figure 4.3A1). For other methods, such as RF and GR, the gene sets of sizes 100-300 are more statistically enriched through the GSAQ approach as compared to larger gene sets. However, all gene sets selected by IG are not enriched with the underlying salinity QTLs (Figure 4.3A1). It can be noted that simple univariate gene selection methods, *i.e.,* t-score and F-score, are equally competitive with multivariate (MRMR) and machine learning approaches such as SVM-RFE and RF for providing salinity trait enriched gene sets (Figure 4.3A1).

Further, the –log10 (*p-values*) from the GSAQ approach for SRC was found to be greater than those of other methods followed by t-score, F-score, MRMR and SVM-RFE. This indicates that gene sets selected by SRC are much more enriched with background salinity QTLs. The superiority of SRC in terms of performance may be expected due to its NP nature. Further for SRC, the –log10 (*p-value*) for the gene set of size 1200 is quite a bit higher than those of other gene sets (Figure

4.3A1), which indicates the maximal enrichment of the same gene set with QTLs. Similar interpretations can be made for other abiotic and biotic stresses (Figures 4.3, 4.4). Similar interpretations can be made for other methods used in GSAQ approach.

Chromosome and QTL-wise distributions of genes

Along with the trait-specific enrichment analysis of gene sets, the proposed GSAQ approach can also be used to get the chromosome- and QTL-wise distributions of genes in selected gene sets. For instance, chromosomal distributions of genes in the gene set of size 1000 across all the five-different abiotic and biotic stresses can be shown using GSAQ. For salinity stress, majority of these salinity responsive genes selected by any gene selection method belong to chromosome numbers 2, 3, 4, 5 and 12. Similar interpretations can also be made for cold, drought, fungal and insect stresses.

The proposed GSAQ approach was able to identify and prioritize QTL candidate genes (*i.e.,* genes having QTL hits) from the selected gene set. In case of salinity stress, most of the QTL candidate genes selected by t-score belong to 8 different QTLs from 13 unique QTLs. For other gene selection methods, the majority of the QTL candidate genes overlapped within the 7 salinity responsive QTLs. Further, it has been found that the QTL with id AQEM001 has the largest number of salinity QTL candidate genes followed by AQEM007 and AQEM009 and this trend is true irrespective of gene selection method used. Similar interpretations can be made for cold, drought, fungal and insect stresses.

Discussion

Traditional strategies for single gene analysis involve expression analysis of a single gene and is mainly focused on identifying individual genes that exhibit differences between two contrasting traits of interest. Although they are useful, but they fail to consider the underlying trait-specific enrichment of the genes that are distributed across an entire network of genes in the selected gene set [8]. The existing GSA approaches mostly focused on whether the selected gene sets are over-represented by differentially expressed genes, known pathways or GO terms through over representation analysis [13,31,38,124]. However, in plant biology, QTLs are considered as a great source of information for conducting an effective breeding experiment, as most of the traits are quantitative in nature and controlled by polygenes. Therefore, we proposed the GSAQ approach as an innovative and efficient way to conduct enrichment analysis of gene sets with trait specific QTLs.

The proposed GSAQ approach is a new way to perform the enrichment analysis of gene sets to establish genotype (polygenes)-phenotype (quantitative trait) association testing with the help of genetically rich trait specific loci data. Further, it is more biologically appealing to establish the association of genes (genotype) in the selected gene set with underlying QTLs (traits/phenotypes). However, in the existing GSVQ approach, the genes are taken as input to the Hypergeometric distribution for performing trait enrichment analysis [2]. This approach violates the basic assumptions (*i.e.,* sampling units must be drawn without replacement) of this distribution and expected to have poor performance in terms of gene set enrichment. Further, it also fails to state the underlying null

hypothesis on which the test is based. Hence, the proposed GSAQ approach is found to be more successful and effective to detect trait specific QTLs enriched gene sets as it is based on statistically strong null hypothesis.

Further, the proposed GSAQ approach is based on testing a competitive null hypothesis using resampling procedure for possible rejection of competitive H_0. In this approach, H_0 was tested against H_1 with the help of the 2×2 table method and gene sampling model. This allows one to statistically test the selected gene set for enrichment with the underlying QTLs (*i.e.,* rejection of null hypothesis of random association of selected genes with QTLs). Further, a *p-value* was computed for a selected gene set, which is more scientific and statistically meaningful to the genome researchers and experimental biologists (as value lies between 0 and 1). The gene sets with lower *p-values* are considered as more enriched with the underlying trait specific QTLs and *vice-versa*. The comparative analysis has shown that the proposed GSAQ approach performs better than existing GSVQ technique for trait specific gene sets enrichment testing. Further, GSAQ approach is more statistically sound, as it satisfied the underlying assumptions of the Hypergeometric distribution and 2×2 contingency tables. Moreover, the developed GSAQ R package is also flexible in detecting QTL enriched gene sets, as four statistically strong options are available to obtain the *p-values* for selected gene sets.

We also demonstrated the performance of the proposed GSAQ approach for performing QTL enrichment test for the selected gene sets on real crop data sets subjected to various complex abiotic and biotic stresses. There are both

challenges and advantages in analyzing these crop datasets. For crops, there are typically limited experimental data available and relatively little literature is available for guidance [2]. The application of GSAQ on complex abiotic and biotic stress scenarios indicated that, it consistently and successfully detects the QTLs enriched gene sets as compared to the existing approach, when the background QTL data is well defined and sufficiently available. It may be noted that the proposed GSAQ approach is a two-stage approach. First, it deals with the selection of gene sets from large gene space by using gene selection methods. Second, it assesses the QTL enrichment significance of gene sets by using the resampling procedure under a gene sampling model and thus provides a suitable statistical framework for testing competitive null hypothesis.

Further, the GSAQ approach has several advantages when compared with single gene-QTL analysis. First, it eases the interpretation of a large-scale experiment by identifying trait-specific enriched gene sets. Therefore, rather than focusing on individual QTL hit genes, researchers can focus on gene sets (polygenes), which tend to be more reproducible and more interpretable (for real world applicability). Further, the multiple testing of hypothesis problem is well tackled in the proposed approach, as it takes the gene set as a functional unit for enrichment analysis. Second, GSAQ is statistically sound, as it is based on a competitive null hypothesis and gene sampling model. It considers the genes present in both selected as well as not selected gene sets, while performing trait specific enrichment analysis. Third, the GSAQ approach helps in prioritizing QTL candidate genes or QTL enriched gene sets under a sound computational

arrangement, which would be very helpful in unraveling genotype-to-phenotype relationships. Gene set enrichment testing is well developed in human genetics, where known biological pathways or ontology are considered. However, in plant biology and breeding, QTL candidate genes or trait specific enriched gene sets identified through this proposed GSAQ technique will be more effective for developing specific trait or stress tolerant crop cultivars. Fourth, the *NQhits* statistic and statistical significance values computed through the GSAQ approach may be considered as biologically relevant criteria for performance analysis of gene selection methods. Previously, subject classification accuracy was a widely used criterion for performance evaluation of gene selection methods [6,18,141,159,163,165,191]. This may be a statistically necessary but may not be a biologically sufficient criterion. Therefore, the proposed GSAQ approach provided two excellent biological relevant criteria for evaluation of gene selection methods under a strong statistical framework. GSAQ approach provides a valuable platform for integrating the GE data with genetically rich QTL data to identify potential QTL enriched gene sets or set of QTL candidate genes, which may act as valuable input or hypothesis for the plant breeders for designing breeding experiments. In this chapter, we have statistically established the credibility of the proposed GSAQ by comparing its performance with the GSVQ on multiple datasets in rice. But, in case of crop biotechnology and breeding, very little amount of work has been done to confirm these results.

"Statistics is the grammar of Science…"

Karl Pearson

CHAPTER 5

DIFFERENTIAL EXPRESSION ANALYSIS OF SINGLE CELL RNA-SEQ DATA: AN OVERVIEW AND COMPARATIVE ANALYSIS

Background

RNA-seq technique measures the aggregated expression levels of thousand(s) of genes from the tissue samples, *i.e.,* a collection of thousand(s) of cells. This technology cannot capture cell-cell heterogeneity since there is no cell-specific information available [192,193]. Hence, scRNA-seq technique is developed for studying the expression dynamics of genes at the single-cell level resolution [194]. Through scRNA-seq, the expression is quantified by mapping reads to a reference genome, followed by counting the number of reads mapped to each gene [195]. Here, individual transcript molecules are attached with a Unique Molecular Identifier (UMI) tag, and subsequently, counting the UMIs yield the number of transcripts for each gene in a cell [196]. Moreover, the scRNA-seq has unique features, such as low library sizes of cells, stochasticity of gene expression, high-level noises, low capturing of mRNA molecules, high dropouts, amplification bias, multi-modality of data. The addition of UMIs during the library preparation step reduces the inherent amplification bias [197] but has no effect on the noises. The noises in scRNA-seq data are mainly due to biological (*e.g.,* stochasticity of gene expression, heterogeneous cell types, cell cycle) and technical factors (*e.g.,*

dropout events, zero-inflation, low input mRNA molecules, low cell capture rates, amplification bias). These biological and technical factors contribute higher proportions of zeros (*i.e.*, zero inflation) or low read counts in the data, characterized as true zeros and dropout zeros, respectively [198–200].

The most commonly performed downstream analysis on scRNA-seq data is DE analysis, which is schematically shown in Figure 5.1.

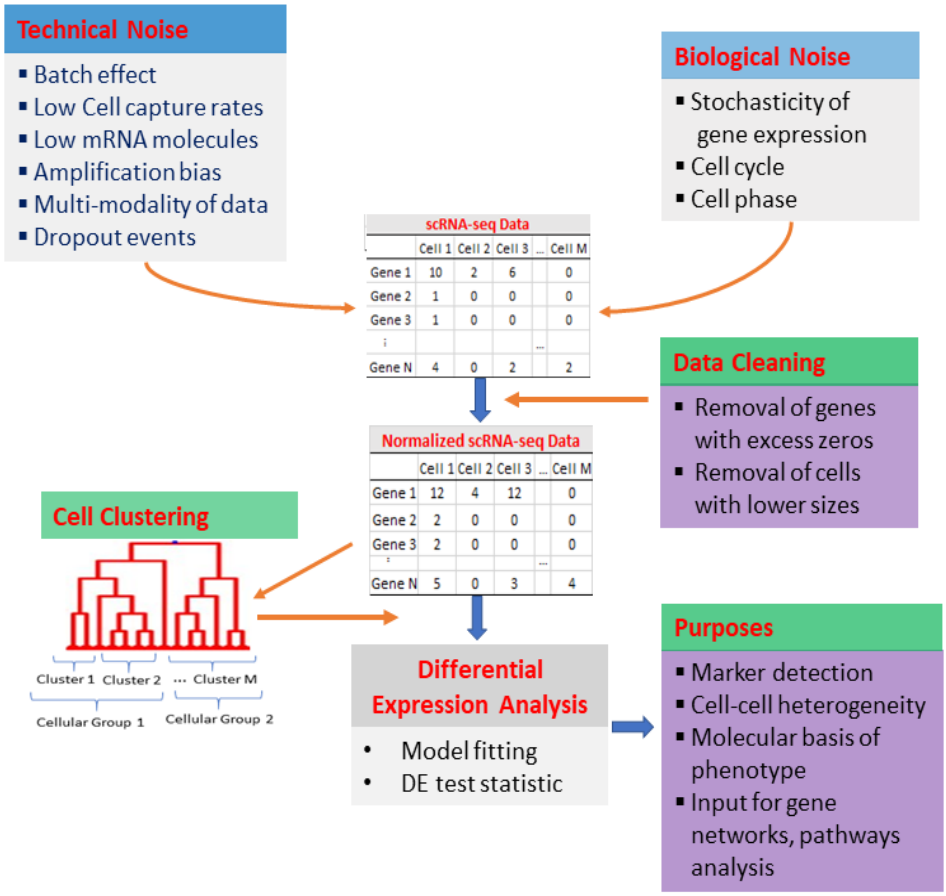


Figure 5.1. A schematic overview of scRNA-seq Differential Expression analysis. scRNA-seq data are inherently noisy with confounding factors. After sequencing, alignment and de-duplication are performed to quantify an initial gene expression profile matrix. Next, normalization is performed with raw expression data using various statistical methods to remove the amplification bias. Additional quality check can be performed when using spike-ins by inspecting the mapping ratio to discard low-quality cells. Finally, the normalized matrix is then subjected to main analysis through clustering of cells to identify subtypes. Cell trajectories can be inferred based on these data and by detecting DE genes between clusters. DE genes can be further to unravel the biological processes of the underlying complex phenotypes through pathway, gene set and network analyses.

The DE analysis is necessary for the identification of gene markers for different cell types, which establishes the molecular basis for phenotypic variation [201]. Further, the detected genes can be used as input for other secondary analyses, such as gene network modeling, pathways, or gene set analysis [202]. Although DE analysis methods for bulk RNA-seq are well reported, these approaches may not be suitable for single-cell data given the special features, such as high-level noise, multi-modality, dropout events, zero inflation (*i.e*., excess of zeros) [203]. For instance, bulk RNA-seq methods, such as edgeR [204] and DESeq2 [205,206] (based on NBD model), are extensively used for the analysis of scRNA-seq data. Further, the utility of such tools may raise serious concerns about their validity due to higher zero dropouts [203], transcriptional bursting [207], lower molecular capturing in cells [208,209], higher dispersion [210], *etc*.

Therefore, dedicated scRNA-seq DE methods are developed, which use different sets of strategies to cope with the above concerns [202,203,208,209,211–213]. For instance, SCDE uses a mixture model (*i.e.,* Poisson for dropout and NB for amplification part) to capture the observed abundance of a given transcript in each cell [214]. There is a lot of DE methods and tools available in the literature, which greatly vary from each other with respect to distributional assumptions of the data, DE test statistic(s), *etc*. [192,193,201,215–217]. Hence, it is pertinent to review the available approaches and tools to understand their statistical theory, unique features, and their limitations. Without sufficient understanding of the underlying statistical principles of these approaches, we may risk drawing erroneous biological interpretations and statistical conclusions. However, there are

minimal studies on classification and rigorous comparative study of the scRNA-seq DE methods in the literature.

In this chapter we, therefore, aim to present a comprehensive review of the up-to-date statistical methods for DE analysis of scRNA-seq data. There are many methodologies developed for bulk RNA-seq, collectively named bulk RNA-seq DE methods, which are extended to single-cell data analysis. Overall, the purpose of these methods is to analyze the data to provide an expansive view of the underlying biological processes, which lead to phenotypic differences (Figure 5.1). The review is organized as follows. In the first part, we overview DE analysis approaches that can be adapted from RNA-seq practice to fit scRNA-seq data as well as those specifically designed for scRNA-seq. While there are plenty of DE approaches, they can be distinguished based on the type of distributional models they fit the data. For instance, the popular DE methods such as DEseq2, edgeR, SAMseq, *etc*. assume that the read counts follow the NB model, while the methods such as DEsingle, DECENT, ZINB-wave, *etc.* assume the read counts follow Zero Inflated Negative Binomial (ZINB) model. Subsequently, we also classify the available approaches into different classes, along with their special features and limitations.

In the second part of this chapter, we attempt to provide a meaningful comparison of several approaches; those are intrinsically statistically different in terms of the model they fit. This includes 19 methods such as DEGseq [218], edgeRLRT [204], edgeRQLF [204], DESeqLRT [206], DESeqNB [205], LIMMA [161], NBPSeq [219], EBSeq [220], BPSC [221], MAST [202], Monocle [213], scDD

[222], NODES [223], DEsingle [211], DECENT [209], T-test [224], Wilcoxon rank sum test (Wilcox) [225], ROTS [226], and EMDomics [227]. Among those, the first 8 methods are designed for bulk-cell RNA-seq, and the next 7 methods are developed for single-cell, and the remaining are general-purpose methods. We compare these methods based on different criteria, such as Area Under Receiver Operating Characteristics (AUROC) curve, FDR, 10 other performance metrics, and runtime on multiple real single-cell datasets. Not surprisingly, the performance of various DE analysis approaches depends on the statistical models they fit and the DE test statistic they use. The findings indicate that the bulk RNA-seq DE methods are competitive and even better compared to some of the single-cell specific methods. Besides, we also assess the performance of the methods under Multiple Criteria Decision Making (MCDM) and combined data setups, which indicated that DECENT and EBSeq are the best options for DE analysis of scRNA-seq data. The similarity analysis of the methods revealed that there exist similarities among the tested methods in terms of detecting common DE genes. These findings were unknown before. Hence, our evaluation provides a proper guideline for selecting the proper DE tool, best performing under particular experimental settings in the context of scRNA-seq.

Overview and Classification of scRNA-seq DE Methods

The available DE analysis approaches used in single-cell data analytics, including bulk RNA-seq DE methods, are listed in Table 5.1. Table 5.1 also presents a comparative overview of the methods in terms of distributional assumptions,

original data motivation (utility), input data type, the test statistic(s), runtime, and their availability platform.

Table 5.1. Description about the potential DE methods used in scRNA-seq study.

SN	Method	Distribution	Utility	Input	DE Test Stat.	Runtime	Availability	Ref.
01	DESeq2	NBD	Bulk cell	Counts	Wald	Low	Bioconductor	[206]
02	edgeR	NBD	Bulk cell	Counts	Quasi-Likelihood F-test, LRT	Low	Bioconductor	[204]
03	LIMMA	Linear Model	Bulk cell	Norm.	Emp. Bayesian Wald t-test	Low	Bioconductor	[161, 228]
04	DEGseq	Poisson Model	Bulk cell	Counts	Z-score	Low	Bioconductor	[218]
05	t-test	t-test	General	Norm.	t-test statistic	Low	CRAN	[224, 229]
06	Wilcox	Wilcoxon test	General	Counts	Wilcoxon signed rank	Low	CRAN	[225, 229]
07	baySeq	NBD	Bulk cell	Counts	Posterior prob.	Low	Bioconductor	[230]
08	NBPseq	NBD	Bulk cell	Counts	Fisher's exact test	Low	CRAN	[219]
09	EBSeq	NBD	Bulk cell	Counts	Bayesian	High	Bioconductor	[220]
10	Cuffdiff	Beta-NBD	Bulk cell	sam		Low	Linux	[231]
11	SAMseq	NP	Bulk cell	Counts	Mann-Whitney statistic	Low	CRAN	[232]
12	Ballgown	Linear Model	Bulk cell	Counts	Lin. Mod. test statistic	Medium	Bioconductor	[233]
13	TSPM	Poisson Model	Bulk cell	Counts		Low	R code	[234]
14	ROTS	NP	Bulk cell	Norm.	*Z*- statistic *(bootstrap)*	Medium	Bioconductor	[226, 235]
15	metagenomeSeq		Bulk cell			Medium		[236]
16	SCDE	Mixture Model	SC	UMI	Bayesian Stat.	High	Bioconductor	[214]
17	scDD	Multi-Modal Bayesian	SC	Norm.	Bayesian Stat.	High	Bioconductor	[222]
18	D3E	NP	SC	UMI	Cramér-von Mises test/ KS test	High	GitHub, Python	[237]
19	BPSC	Beta-Poisson	SC	UMI	LRT	Medium	GitHub	[221]
20	MAST	Hurdle	SC	Norm.	LRT	Medium	Bioconductor	[202]
21	Monocle	GAM	SC	Norm.	LRT	Medium	Bioconductor	[213]
22	DEsingle	ZINB	SC	UMI	LRT	High	Bioconductor, GitHub	[211]
23	DECENT	ZINB	SC	UMI	LRT	High	GitHub	[209]
24	DESCND	PD	SC	UMI		High	GitHub	[208]
25	EMDomics	NP	SC	Norm.	Euclidean distance	High	Bioconductor	[227]
26	Sincera	NP	SC	Norm.	Welch t-test(L) Wilcox (S)	High	GitHub	[238]
27	ZIAQ	Logistic Regression	SC	Norm.	Fisher's test	Medium	GitHub	[239]

28	sigEMD	NP	SC	Norm.	Distance measure	High	GitHub	[240]
29	TASC	Logistic, Poisson Models	SC	UMI	LRT	High	GitHub	[241]
30	ZINB-Wave	ZINB	SC	UMI	LRT	High	Bioconductor, GitHub	[203, 242]
31	NODES	Wilcox	SC	Norm.	Wilcoxon test	Medium	*Dropbox	[223]
32	BASiCS	Poisson-Gamma	SC	Norm.	Posterior prob.	High	Bioconductor	[243]
33	NBID	NBD	SC	UMI	LRT	Medium	R code	[244]
34	tradeSeq	NBD (GAM)	SC	UMI	Wald test	Medium	GitHub	[245]
35	SC2P	ZIPD	SC	UMI	Posterior prob.	High	GitHub	[246]

Bulk: bulk RNA-seq; SC: Single Cell methods; ZINB: Zero Inflated Negative Binomial; UMI: Unique Molecular Identification counts; scRNA-seq: Single cell RNA-seq; Norm.: Normalized counts (Continuous); Ref.: Reference cited; GAM: Generalized Additive Model; LRT: Likelihood Ratio Test; L: Large number of Samples; S: Small number of Samples; KS: Kolmogorov-Smirnov's test

Instead of reviewing them individually, we classified these methods based on different factors, which is shown in Figure 5.2.

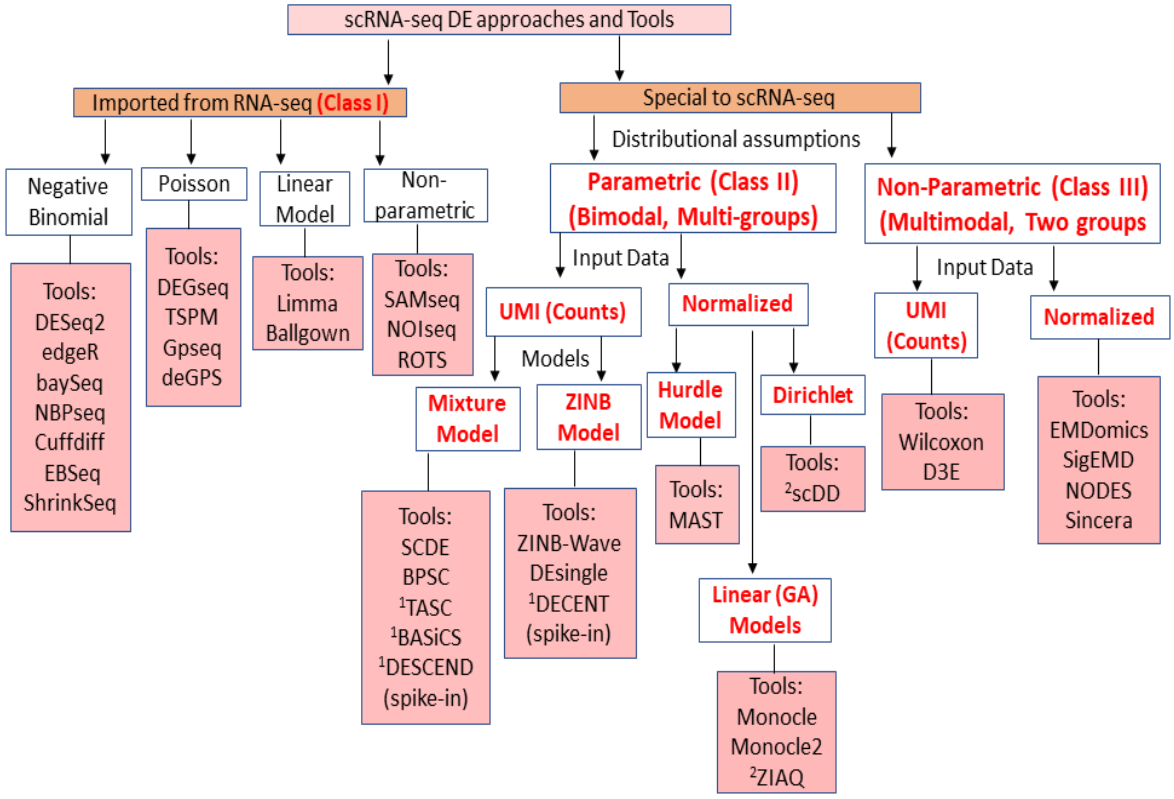


Figure 5.2. Schematic Representation of Classification of DE Methods and Tools. Schematic overview illustrating the breakup of the DE methods that can be adapted from RNA-seq practice to fit scRNA-seq data (Class I) as well as those specifically designed for single-cell (Classes II, III) based on different distribution models that they fit for DE analysis. Different example tools belonging to each category are listed in pink color boxes.[1]Methods use the external RNA spike-ins and [2]Parametric approaches but can handle multi-modality of the data.

In other words, Figure 5.2 illustrates different distribution models used to fit the count data, DE test statistic(s), ability to use external spike-ins, *etc.* and utility of the different scRNA-seq DE methods. The available methods can be classified based on origin, *i.e*., methods originally developed for bulk RNA-seq but later extended to scRNA-seq and methods exclusively designed for single-cell (Figure 5.2). Further, the bulk RNA-seq methods can be classified into Parametric and NP (Figure 5.2). The former class assumes that the data follows certain count data models, while the latter is distribution-free. For instance, the parametric class methods mostly assume the read counts are obtained from Poisson or NB distribution and based on this, software packages, such as edgeR [204], DESeq2 [205,206], BaySeq [230], DEGseq [218], TSPM [234] are developed. On the contrary, NP (*i.e.,* distribution-free) DE methods estimate the parameters that can quantify the distribution of expression profiles and make comparisons between case *vs.* control groups. The tools for this category includes SAMseq [232], TOISeq, ROTS [226,235], to name a few, which are developed for bulk RNA-seq, but later extended to scRNA-seq (Figure 5.2).

The bulk RNA-seq DE methods suffer from serious limitations, as listed in Table 5.2, when they are extended to the scRNA-seq. Likewise, the methods specially developed for single-cell data can be grouped into parametric and NP based on the assumption of the underlying distributions of the UMI counts data. The parametric methods assume that the UMI counts follow count models, such as zero inflated models, (ZINB, Zero Inflated Poisson (ZIPD)), Mixture Models, (Beta-Poisson, Poisson-NB, NB-Logistic, and Hurdle models). The R packages,

such as DEsingle, DECENT, ZINB-wave, BPSC, SCDE, and MAST, listed a few, belong under this category. Further, the NP methods are implemented in software packages, such as D3E [237], sigEMD [240], Sincera [238], NODES [223], and EMDomics [227]. These approaches estimate the parameters that can quantify the distribution of the distribution of expression profiles and can handle the multi-modality of scRNA-seq data but limited only to two groups comparisons. The special features, pros, and limitations for various classes of methods are listed in Table 5.2.

Table 5.2. Classification of methods used for detection of DE genes in scRNA-seq data.

SN.	Classes	Descriptions
01	Class I	**Underlying Models:** Negative Binomial Model; Linear Model; Poisson Model; Bayesian Model*
		Features: Computationally simple; Require less run time; Applicable to both counts and normalized data
		Limitations: • Do not consider the multi-modality, dropout events, zero-inflation • Overestimate the dispersion parameter • Underestimate the mean (difference in mean across cellular groups) • Less statistical power to detect DE genes • Do not consider the higher technical and biological variations • Cannot handle the long-tailed (skewed) distributions • Cannot handle high sparsity
		Tools: DEseq2[205,206], edgeR[204], Limma[161], SAMseq[232], DEGSeq[218], baySeq[230], NBPseq[219], Cuffdiff[231], Ballgown[233], TSPM[234], metagenomeSeq[236], ROTS[226,235], NOISeq[247] EBSeq[220], ShrinkSeq[248], GPseq[249], DeGPS[250]
02	Class II	NP methods
		Features: • Distribution free approaches • Considers the multi-modality of the data • Computationally not cumbersome (less runtime) • Without fitting the distribution of genes and estimate the parameters

		• Performs DE analysis with distance like metrics across two conditions for genes • Performed well when lesser proportions of zeros **Limitations:** • Mostly focused on two cellular groups comparison • Computationally complex for multi-groups • Performance severely affected due to high dropouts (some methods exclude dropouts) • Cannot separate between true/biological and false/dropout zeros • Sensitive to sparsity • Methods like D3E, scDD failed to consider UMI count nature of data • Cannot separate technical from biological sources of variation **Tools:** D3E[237], scDD[222], sigEMD[240], NODES[223], EMDomics[227], Sincera[238], ZIAQ[239], Wilcoxon signed rank test
03	**Class III**	**Models:** • Zero inflated Models; Hurdle Models; Mixture Models; GLM, GAM **Features:** • Parametric approaches • Capture the bimodality of the scRNA-seq data • Easily applicable to multi-cellular groups • Considers the zero inflations, dropout events in scRNA-seq data • Methods like TASC, DECENT, etc. make use of external spike-in data for model building • Mostly uses the GLM framework to compute DE statistics **Limitations:** • Cannot capture the multimodality (> 2) of scRNA-seq data • Methods like MAST failed to consider UMI count nature of data and excludes the dropout events • Methods like SCDE and MAST does not differentiate between true/biological and dropout zeros during the model building • Computationally intensive and require more runtime • Most of them do not distinguish biological from technical factors that are causing zero-inflation. • Assumes the dropout events to be linear, however, the effect of dropout events is likely to be non-linear, especially for genes with low to moderate expression **Tools:** SCDE [214], NBID [244], MAST[202], Monocle[212], Monocle2[251], BPSC[221], ZINB-Wave[203], DEsingle[211], DECENT[209], DESCEND[208], TASC[241], BASiCS[243], Random Hurdle Model [252], SC2P [246]

Besides, classification of DE methods can be made based on the nature of input data to the concerned tools, such as discrete/counts (UMI reads) or continuous (Fragments Per kilo-base per Million reads (FPKM) or, Counts-Per-

Million reads (CPM), normalized data). The methods which are specialized to handle the UMI counts include DEsingle, DECENT, DESCEND, ZINB-wave, to name a few. Other methods based on continuous data or transform the original UMI counts include scDD, MAST, Monocle2, EMDomics, ROTS, *etc*. (Table 5.1). However, such methods ignore the original nature of UMI counts, and subsequently, there is a chance of losing some information. Another classification of DE methods can be possible through the use of the type of test statistic(s) they use for DE testing. This includes methods, *e.g.,* DESeq2, edgeR, DECENT, BPSC, MAST, Monocle2, *etc.* based on the Likelihood Ratio Test (LRT) statistic computed in the Generalized Linear Model (GLM) framework. The other class of DE methods, such as D3E, NODES, EMDomics, Sincera, sigEMD, where the DE test statistic(s) are computed from the NP testing procedure. Moreover, the UMI data provides an opportunity to integrate the molecular capturing process with the parametric DE testing models, which improve the performance of the DE methods [209]. In other words, the available scRNA-seq DE methods can also be classified based on the requirement of external spike-ins (*e.g.,* External RNA Controls Consortium (ERCC) spike-ins) for fitting the models. The class of methods explicitly considers technical variation and molecular capturing process based on external spike-ins data. This includes methods such as TASC [253], BASiCs [243], DECENT [209], and DESCEND [208]. Here, the spike-ins data is used for the computation of cell capture rates, subsequently integrated into the count data modeling process.

Real ScRNA-seq datasets

To assess the performance of the methods, we used the publicly available real scRNA-seq datasets to study their real data behaviors. This process starts with the collection of these scRNA-seq datasets from the GEO NCBI database (https://www.ncbi.nlm.nih.gov/geo). In our comparative analysis, we included the 11 UMI count gene expression datasets derived from 9 independent scRNA-seq studies. The main rationale behind selecting the count expression data, as they are well quality checked and preprocessed by the authors of these original publications. Further, these datasets were generated over protocols, including SmartSeq, DropSeq, NextSeq, HiSeq, MARSseq, and SCRBSeq. The selected datasets include scRNA-seq data from lung cancer cells, pluripotent stem cells, liver cells, adipose stem/stromal cells, HEK cells, breast cancer cells from humans, and embryonic stem cells, blood cells, embryonic fibroblasts cells, and cells from mice. A brief description of the selected real datasets is given in Table 5.3. For instance, Islam data consists of 22928 genes over 92 cells (48: mouse embryonic stem cells; 44 mouse embryonic fibroblasts cells), available at the GEO database with accession GSE29087 was taken in this study. Then, to reduce the dimension of the data, we filtered out the low expressed genes, *i.e.,* genes which do not have non-zero expressions in at least 5 cells and cell library sizes below 1000. Further, the reference genes for the same cell lines were collected from the Microarray study available at http://carlosibanezlab.se//Data/Moliner_CELfiles.zip [254] to assess the performance of the methods. Similar descriptions about other real datasets including Tung data [197], Soumillon1 data [255], Soumillon2 data [255],

Soumillon3 data [255], Klein data [256], Gierahn [257], Chen data [244], Savas data [258], Grun data [259], Ziegenhain data [260] are given in Table 5.3.

Table 5.3. List of the scRNA-seq datasets used in this study.

Data	Description	Ref. gen.	Protocol	#Genes	#Cell	Ref
Tung	Human induced Pluripotent stem cell lines.	Bulk RNA-seq	HiSeq	18938	576	[197]
Islam	single-cell transcriptional landscape by highly multiplex RNA-Seq	Microarray	Smartseq	22928	92	[196]
Soumillon1	Differentiating adipose cells by scRNA-Sequencing (Day 1 *vs* 2)	scRNA-seq	HiSeq	23895	1835	[255]
Soumillon2	Differentiating adipose cells by scRNA-Sequencing (Days 1 *vs* 3)	scRNA-seq	HiSeq	23895	2268	[255]
Soumillon3	Differentiating adipose cells by scRNA-Sequencing (Days 2 *vs* 3)	scRNA-seq	HiSeq	23895	1613	[255]
Klein	Mouse embryonic stem cells	scRNA-seq	Droplet	24174	1481	[256]
Gierahn	Single-cell RNA sequencing experiments of HEK cells	scRNA-seq	NextSeq	24176	1453	[257]
Chen	ScRNA-seq of Rh41 using 10x Genomics	Bulk RNA-seq	HiSeq	33694	7261	[244]
Savas	Breast cancer cells using 10x Genomics	Bulk RNA-seq	HiSeq	33694	6311	[258]
Grun	Mouse embryonic stem single cells using CEL-seq technique	scRNA-seq	HighSeq	12467	320	[259]
Ziegenhain	Sc-RNA sequencing of Mouse embryonic stem cells	scRNA-seq	**A***	39016	583	[260]

#genes: number of genes, #cells: number of cells; A: CEL-seq2, Drop-seq, MARS-seq, SCRB-seq, Smart-seq and Smart-seq2; Ref.: cited reference; Ref. gen.: type of study from which reference genes are obtained.

Notations: Y_{ij}: *rv* represents observed read (UMI) counts of i^{th} (i = 1, 2, …, N) gene in j^{th} (j = 1, 2, …, M) cell; N: total number of genes; M: total number of cells; μ_{ij}: mean of i^{th} gene in j^{th} cell for NB distribution (count part of the model); θ_{ij} $(= \varphi_{ij}^{-1})$ and φ_{ij}: size and dispersion parameters respectively of i^{th} gene in j^{th} cell for NB distribution; π_{ij}: mixture probability (zero inflation probability) of i^{th} gene in j^{th} cell; s_j: size factor of j^{th} cell; Z_{ij}: *rv* represents the true (unknown) concentration of reads for i^{th} gene of j^{th} cell; $\boldsymbol{X}$: design matrix for cell group information, the j^{th} row of $\boldsymbol{X}$,

$X_j = [X_{j1}, X_{j2}, \ldots, X_{jN}]$; W_{ij}: indicator *rv* representing the rate of expression for i^{th} gene in j^{th} cell, *i.e.* $W_{ij} = 0: Y_{ij} = 0$; $W_{ij} = 1: Y_{ij} > 0$.

Count Data Models for scRNA-seq Data

NBD Model

The PMF of the NB distribution is expressed as:

$$f_{NB}(y) = P[Y_{ij} = y] = \frac{G(y+\theta_{ij})}{G(y+1)G(\theta_{ij})}\left(\frac{\theta_{ij}}{\theta_{ij}+\mu_{ij}}\right)^{\theta_{ij}}\left(\frac{\mu_{ij}}{\theta_{ij}+\mu_{ij}}\right)^{y} \quad \forall\, y = 0, 1, 2, \ldots \quad (5.1)$$

where, $\mu_{ij} \geq 0$; $\theta_{ij} > 0$ are the parameters of NB distribution, *G*(.): Gamma function. Then, the expected value and variance of Y_{ij} is shown as:

$$E(Y_{ij}) = \mu_{ij} \quad (5.2)$$

$$V(Y_{ij}) = \mu_{ij} + \frac{{\mu_{ij}}^2}{\theta_{ij}} = \mu_{ij} + \varphi_{ij} \quad (5.3)$$

If $\varphi_{ij} \to 0\ (No\ dispersion) \ \dot{\Rightarrow}\ NB(\mu_{ij}, \theta_{ij}) \ \to Poisson(\mu_{ij})$

ZINB Model

For any $\pi_{ij} \in [0, 1]$, Y_{ij} is assumed to follow a ZINB distribution [203,209,211]. The PMF of the ZINB Distribution is expressed as follows.

$$f_{ZINB}(y) = P[Y_{ij} = y] = \pi_{ij}\delta_0(y) + (1 - \pi_{ij})f_{NB}(y) \quad \forall\, y = 0, 1, 2, \ldots \quad (5.4)$$

where, $f_{NB}(.)$: PMF of NB distribution (Eq. 5.1); $\delta_0(.)$: Dirac's delta function. Here, $\delta_0(.)$ used to model the excess zeros in the data, and its PMF is expressed as:

$$\delta_0(Y_{ij} = y) := \begin{cases} 1; & y = 0 \\ 0; & y \neq 0 \end{cases} \quad (5.5)$$

Now, the PMF of the ZINB distribution to model the read counts from scRNA-seq data is given in Eq. 5.6.

$$P[Y_{ij} = y] = \begin{cases} \pi_{ij} + (1 - \pi_{ij}) \left(\frac{\theta_{ij}}{\theta_{ij}+\mu_{ij}}\right)^{\theta_{ij}} & when\ y = 0 \\ (1 - \pi_{ij}) \frac{G(y+\theta_{ij})}{G(y+1)G(\theta_{ij})} \left(\frac{\theta_{ij}}{\theta_{ij}+\mu_{ij}}\right)^{\theta_{ij}} \left(\frac{\mu_{ij}}{\theta_{ij}+\mu_{ij}}\right)^{y} ; & y > 0 \end{cases} \quad (5.6)$$

Now, $Y_{ij} \sim ZINB(\pi_{ij}, \mu_{ij}, \theta_{ij})$, then the expected value and variance of Y_{ij} can be obtained as follows:

$$E(Y_{ij}) = (1 - \pi_{ij})\mu_{ij} \quad (5.7)$$

$$V(Y_{ij}) = (1 - \pi_{ij})\mu_{ij}\left(1 + \pi_{ij}\mu_{ij} + \frac{\mu_{ij}}{\theta_{ij}}\right) \quad (5.8)$$

If $\pi_{ij} = 0 \dot{\Rightarrow} ZINB(\pi_{ij}, \mu_{ij}, \theta_{ij}) \rightarrow NB(\mu_{ij}, \theta_{ij})$

If $\varphi_{ij} \rightarrow 0\ (No\ dispersion) \dot{\Rightarrow}\ ZINB(\pi_{ij}, \mu_{ij}, \theta_{ij}) \rightarrow ZIP(\pi_{ij}, \mu_{ij})$

Poisson Distribution

Poisson Distribution (PD) are also extensively used for analysis of count data obtained from bulk RNA-seq or scRNA-seq experiments. The PMF of PD can be expressed as:

$$f_{PD}(y) = P[Y_{ij} = y] = \frac{e^{-\mu_{ij}}\mu_{ij}^{y}}{G(y+1)} \quad \forall\ y = 0, 1, 2, \ldots \quad (5.9)$$

$$E(Y_{ij}) = Var(Y_{ij}) = \mu_{ij} \quad (5.10)$$

Zero Inflated Poisson Distribution (ZIPD)

Poisson model has very strict assumptions, *i.e.,* mean equals the variance, which is often violated in scRNA-seq data analysis. When the variance is too large because there are many 0s as well as a few very high values for expression counts [261]. In this case, a better solution is often the ZIPD model.

The PMF of ZIPD distribution can be expressed as:

$$f_{ZIPD}(y) = P[Y_{ij} = y] = \pi_{ij} I(y = 0) + (1 - \pi_{ij}) f_{PD}(y) \qquad \forall\ y = 0, 1, 2, \ldots \quad (5.11)$$

$$= \begin{cases} \pi_{ij} + (1-\pi_{ij})e^{\mu_{ij}} & when\ y = 0 \\ (1-\pi_{ij})\frac{e^{-\mu_{ij}}\mu_{ij}{}^{y}}{G(y+1)}; & y > 0 \end{cases} \quad (5.12)$$

The mean and variance of ZIPD model is shown in Eq. 5.13 and 5.14, respectively.

$$E(Y) = (1-\pi_{ij})\mu_{ij} \quad (5.13)$$

$$Var(Y) = (1-\pi_{ij})\mu_{ij}(1+\pi_{ij}\mu_{ij}) \quad (5.14)$$

Hermite Distribution

Hermite Distribution (HD) can be used to model the counts data [262]. Further, the PMF of HD is given in Eq. 5.15.

$$f_{HD}(Y_{ij} = y|\alpha_{ij}, \beta_{ij}) = e^{-(\alpha_{ij}+\beta_{ij})} \sum_{k=0}^{[\frac{y}{2}]} \frac{\alpha_{ij}{}^{y-2k}\beta_{ij}{}^{k}}{G(y-2k+1)G(k+1)} \forall\ y = 0, 1, 2, \ldots \quad (5.15)$$

where, [.]: integral part. The mean, variance, and dispersion index (*i.e.*, ratio between variance and mean) of rv Y_{ij}~HD (α_{ij} , β_{ij}) is given in Eq. 5.16 – 5.18.

$$E(Y_{ij}) = f(\alpha_{ij}, \beta_{ij}) = (\alpha_{ij} + 2\beta_{ij}) \quad (5.16)$$

$$Var(Y_{ij}) = (\alpha_{ij} + 4\beta_{ij}) \quad (5.17)$$

$$\varphi = g(\alpha_{ij}, \beta_{ij}) = 1 + 2\beta_{ij}/(\alpha_{ij} + 2\beta_{ij}) \quad (5.18)$$

The good-ness of fit of the above count data models were assessed through Akaike Information (AIC) and Bayesian Information (BIC) Criteria.

Statistical Tests for Zero inflation and Overdispersion for scRNA-seq Data

For simplicity we assume that the parameters for each gene remain same over the cells, *i.e.*, $\mu_{i1} = \mu_{i2} = \cdots = \mu_{iM} = \mu_i$; $\theta_{i1} = \theta_{i2} = \cdots = \theta_{iM} = \theta_i$; $\pi_{i1} = \pi_{i2} = \cdots = \pi_{iM} = \pi_i$. For testing the statistical significance of the dispersion parameter of i^{th} gene, θ_i, we adopt the following LRT procedure. Here, for the testing purpose, we define the following null hypothesis.

Hypothesis for overdispersion: $H_{10}: \theta_i = 0\ vs.\ H_{11}: \theta_i \neq 0$

Hypothesis for Zero Inflation: $H_{20}: \pi_i = 0\ vs.\ H_{21}: \pi_i \neq 0$

where, $H_{.0}$: null hypothesis; $H_{.1}$: alternate hypothesis. Here, H_{10} tells us that *i*th gene is not overdispersed, means the mean is same as the variance and subsequently, the scRNA-seq count data is obtained from a Poisson model. Further, if we fail to reject H_0, then we can say the UMI counts data is not overdispersed and simply fitting a Poisson model will give satisfactory results. Further, H_{20} speaks that *i*th gene is not zero inflated, and the scRNA-seq data structure is same as bulk RNA-seq data. If we fail to reject H_0, then the RNA-seq DE tools can be used for DE analysis of scRNA-seq data with the expectation of satisfactory results.

The above tests can be tested through the LRT statistic(s) given in Eq. 5.19.

$$-2ln\alpha = -2\{l(\boldsymbol{\Omega}_i = \widehat{\boldsymbol{\Omega}}_{i0};\ Y_{ij}) - l(\boldsymbol{\Omega}_i = \widehat{\boldsymbol{\Omega}}_i;\ Y_{ij})\} \quad (5.19)$$

where, $\widehat{\boldsymbol{\Omega}}_{i0}$: MLE of $\boldsymbol{\Omega}_i$ for *i*th gene under the constraint of H_0 and $\widehat{\boldsymbol{\Omega}}_i$: unconstrained MLE of $\boldsymbol{\Omega}_i$ for *i*th gene, $\boldsymbol{\Omega}_i$: parametric space for *i*th gene, *i.e.* $\boldsymbol{\Omega}_i = \{\mu_i, \theta_i, \pi_i\}$. The test statistic in Eq. 5.19 is asymptotically distributed as Chi-square distribution with 1 degree of freedom (df) under H_0.

Methods for scRNA-seq DE Analysis

NBD Model based Methods

DESeq

DESeq [205] assumes that Y_{ij} follows a NB model (Eq. 5.1). In other words, the read counts are modeled by the NB distribution with μ_{ij} and $V(Y_{ij})$ estimated from the scRNA-seq data. For each gene, the μ_{ij} (Eq. 5.2) is modeled as the product of

the $E(Z_{ij})$, and s_j (accounts for sequencing depth of the cell). Further, the Y_{ij} can be described with the NB GLM framework through the following expressions.

$$Y_{ij} \sim NB(\mu_{ij}, \theta_{ij}) \tag{5.20}$$

$$\mu_{ij} = s_j E(Z_{ij}) \tag{5.21}$$

$$log_2 E(Z_{ij}) = \beta_{0i} + \beta_{1i} X_j \tag{5.22}$$

where, X_j is simply the binary indicator of cellular group, β_{0i}: logarithm of mean parameter for *i*th gene in the reference cell type, β_{1i}: log fold-change parameter for *i*th gene. DESeq first estimates the size factors that account for the differences in the library size, then estimates the dispersion, and lastly, fits a GLM for each gene. The DESeq uses various test statistic(s) to compute the *p-value* and size effect estimate for the log2 FC. For DESeq, we used two methods based on LRT and NB test statistic(s) through executing *nbinomTest*, and *DESeq* functions respectively implemented in DESeq2 R package [206].

edgeR

Like DESeq, edgeR [204] also models Y_{ij} using a NB distribution (Eq. 5.1). For each gene, the $\mu_{ij}$is assumed to be the product of the total number of reads and the (unknown) relative abundance of that gene in the current experimental condition. Here, $V(Y_{ij})$ is a function of μ_{ij}, as shown in Eq. 5.2 and which requires the estimation of the overdispersion parameter (φ_{ij}). So, edgeR estimate φ_{ij} using a conditional Maximum Likelihood Estimation (MLE) procedure, conditioning on the total read count of each gene and an empirical Bayes procedure to shrink the dispersions toward a consensus value [263]. For each gene, DE test statistic(s) are computed through the GLM based LRT [264] or Quasi-Likelihood F test (QLF).

Here, we used two methods based on LRT, and QLF test statistic(s), *i.e.,* edgeRLRT and edgeRQLF through executing *glmLRT* and *glmQLFTest* implemented in edgeR R package [204].

NBPSeq

NBPSeq [219] method (or NBSeq) was originally developed for RNA-seq data to detect the DE of genes, which assumes the read counts follow NB distribution. The DE testing procedure is based on the NBP parameterization of the NB distribution and uses the extended version of the exact test proposed by Robinson and Smyth (2007) [263]. Through this test, the constant dispersion parameter is used to model the count variability between biological replicates and introduced an additional parameter to allow the dispersion parameter to depend on the mean. To implement the NBPSeq method, we executed the *nbp.test* function implemented in NBPSeq R package [219].

EBSeq

EBSeq [220] assumes the true (unknown) read counts follow the NB model and uses a Beta prior distribution to model the fluctuations in technical and biological variations. For RNA-seq data with two biological conditions, EBSeq tests the hypothesis, $H_0: \mu_{i1} = \mu_{i2}$, using Bayesian approaches through incorporating prior probability of DE of counts (modelled by the mixture distribution). Here, means and variances of genes are obtained through the method-of-moments, and the four global hyperparameters are computed using Expected Maximization (EM) algorithm. With these parameter estimates, the posterior probability of DE of genes obtained using Bayes' rule and subsequently DE genes are detected. To execute

this method, the *EBTest* function implemented in EBSeq R package [220] was used.

Poisson Model based Method

DEGSeq

DEGSeq [218] assumes the read counts follow a PD model [265], PMF given in Eq. 5.9. The model parameters were estimated using the MLE method by maximizing the concave joint likelihood function [265]. Further, with the estimates of PD parameters, the DE genes in bulk RNA-seq data are identified through Fisher's exact and the LRT statistic(s) [265]. Here, we used only the LRT statistic to detect DE genes in scRNA-seq data through executing the *DEGexp* function implemented in DEGSeq R package [218].

DEsingle

DEsingle [211] is a Zero Inflated Model (ZIM) based approach that employs the ZINB model given in Eq. 5.6 to discriminate the observed zero values into two parts dropout and true zeros (*i.e.,* from NB distribution). Under this model formulation, DEsingle is designed to overcome the issues of the excessive zeros observed in the scRNA-seq data. To detect DE genes between two cell groups, DEsingle first calculates the MLE of two ZINB populations parameters in Eq. 5.6. Then detects the DE genes using the LRT statistic through the constrained MLE of the two models' parameters under the null hypothesis. Here, the *p-values* for the genes were computed through executing *DEsingle* function implemented in DEsingle R package [211].

DECENT

DECENT [209] is based on ZIM, precisely use the ZINB model given in Eq. 5.6 for fitting scRNA-seq data, which also explicitly and accurately models the molecular capture process using a Beta-Binomial model. Here, the unobserved true UMI counts, Z_{ij}, are assumed to follow ZINB model (Eq. 5.6). Further, DECENT assumes the following models for different processes.

$$Z_{ij}; \pi_{ij}, s_j, \mu_{ij}, \theta_{ij} \sim ZINB(\pi_{ij}, s_j\mu_{ij}, \theta_{ij}) \tag{5.23}$$

$$Y_{ij}\big|Z_{ij} = k; p_{ij} \sim B(k, p_{ij}) \tag{5.24}$$

$$p_{ij} \sim Beta(a_{ij}, b_{ij}) \tag{5.25}$$

where, p_{ij} be the transcriptional capture rate for *i*th gene of *j*th cell, *B*(.): Binomial distribution, $a_{ij}, and\ b_{ij}$ in Eq. 5.25 are the parameters of the beta distribution. DECENT uses the Expected Conditional Maximization (ECM) algorithm to calculate MLE of the ZINB model parameters (Eq. 5.23) using the observed data through integrating molecular capturing procedure in the presence of external RNA-spike ins. To detect DE genes, DECENT uses the GLM framework in Eq. 5.26 to model the μ_{ij}.

$$log\ \mu_{ij} = \beta_{0i} + \beta_{1i}X_j + \tau_i'U_j \tag{5.26}$$

where, β_{0i}, β_{1i}, X_j has the usual meaning as in Eq. 5.22 and τ_i: the regression coefficient of *i*th gene for *j*th cell-level auxiliary, U_j. The *p-values* for each gene are computed through LRT statistic under the GLM (Eq. 5.19), which is executed through *decent* function implemented in DECENT R package [209].

Mixed Model based Methods

BPSC

BPSC [221] is an analytical method based on Beta-Poisson (BP) mixture model, designed to capture the distributional features of the scRNA-seq data, *i.e.,* non-integer expression or low expression values. Through this, the normalized data, such as FPKM or CPM, are modeled by using a four parameters BP model given in Eq. 5.27.

$$BP_4\left(Y_{ij}\middle|\alpha,\beta,\vartheta_1,\vartheta_2\right) = \vartheta_2 P\left(Y_{ij}\middle|\vartheta_1 Beta(\alpha,\beta)\right) \tag{5.27}$$

where, Y_{ij}: normalized value of the read counts; *P*(.): Poisson PMF; $\alpha,\beta,\vartheta_1,\vartheta_2$ are the parameters of the BP model. The expected value and variance of Y_{ij} is expressed in Eq. 5.28 and 5.29, respectively.

$$E\left(Y_{ij}\right) = \mu_{ij} = \vartheta_1\vartheta_2\frac{\alpha}{\alpha+\beta} \tag{5.28}$$

$$V\left(Y_{ij}\right) = \mu_{ij}\vartheta_2 + {\mu_{ij}}^2\frac{\beta}{\alpha(\alpha+\beta+1)} \tag{5.29}$$

The MLEs of the parameters in Eq. 5.29 are estimated using the iterative weighted least-squares algorithm [221]. The DE analysis of the genes was carried out under the GLM frameworks given in Eq. 5.19 and 5.22. Further, *p-values* for the genes are computed through the LRT statistic by executing *BPglm* function implemented in the BPSC R package [221].

scDD

scDD [222] method based on Logistic- Dirichlet mixture model, which is designed to model the scRNA-seq data under a Bayesian modeling framework. It models the excess zeros in scRNA-seq data using logistic regression and also models the non-zero counts using the conjugate Dirichlet model of normal distributions. Here, the UMI counts are transformed to CPM measures through *cpm* function implemented in edgeR R package [204] followed by log-transformation. scDD uses

a Bayesian modeling approach to detect DE genes between the two cellular groups. For this purpose, it computes an approximate Bayes factor score that compares the probability of DE with the probability of non-DE for each gene. The empirical gene *p-values* for the DE tests are computed using a permutation method. To execute this method, we used *scDD* function implemented in scDD R package [222].

Normal based methods

LIMMA

LIMMA [161,228], based on linear modelling, was originally designed for Microarrays but recently extended to bulk RNA-seq data. For expression counts, LIMMA uses Voom transformations [228]. It considers gene-specific linear models to model the transformed expression values of counts, Y_{ij}^{v}, given as:

$$E\left(Y_{ij}^{v}\right) = \boldsymbol{X}\boldsymbol{\omega_i} \tag{5.30}$$

$$Var\left(Y_{ij}^{v}\right) = L_i\sigma_i^2 I \tag{5.31}$$

where, $\boldsymbol{\omega_i}$: regression coefficient vector for *i*th gene, L_i: known weight matrix for *i*th gene, and σ_i^2: variance of *i*th gene. For performing the DE analysis of scRNA-seq data, the empirical Bayes approach was used by incorporating the expected value-variance relationship [161]. In this study, *voom*, *lmFit*, and *eBayes* functions implemented in limma R package are executed for data transformations, model fitting, and DE analysis, respectively.

MAST

MAST [202] uses a hurdle model approach for DE analysis and assumes conditional independence between expression rate (W_{ij}) and expression levels

(Y_{ij}) for each gene. It fits a logistic regression for W_{ij} and fits a Gaussian linear model for the continuous variable (Y_{ij} | W_{ij} = 1), which can be summarized as:

$$logit\left[\Pr\left(W_{ij} = 1\right)\right] = \boldsymbol{X_j \beta_i} \tag{5.32}$$

$$Pr(Y_{ij} = y | W_{ij} = 1) = N(\boldsymbol{X_j \beta_i}, \sigma_i^2) \tag{5.33}$$

In order to improve the inference for genes with sparse expression, the model parameters are fitted using an empirical Bayesian framework [202]. Finally, DE testing for genes is performed across the two cellular groups through the LRT statistic(s), given in Eq. 5.19. For this purpose, we executed *zlm*, and *summary* functions for hurdle model fitting and DE analysis respectively implemented in MAST R package [202].

Monocle

Monocle [212,213] (updated as Monocle2 [213]), a specially designed method for DE analysis, *i.e.* identifying DE genes that vary across different cell types or pseudo-times in scRNA-seq data. It uses a generalized additive model (GAMs) to model μ_{ij} under the GLM framework, *i.e.* relating μ_{ij} to one or more predictors through GAMs for each gene and is expressed as:

$$log\mu_{ij} = \beta_{0i} + f_1(x_1) + f_2(x_2) + \cdots + f_M(x_M) \tag{5.34}$$

where, β_{0i}: regression co-efficient; x_j: predictor variable that represents group memberships of the cells; $f_j(.)$: smoothing functions, *e.g.,* cubic splines. Specifically, Y_{ij} across the cells are modeled using a Tobit model (approximately); thus, Monocle's GAM becomes:

$$\mu_{ij} = s\left(\delta_t\left(b_x, f_j\right)\right) + \varepsilon \tag{5.35}$$

where, $\delta_t(b_x, f_j)$: pseudo-time or cell type of a cell; f_j: cubic smoothing function (with three effective df), and ε: error term, follow a standard normal distribution. Further, Monocle performs DE testing of genes across cell groups through LRT statistic(s) through comparing full GLM with additional effects to a reduced GLM based on the NB model. For this purpose, *differentialGeneTest* function implemented in monocle R package [213] is executed.

T-test

T-test [224] is a general-purpose method, used to compare the mean expressions of genes across two cellular groups. Traditionally, the scRNA-seq UMI data violates the T-test's normality assumptions, so, we used TMM method to transform the data. The test statistic for the T-test is expressed as:

$$t_i = \frac{\bar{y}_{i1} - \bar{y}_{i2}}{\sigma_i} \tag{5.36}$$

$$\sigma_i = \sqrt{\frac{S_{i1}^2}{M_1} + \frac{S_{i2}^2}{M_2}} \tag{5.37}$$

where, $\bar{y}_{ik}, S_{ik}^2$ be the mean and variance of the normalized expression values of i^{th} gene for k^{th} (k=1,2) cell group, M_k: number of cells in k^{th} cell group. Empirically, scRNA-seq data are highly (positively) skewed, but the T-test is known to have certain robustness against skewness. Therefore, it is worthy to compare its performance against sophisticated bulk and single-cell methods. This method is executed through *t.test* function implemented in stats R package.

NP methods

EMDomics

EMDomics [227], general-purpose NP method based on Earth Mover's Distance (EMD), developed for DE analysis of genomics data to test the mean expressions difference of genes between two cell groups significantly different from zero. Let, $P_i = \{(p_{i1}, w_{p1}), (p_{i2}, w_{p2}) \ldots, (p_{iM_1}, w_{pM_1})\}$ and $Q_i = \{(q_{i1}, w_{q1}), (q_{i2}, w_{q2}) \ldots, (q_{iM_2}, w_{pM_2})\}$ be the signatures of *i*th gene across two cell groups; p_{im} (*m = 1*, *2*, …, *M*$_1$) and q_{in} (*n = 1*, *2*, …, *M*$_2$) are the centers of *m*th and *n*th histogram in two cell groups; w_{pm} and w_{qn} are weights for *m*th and *n*th cell in two groups. The EMD score for *i*th gene is computed through Eq. 5.38.

$$EMD_i = \frac{\sum_{m=1}^{M_1}\sum_{n=1}^{M_2} f_{mn}^i d_{mn}^i}{\sum_{m=1}^{M_1}\sum_{n=1}^{M_2} f_{mn}^i} \quad (5.38)$$

where, d_{mn}^i: Euclidean distance between *m*th and *n*th cell across two groups for *i*th gene and f_{mn}^i: coefficient of flow from *m*th to *n*th cell for *i*th gene and determined through minimizing the cost function in Eq. 5.39.

$$Cost^i(P, Q, F) = \sum_{m=1}^{M_1}\sum_{n=1}^{M_2} f_{mn}^i d_{mn}^i \quad (5.39)$$

Here, the EMD test statistic reflects the overall difference between two normalized distributions (for two cell groups), usually assessed through statistical significance using permutation test. For this purpose, *calculate_emd* function implemented in EMDomics R package [227] was executed.

NODES

NODES [223], a NP method used for detecting DE genes across two cell groups through using normalized scRNA-seq data. Here, normalization is done through Pseudo-Count Quantile Normalization method [223]. The test statistic for *i*th gene (d_i) is given in Eq. 5.40.

$$d_i = \frac{|\bar{y}_{i1} - \bar{y}_{i2}|}{a_0 + \sigma_i} \quad (5.40)$$

where, a_0: computed as a fixed percentile (*e.g.,* 50th) of the standard errors $\{\sigma_i; i = 1,2, \ldots, N\}$, and $\bar{y}_{i1}, \bar{y}_{i2}, \text{and } \sigma_i$ are defined in Eq. 5.37. The *p-values* for the genes are computed using permutation test through executing the *NODES* function implemented in NODES R package [223].

Wilcoxon signed rank test (Wilcox)

Wilcox method [225] (Mann-Whitney test) is a NP method used to test whether the expression of the genes across the two cell groups significantly different or not. The test's main idea is to compare the ranks for the observations that come from the two cell groups. This rank-based test mostly ignored the magnitude of expression deviation of genes between the two cell groups but maybe a potential method compared to others. To execute this method, we used *wilcox.test* function available in stats R package.

ROTS

Like T-test, Wilcox, and EMDomics, ROTS [226] does not have any single-cell or sequencing-specific functions. It optimizes the parameters among a family of modified t-statistics by maximizing the detections' reproducibility across bootstrap samples. In other words, ROTS maximizes the scaled reproducibility, in Eq. 5.41, over the parameters $\alpha = (\alpha_1, \alpha_2)$; $\alpha_1 \epsilon [0, \infty), \alpha_2 \epsilon \{0,1\}$ and *k* (> 0).

$$\frac{R_k(d_\alpha) - R_k^0(d_\alpha)}{S_k(d_\alpha)} \quad (5.41)$$

where, $S_k(d_\alpha)$: estimated standard deviation of the bootstrap distribution of the observed reproducibility, $R_k(d_\alpha)$ and $R_k^0(d_\alpha)$: reproducibility for observed and

random data. It is calculated as the average reproducibility over randomized data sets, which are permuted from the real samples. The reproducibility is defined as:

$$R_k(d_\alpha) = \frac{1}{B}\sum_{b=1}^{B} R_k^{(b)}(d_\alpha) \quad (5.42)$$

where *B* is the number of bootstrap samples, and $d_\alpha$is the test statistic defined in Eq. 5.40. This method was executed through *ROTS* function implemented in ROTS R package [226].

Comparative Performance Evaluation

Performance metrics

Under this setting, we evaluate the performance of the 19 tested methods for identifying genuine DE genes through 13 performance metrics, such as the number of True Positives (TP) genes, False Positive (FP), True Negative (TN), False Negative (FN), True Positive Rate (TPR), False Positive Rate (FPR), FDR, Positive Prediction Rate (PPV), Negative Prediction value (NPV), Accuracy (ACC), F1 score (F1), and AUROC, defined in Eq. 5.43 – 5.50, and runtime criteria. We evaluate the performance of the 19 methods on 11 real publicly available scRNA-seq datasets (Table 5.3). Further, the performance metrics (Eq. 5.43–5.50) are computed by comparing the DE genes obtained through each method with the reference genes (*i.e.,* true DE genes) for each dataset. For instance, we defined TP in Eq. 5.43 as the genes that are called the true DE genes and FP as the genes that were found significant but were not true DE genes. Similarly, TN were defined as genes that were not true DE and were not found significant, and FN were defined as genes that were true DE but were not found significant.

$$FPR = 1 - Specificity = \frac{FP}{FP+TN} \quad (5.43)$$

$$FDR = \frac{FP}{FP+TP} \quad (5.44)$$

$$PPR = \frac{TP}{TP+FP} \quad (5.45)$$

$$TPR = Sensitivity = \frac{TP}{TP+FN} \quad (5.46)$$

$$NPV = \frac{TN}{TN+FN} \quad (5.47)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.48)$$

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (5.49)$$

AUROC = Area under *Sensitivity* vs. (1-*Specificity*) curve (5.50)

The criteria defined in Eq. 5.43, and 5.44, FP, FN and runtime have "–" impact on the performance of the tested methods, while the criteria in Eq. 5.45 – Eq. 5.50, TP, and TN have "+" impact. For instance, a higher value of ACC or TPR ("+") indicates that the method performs better and *vice-versa*. Similar interpretations can be made for other criteria.

Performance evaluation under MCDM setup

We emphasized to comparative performance analysis of the 19 methods under the simultaneous consideration of all the 13 criteria. In operational research, such a performance evaluation setting is called as MCDM setup [266], where the main idea is to consider a set of criteria and choose the best performing method over a list of methods [267]. Under this MCDM set up, Technique for Order Performance by Similarity to Ideal Solution (TOPSIS) [268] has been extensively used [18]. However, we used this approach for the first time in single-cell data analytics. Here, the basic idea is to choose the best method out of the 19 tested methods based on the simultaneous consideration of the 13 decision criteria, Eq. 5.43-5.50.

Through TOPSIS, it is expected that the best identified method should have shortest geometric distance from the positive ideal solution (PIS) and the longest geometric distance from the negative ideal solution (NIS) [269]. The detailed method and major analytical steps for the MCDM-TOPSIS analysis are given as follows.

Let ***U*** be the resultant decision matrix used under MCDM setup, *i.e.* ***U*** = ((u_{rs})), where u_{rs} represents the value of M_r (r^{th} method) (r = 1, 2, …, 19) under C_s (s^{th} decision criteria) (s = 1, 2, …, 13) and W_s's are the criteria weights indicate the relative importance among them. Further, the W_s are calculated using the entropy technique through the following steps.

Step 1: Normalization of the decision matrix (***U***)**:** The resulted values in ***U*** are first transformed to normalized values (P_{rs}) through: $P_{rs} = {}^{u_{rs}}/_{\sum_{r=1}^{19} u_{rs}}$ (5.51)

Step 2: Calculation of entropy measure (E_s) for s^{th} criterion is calculated using: $E_s = -a \sum_{r=1}^{13} P_{rs} ln\,(P_{rs})$ (5.52)

where a = 1/ln19. Further, the degree of diversity (D_s) for s^{th} criterion can be computed as: $D_s = 1 - E_s$ (5.53)

Step 3: Calculation of weights (W_s) for each criterion: W_s are computed for s^{th} criterion through: $W_s = D_s / \sum_{s=1}^{13} D_s$ (5.54)

After obtaining criteria weights, they are incorporated in the usual TOPSIS technique to calculate the overall scores for each tested method. The major steps for the TOPSIS technique in this context are briefly given as:

[**1**] Construct the normalized decision matrix (**Z**) by vector normalization:

$$z_{rs} = {}^{u_{rs}}/_{\sum_r u_{rs}^2} \quad (5.55)$$

[**2**] Calculate weighted normalized decision matrix using: $v_{rs} = W_s \times z_{rs}$ (5.56)

[**3**] Determine the PIS, $\boldsymbol{V}^+$, and NIS, $\boldsymbol{V}^-$, by using:

$\boldsymbol{V}^+ = \{V_1^+, V_2^+, \ldots, Vs^+\} = \{\langle \max(v_{rs}|r = 1, 2, \ldots, 19)|s \in C_- \rangle, \langle \min(v_{rs}|r = 1, 2, \ldots, 19)|s \in C_+ \rangle\}$

$\boldsymbol{V}^- = \{V_1^-, V_2^-, \ldots, V_S^-\} = \{\langle \max(v_{rs}|r = 1, 2, \ldots, 19)|s \in C_+ \rangle, \langle \min(v_{rs}|r = 1, 2, \ldots, 19)|s \in C_- \rangle\}$

where $C_+ = \{s = 1, 2, \ldots, 8|s$ associated with the criteria having a positive impact and $C_- = \{s = 1, 2, \ldots, 5|s$ associated with the criteria having a negative impact.

[**4**] Calculate the L_2 distance for PIS (d_r^+) and NIS (d_r^-) using:

$$d_r^+ = \left(\sum_{s=1}^{8}(v_{rs} - V_s^+)^2\right)^{1/2} \quad (5.57)$$

$$d_r^- = \left(\sum_{s=1}^{5}(v_{rs} - V_s^-)^2\right)^{\frac{1}{2}} \quad (5.58)$$

[**5**] Determine the relative closeness of the tested method to the ideal solution using Eq. 5.59.

$$R_r = \frac{d_r^-}{d_r^- + d_r^+} \quad \forall\, r = 1, 2, \ldots, 19 \quad (5.59)$$

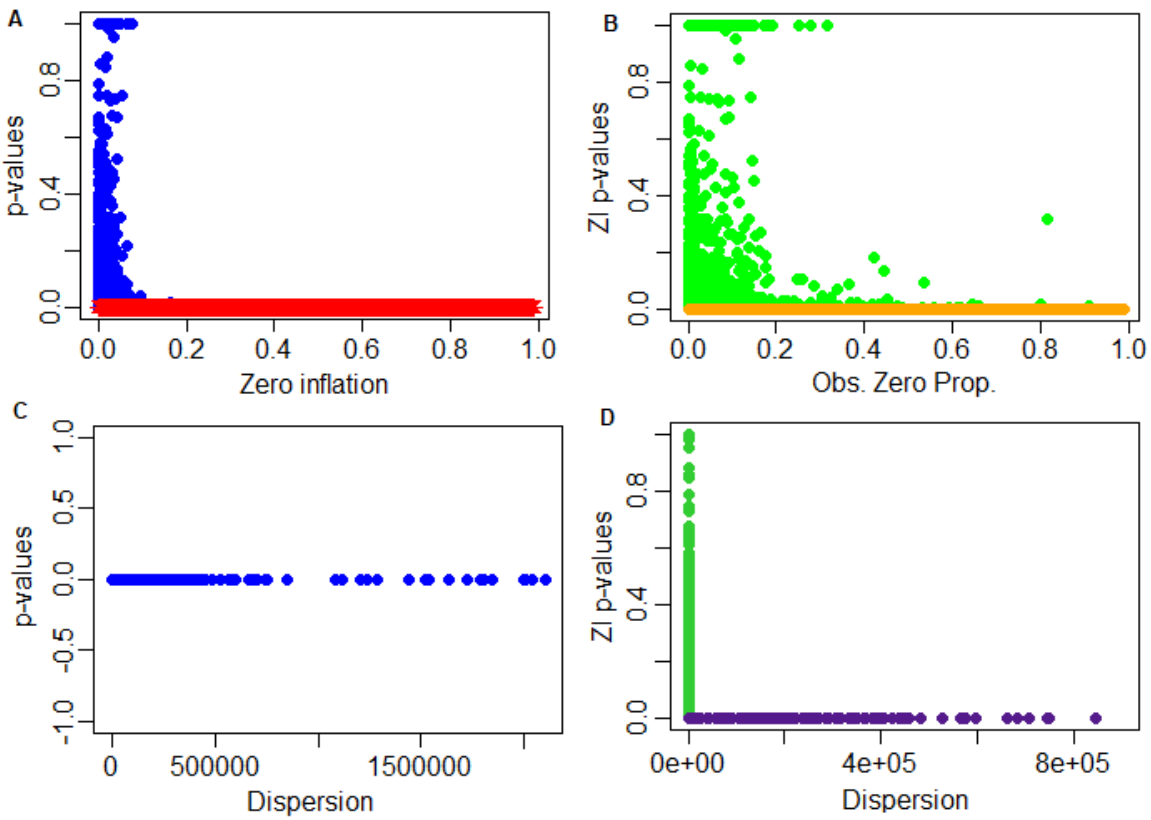
Through this, the methods with higher R_r ($0 \leq R_r \leq 1$) are preferred and considered as better over the multiple criteria and *vice-versa*.

Results and Discussion

Count data models for fitting of scRNA-seq data

The results from the statistical tests for zero inflation and overdispersion are shown in Figure 5.3. The statistical significance values computed for the genes through LRT statistic(s), given in Eq. 5.3, found to be significant for most of the genes. The findings indicated that most of the genes are zero inflated and overdispersed (Figure 5.3).

Figure 5.3. Overdispersion and zero inflation analysis of scRNA-seq data.



Though it was well established, still our analytical findings showed that the scRNA-seq data is zero inflated and overdispersed. Most of the DE methods and tools assume certain count models for fitting the underlying data. Hence, we considered 5 count models, such as NB, ZINB, PD, HD, and ZIPD [270,271] to show their suitability and goodness of fit for zero inflated and overdispersed count (scRNA-seq) data.

Table 5.4. Fitting of well-known discrete models to over-dispersed and zero-inflated cyst count data.

Read	Obs. Freq.	Exp. Freq. NBD	Exp. Freq. ZINBD	Exp. Freq. PD	Exp. Freq. ZIPD	Exp. Freq. HD
0	65	63.29	64.99	25.1	65.03	45.36
1	14	17.56	14.01	37.32	5.1	13.75
2	10	8.98	9.11	27.74	8.87	28.92
3	6	5.72	6.27	13.74	10.28	8.35

4	4	3.91	4.44	5.11	8.93	9.19
5	2	2.79	3.2	1.52	6.21	2.53
6	2	2.04	2.33	0.38	3.6	1.94
7	2	1.52	1.71	0.08	1.79	0.51
8	1	1.15	1.26	0.01	0.78	0.31
9	1	0.88	0.93	0	0.3	0.08
10	1	0.68	0.69	0	0.1	0.04
11	2	0.52	0.52	0	0.03	0.01
12	1	0.41	0.38	0	0.01	0
Total	111	110.95	110.84	111	111.03	110.99
Parameters (MLE)		μ=1.49 θ=0.31	$\mu = 2.285$ $\theta = 0.698$ $\pi = 0.349$	$\mu = 1.486$	$\mu = 3.476$ $\pi = 0.572$	$\mu = 1.487$ $\varphi = 1.796$
#Parameters		2	3	1	2	2
Likelihood		-175.22	-172.8	-263.25	-191.9	-202.84
AIC		354.44	351.60	528.50	387.80	409.68
BIC		354.53	351.74	528.55	387.89	409.77

#Parameters: number of parameters; μ: Mean; θ: size; π: zero-inflation probability; φ: dispersion index (ratio of variance to mean); AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; Obs. Freq: Observed Frequency; Exp. Freq. NBD: computed expected frequency through NB model; Exp. Freq. ZINB: computed expected frequency through ZINB model; Exp. Freq. PD: computed expected frequency through Poisson model; Exp. Freq. ZIPD: computed expected frequency through ZIPD model; Exp. Freq. HD: computed expected frequency through HD model

Table 5.5. Fitting of well-known discrete models to over-dispersed and zero-inflated European red mite data.

Read	Obs. Freq.	NBD	ZINBD	PD	ZIPD	HD
0	70	68.49	69.1	47.65	69	64.65
1	38	37.6	35.01	54.64	28.67	34.71
2	17	20.1	20.65	31.33	25.68	29.02
3	10	12.7	11.21	11.97	15.34	12.25
4	9	5.69	5.91	3.43	6.87	6.07
5	3	3.02	3.06	0.79	2.46	2.14
6	2	1.6	1.57	0.15	0.74	0.81
7	1	0.85	3.79	0.02	1.19	0.25
8	0	0.6	0.1	0.1	0.67	0.02
Total	150	150.65	150.4	150.08	150.62	149.92
Parameter Estimates (MLE)		μ=1.147 $^{\theta}$=1.025	μ=1.283 $^{\theta}$=1.39 Π=0.107	μ =1.146	$\mu = 1.791$ $\pi = 0.367$	$\mu = 1.147$ $\varphi =$ 1.757
#Parameter		2	3	1	2	2

Likelihood	-224.71	-223.43	-242.8	-226.44	-225.98
AIC	453.40	452.80	487.72	456.80	455.95
BIC	453.75	453.33	487.90	457.15	456.31

Our analytical results indicated that the expected frequencies computed from the ZINB were much closer to their observed counterparts, followed by NB models as compared to other models (Tables 5.4, 5.5). Further, the AIC and BIC values for ZINB were lowest followed by NB model for the given zero inflated and over dispersed datasets as compared to PD, ZIPD and HD (Tables 5.4, 5.5). This indicates, for fitting over-dispersed and zero inflated datasets like scRNA-seq, ZINB model provides a better fit as compared to other count models (Figure 5.4).

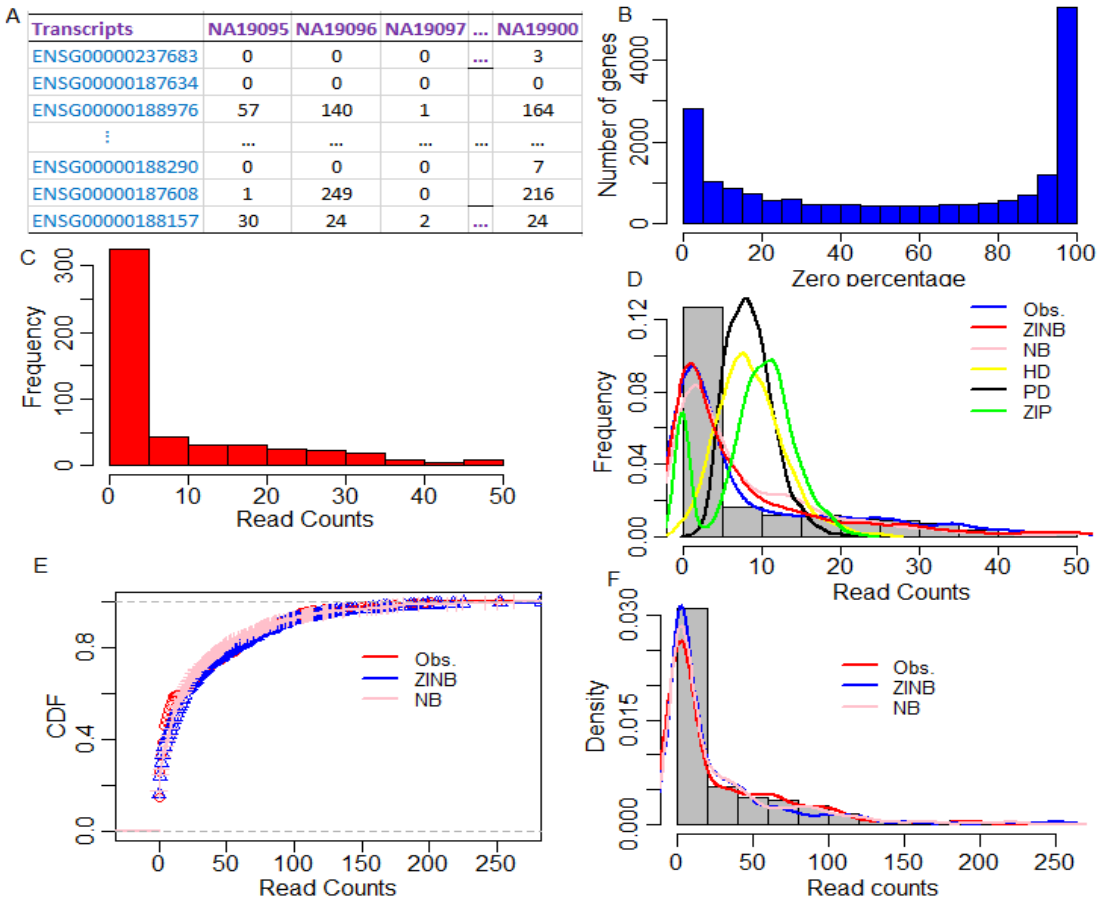


Figure 5.4. Data Characteristics, Distributions and Fitting of Various Count Data Models. (A) Glimpse of the Tung's scRNA-seq (UMI) read count data matrix. Here, rows represent the genes and columns represent the cell lines. The values represent the number of read of

sequences mapped to each gene. (B) Distribution of zero percentages of genes in scRNA-seq data. X-axis represents the various zero percentages and Y-axis represents the number of genes. Here, the approx. (C) Distribution of scRNA-seq read counts of ENSG00000176022 gene (from Tung's data). X-axis represents the reads and Y-axis represents the frequency of the reads. (D) Fitting of Various Discrete Data Models to scRNA-seq read counts of ENSG00000162585 gene (from Tung's data). X-axis represents the read counts and Y-axis represents the density. The fitting of observed density and densities from the NB, ZINB, PD, ZIP and Hermite HD to the observed data are shown in different colors. (E) Cumulative Distribution Function (CDF) plot for scRNA-seq data of ENSG00000176022 gene (Tung's data). Here, X-axis represents the read counts and Y-axis represents the cumulative density of read counts. Observed CDF, and CDFs from NB and ZINB models are shown. (F) Density plots for scRNA-seq data of ENSG00000176022 gene (Tung's data). Observed density plot, and density plots from NB and ZINB models are shown.

At this stage, we inferred that ZINB and NB model best suit for fitting the scRNA-seq count data as compared to other models (Tables 5.4, 5.5, Figure 5.4). To be more specific, we also tested the NB and ZINB models' ability to estimate the mean and dispersion parameters for scRNA-seq data through simulation. The results are shown in Table 5.6. Our analytical results indicated that the NB model underestimated the mean and overestimated the dispersion parameter for scRNA-seq data (Table 5.6).

Table 5.6. Comparative analysis of NBD and ZINBD for estimation of parameters from Tung's data.

		NBD				ZINBD			
Parameter	**True value**	**MLE**	**Bias**	**MSE**	**95% CI**	**MLE**	**Bias**	**MSE**	**95% CI**
Mean (μ)	**2.28**	1.483	-0.797	0.649	(1.325, 1.641)	2.328	0.046	0.132	(2.254, 2.410)
Dispersion (θ^{-1})	**1.45**	3.315	1.865	3.597	(2.943, 3.687)	1.433	0.048	0.263	(1.333, 1.534)
Zero inflation prob. (π)	**0.35**	-	-	-	-	0.353	0.003	0.011	(0.331, 0.371)

Number of cells: 500; number of simulations: 100; MSE: Mean Standard Error; CI: Confidence Interval

Contrarily, the ZINB model provided better estimates of mean and dispersion, which are close to their true values for scRNA-seq data. Further, ZINB model has lower bias and MSE as compared to NBD model (Table 5.6). It is interesting to note that 95 % confidence interval of parameters for NBD does not

contain the true values of the parameters. While this observation was quite satisfactory for ZINB model. This indicated the better suitability of ZINB model for modeling the zero inflated and overdispersed scRNA-seq count data and provides better estimates of the parameters. The reason may be attributed as NBD thus accommodates excess zeros by underestimating the mean and overestimating the dispersion parameters. This phenomenon may jeopardize the statistical power of NBD based DE tools to discover DE genes in the presence of zero inflation, when applied to scRNA-seq data.

Comparative performance analysis of scRNA-seq DE methods

We compared the performance of the 19 methods for detecting DE genes on 11 real publicly available scRNA-seq datasets (Table 5.3) under the condition of comparing two groups of cells. However, the real studies involve more complex experimental designs, which some of the tested methods do not accommodate. Specifically, the T-test, Wilcox, ROTS, DEsingle, scDD, NODES are limited to two-group comparisons, whereas EMDomics can perform a limited number of analysis types. The remaining methods implement statistical frameworks that can accommodate more complex designs, including comparison across the multiple cellular groups, accommodation of cell covariates and adjustments for batch effects, and cell capture rates. To make the comparisons fully reproducible, we provide the R codes, processed scRNA-seq datasets, and reference genes in https://github.com/sam-uofl/RoopSeq.

Comparative assessment based on performance metrics

The single-cell datasets and their respective comparison designs were used to detect the DE genes through each of the 19 tested methods. For instance, Islam data [196], the experimental design involves DE analysis of genes between 48 mouse embryonic stem cells and 44 mouse embryonic fibroblast cells through the methods. In other words, we selected the DE gene sets of sizes 200, 400, …, 3000 through the tested methods from the Islam data (Table 5.3). Then, the performance metrics such as TP, FP, PPR, TPR, FPR, ACC, and F1, were computed by comparing the detected DE genes with the reference genes for each dataset, and the results are shown in Figures 5.5 – 5.15.

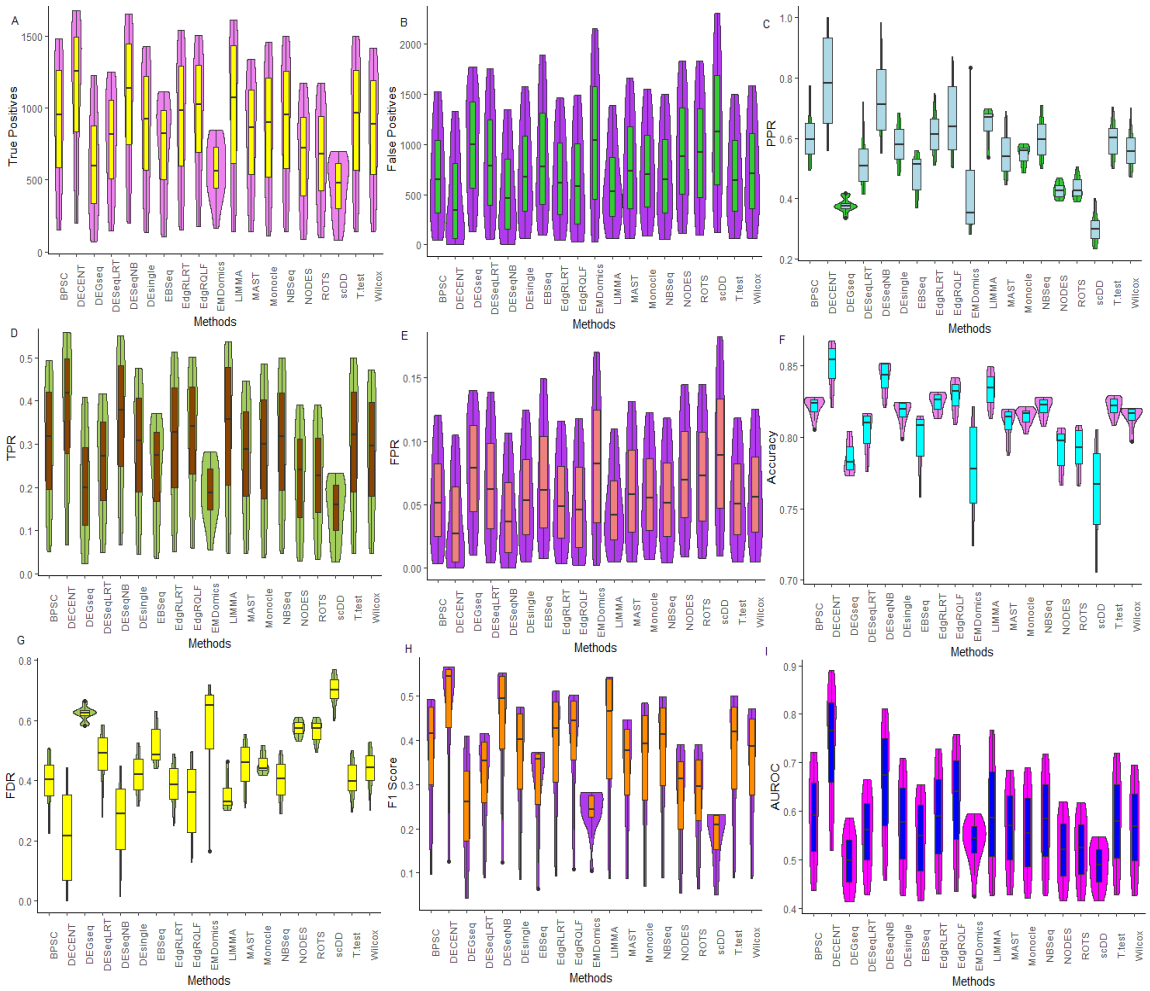


Figure 5.5. Comparative performance evaluation of the methods through the performance metrics for Soumillion2 data. The tested methods are evaluated on

the Soumillion2 data through performance metrics, *i.e.*, TP, FP, TPR, FPR, PPR, FDR, Accuracy, F1 score, and AUROC. The 19 tested methods are shown in the X-axis. The Violin plots are shown for the evaluation of the methods through (A) TP; (B) FP; (C) PPR; (D) TPR; (E) FPR; (F) Accuracy; (G) FDR; (H) F1 score; and (I) AUROC. The violin plot shows the full distribution of the performance metrics computed through each tested method. The box represents inter-quartile range, the horizontal line represents median, the bars on the boxes shows the 1.5 x inter-quartile range.

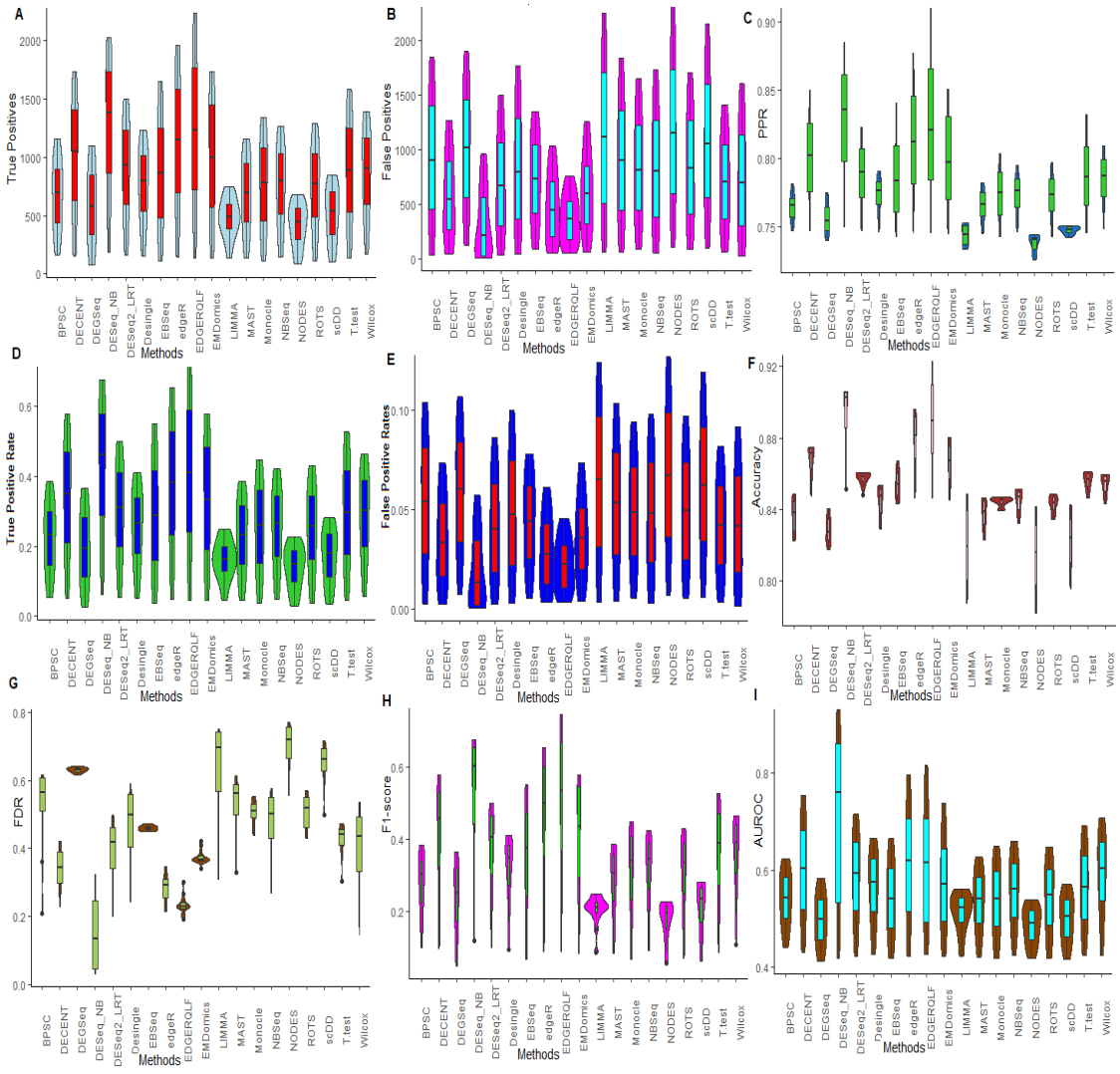


Figure 5.6. Comparative performance evaluation of the methods through the performance metrics for Islam data. The tested DE methods are evaluated through the performance evaluation metrics, such as TP, FP, TPR, FPR, PPR, FDR, Accuracy, F1 score, and AUROC. The 19 tested methods are shown in the X-axis. The Violin plots are shown for comparative evaluation of tested methods through (A) TP; (B) FP; (C) PPR; (D) TPR; (E) FPR; (F) Accuracy; (G) FDR; (H) F1 score; and (I) AUROC. The violin plot shows the full distribution of the performance metrics computed through each tested method. The box represents inter-quartile range, the horizontal line represents median, the bars on the boxes shows the 1.5 x inter-quartile range.

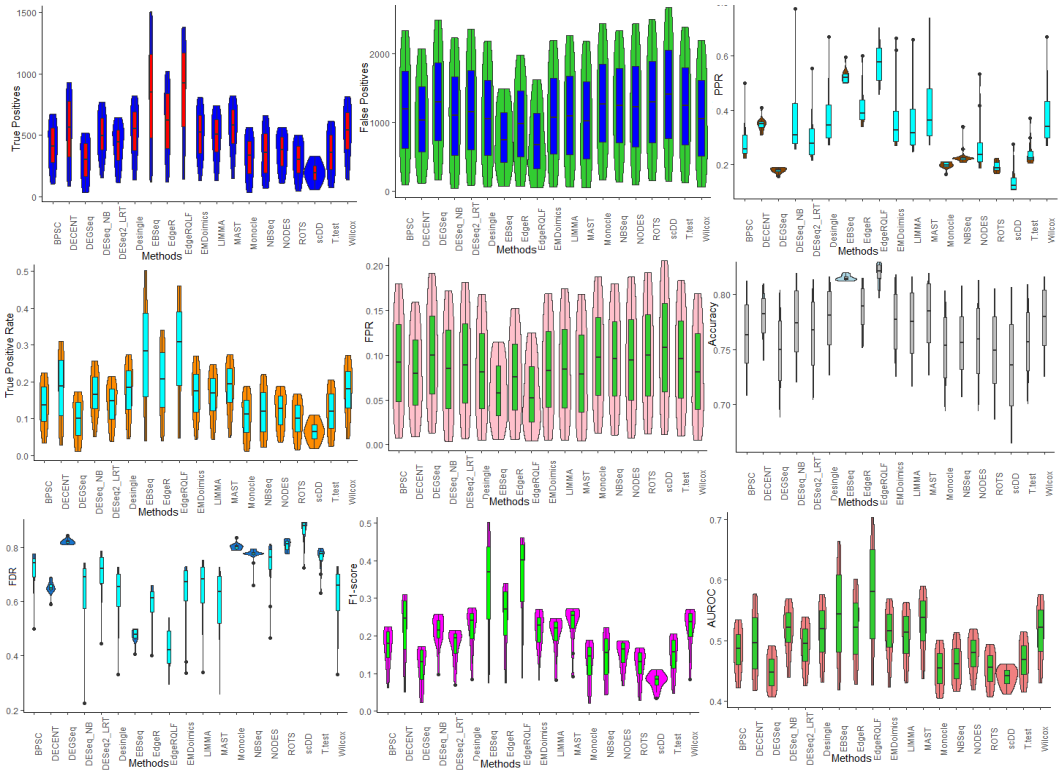


Figure 5.7. Comparative performance evaluation of the DE methods through the performance metrics for Tung data.

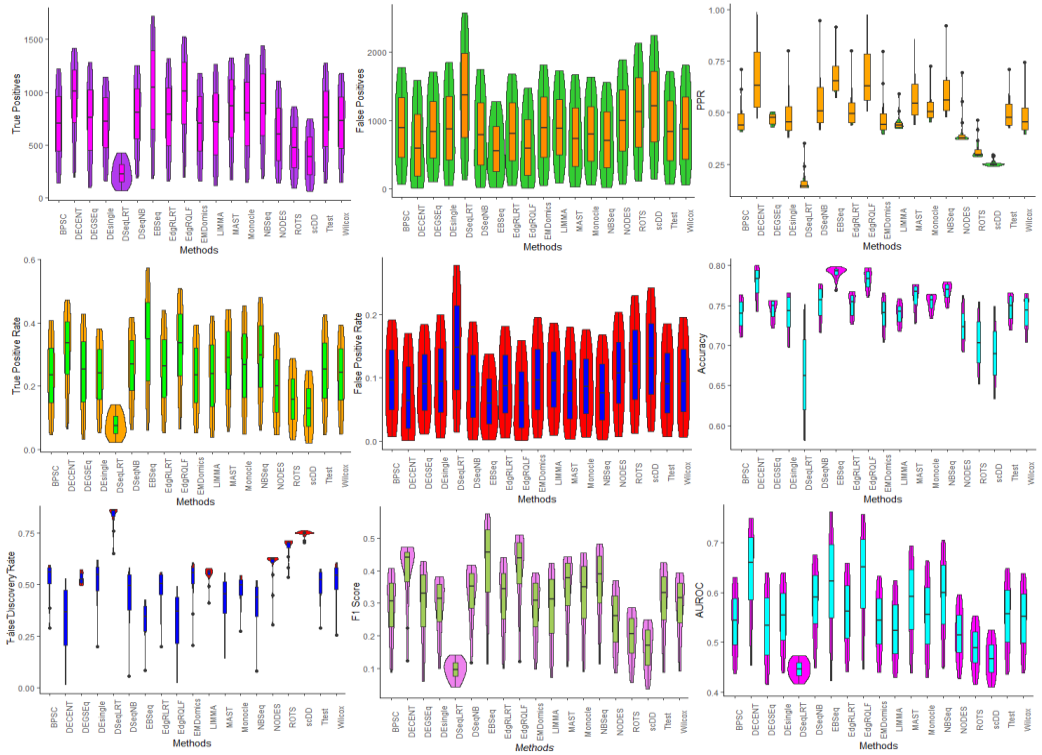


Figure 5.8. Comparative performance evaluation of the DE methods through the performance metrics for Soumillon1 data.

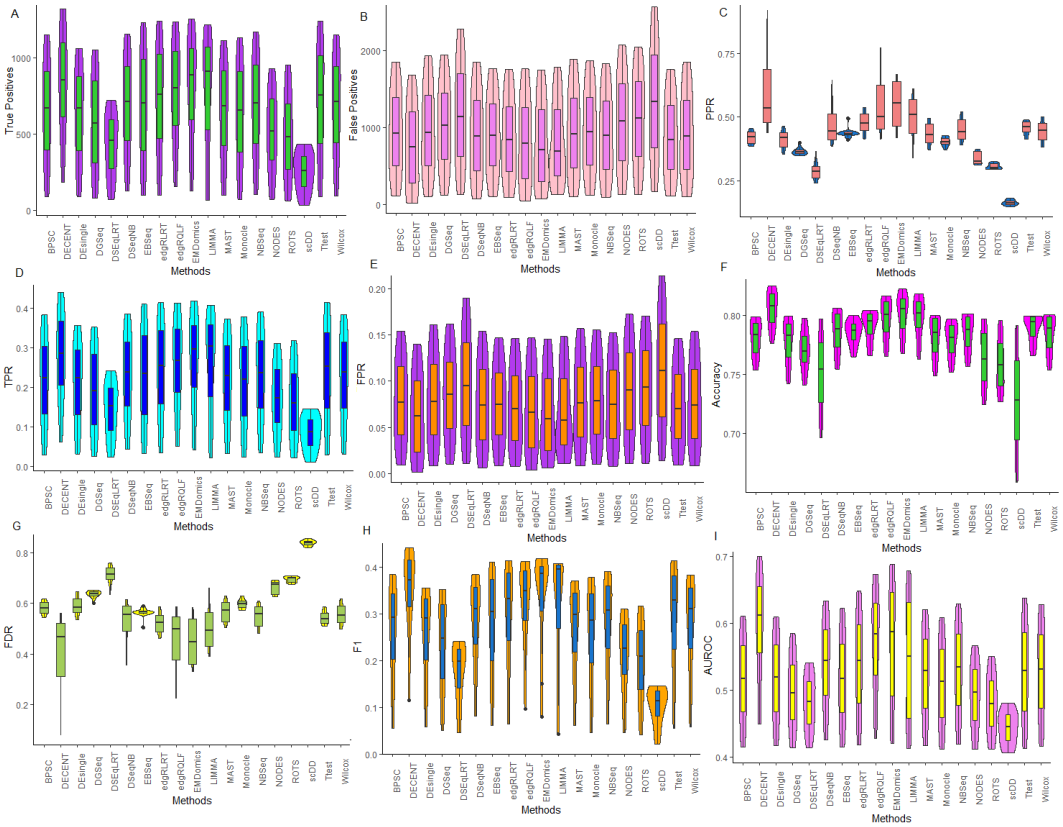


Figure 5.9. Comparative performance evaluation of the DE methods on Soumillon3 data.

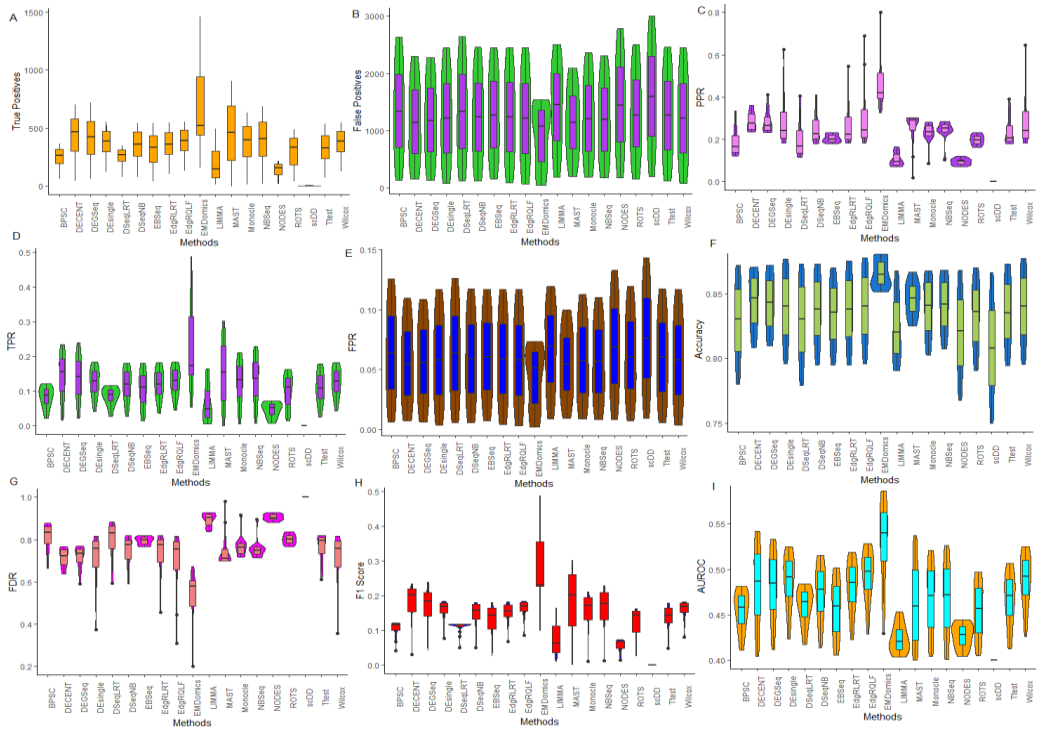


Figure 5.10. Comparative performance evaluation of the DE methods on Klein data.

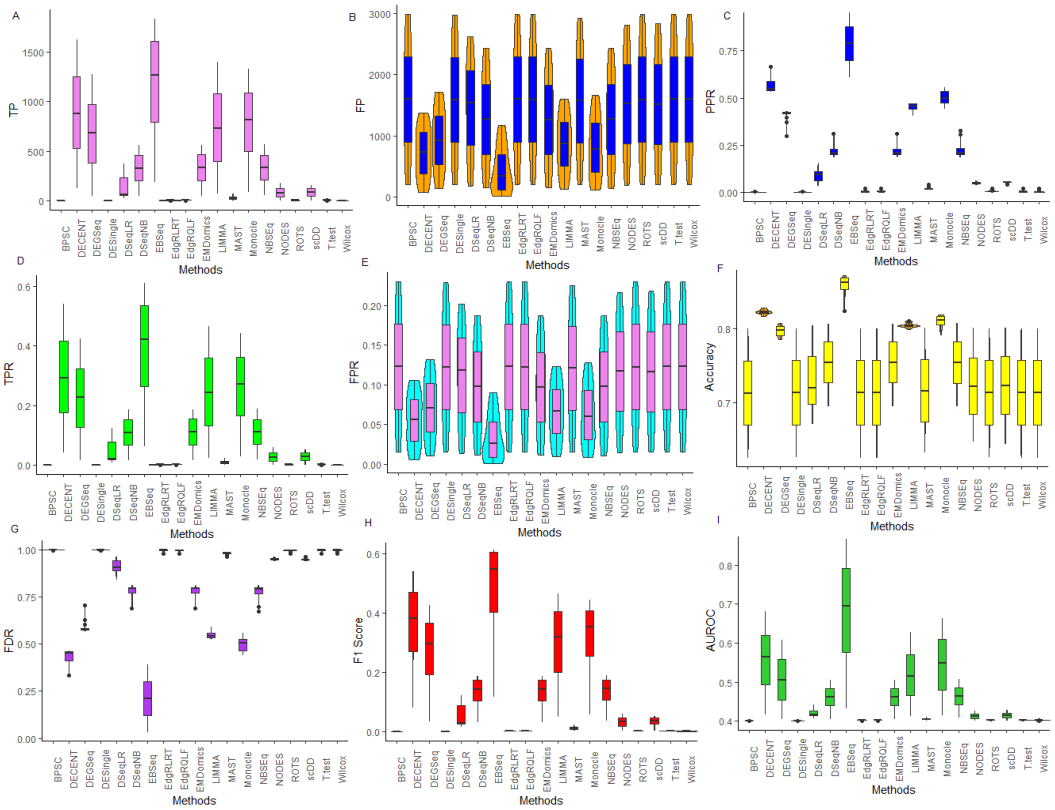


Figure 5.11. Comparative performance evaluation of the DE methods on Gierahn data.

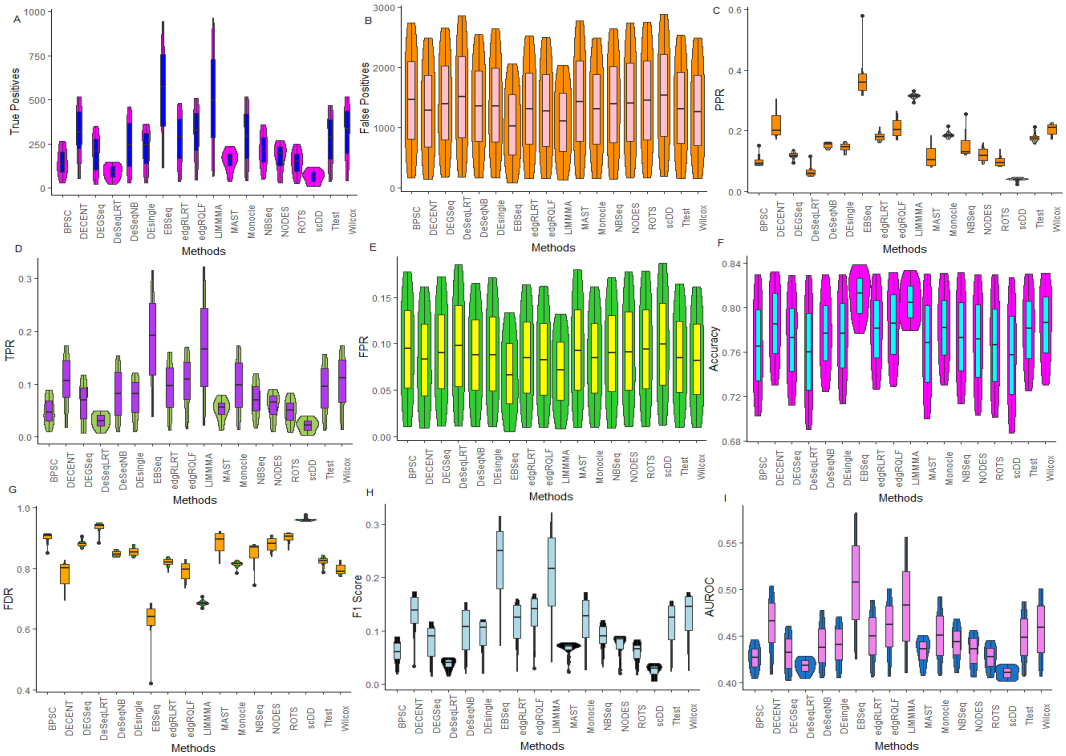


Figure 5.12. Comparative performance evaluation of the DE methods on Chen data.

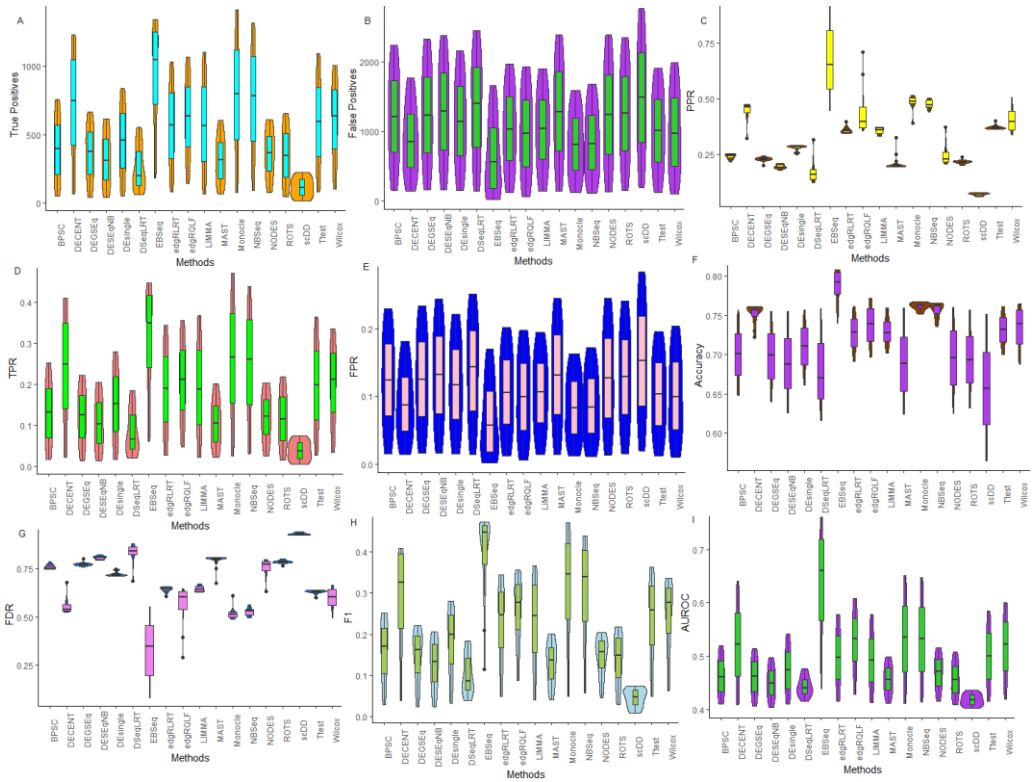


Figure 5.13. Comparative performance evaluation of the DE methods on Savas data.

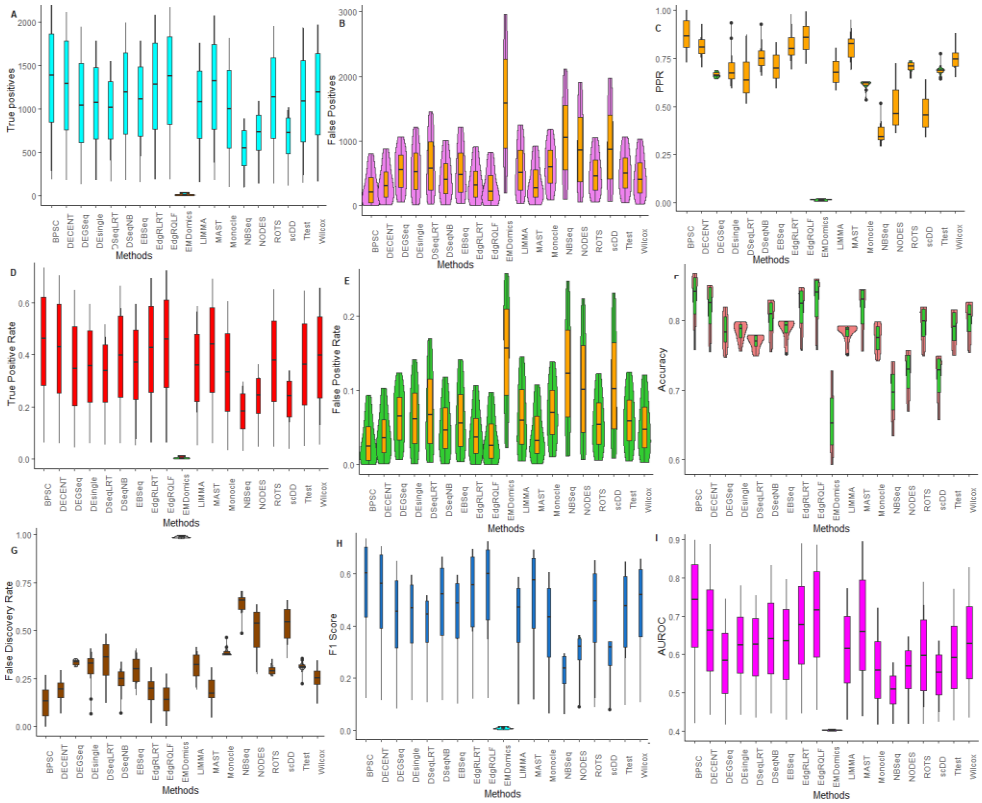


Figure 5.14. Comparative performance evaluation of the DE methods on Grun data.

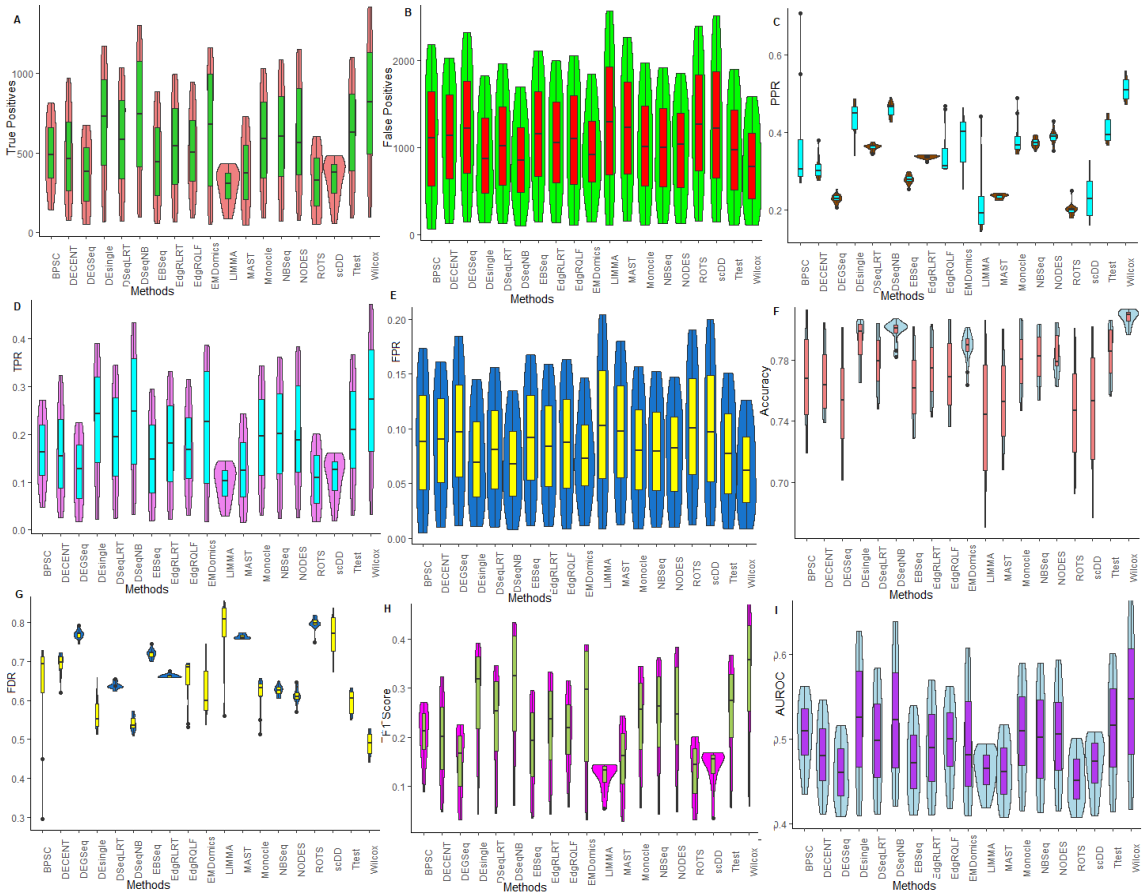


Figure 5.15. Comparative performance evaluation of the methods through the performance metrics for Zigenhein data. The tested DE methods are evaluated on the Zigenhein scRNA-seq data through the performance evaluation metrics, such as TP, FP, TPR, FPR, PPR, FDR, Accuracy, F1 score, and AUROC. The 19 tested methods are shown in the X-axis. The Violin plots are shown for comparative evaluation of tested methods through (A) TP; (B) FP; (C) PPR; (D) TPR; (E) FPR; (F) Accuracy; (G) FDR; (H) F1 score; and (I) AUROC. The violin plot shows the full distribution of the performance metrics computed through each tested method. The box represents inter quartile range, the horizontal line represents median, the bars on the boxes shows the 1.5 x inter-quartile range.

In this comparison setting for Soumillon2 data, the DECENT provided the highest (median) TP values, followed by DESeqNB, LIMMA, edgeRQLF (Figure 5.5A). Similar findings were observed when assessed through TPR. Further, we found lowest values of the FP and FPR for these methods compared to others (Figure 5.5B). For instance, for the DE gene set of size 3000, the DECENT detected 1674 genes as truly DE, followed by DESeqNB (1653) and LIMMA (1612)

(Table 5.7). In other words, DECENT detected the fewer FP genes with higher probabilities along with DESeqNB, LIMMA, and edgeRQLF as compared to others. The accuracy-based performance analysis of the tested DE methods indicated that the DECENT was found to detect true (both positive and negative) genes more accurately, followed by DESeqNB and edgeRQLF compared to others (Figure 5.5C). Among the tested methods, EMDomics, and scDD were found to have the lowest rates of sensitivities and specificities for detecting true DE genes, therefore performed worst for Soumillon2 data (Figure 5.5). Similar interpretations can be made about the tested methods through PPR and F1 score (Figure 5.5, Table 5.7).

Table 5.7. Evaluation of DE methods based on performance metrics for Soumillon2 data.

Methods	TP	FP	TPR	FPR	FDR	PPR	NPV	ACC	F1	AUC
				DEG = 1000						
BPSC	639	361	0.213	0.029	0.361	0.639	0.839	0.826	0.320	0.529
DECENT	914	86	0.305	0.007	0.086	0.914	0.857	0.861	0.457	0.641
DEGseq	368	632	0.123	0.050	0.632	0.368	0.820	0.791	0.184	0.461
DESeqNB	813	187	0.271	0.015	0.187	0.813	0.851	0.848	0.407	0.586
DESeqLRT	555	445	0.185	0.035	0.445	0.555	0.833	0.815	0.278	0.509
DEsingle	620	380	0.207	0.030	0.380	0.620	0.837	0.823	0.310	0.513
EBSeq	553	447	0.184	0.035	0.447	0.553	0.833	0.815	0.277	0.487
edgeRLRT	650	350	0.217	0.028	0.350	0.650	0.839	0.827	0.325	0.525
edgeRQLF	761	239	0.254	0.019	0.239	0.761	0.847	0.842	0.381	0.557
EMDomics	461	539	0.154	0.043	0.539	0.461	0.827	0.803	0.231	0.521
LIMMA	683	317	0.228	0.025	0.317	0.683	0.842	0.832	0.342	0.516
MAST	591	409	0.197	0.032	0.409	0.591	0.835	0.820	0.296	0.511
Monocle	573	427	0.191	0.034	0.427	0.573	0.834	0.817	0.287	0.497
NBSeq	638	362	0.213	0.029	0.362	0.638	0.839	0.826	0.319	0.519
NODES	433	567	0.144	0.045	0.567	0.433	0.825	0.800	0.217	0.474
ROTS	472	528	0.157	0.042	0.528	0.472	0.827	0.805	0.236	0.479
scDD	325	675	0.108	0.053	0.675	0.325	0.817	0.786	0.163	0.459
T-test	627	373	0.209	0.030	0.373	0.627	0.838	0.824	0.314	0.514
Wilcox	594	406	0.198	0.032	0.406	0.594	0.836	0.820	0.297	0.509
				DEG = 2000						
BPSC	1131	869	0.377	0.069	0.435	0.566	0.863	0.825	0.452	0.631
DECENT	1400	600	0.467	0.047	0.300	0.700	0.883	0.859	0.560	0.784
DEGseq	746	1254	0.249	0.099	0.627	0.373	0.835	0.776	0.298	0.523

DESeqNB	1341	659	0.447	0.052	0.330	0.671	0.878	0.852	0.536	0.716
DESeqLRT	948	1052	0.316	0.083	0.526	0.474	0.850	0.801	0.379	0.595
DEsingle	1107	893	0.369	0.071	0.447	0.554	0.861	0.822	0.443	0.618
EBSeq	928	1072	0.309	0.085	0.536	0.464	0.848	0.799	0.371	0.588
edgeRLRT	1167	833	0.389	0.066	0.417	0.584	0.866	0.829	0.467	0.634
edgeRQLF	1189	811	0.396	0.064	0.406	0.595	0.867	0.832	0.476	0.679
EMDomics	661	1339	0.220	0.106	0.670	0.331	0.828	0.765	0.264	0.557
LIMMA	1321	679	0.440	0.054	0.340	0.661	0.877	0.849	0.528	0.633
MAST	1028	972	0.343	0.077	0.486	0.514	0.855	0.812	0.411	0.606
Monocle	1080	920	0.360	0.073	0.460	0.540	0.859	0.818	0.432	0.597
NBSeq	1138	862	0.379	0.068	0.431	0.569	0.863	0.826	0.455	0.625
NODES	836	1164	0.279	0.092	0.582	0.418	0.841	0.787	0.334	0.554
ROTS	841	1159	0.280	0.092	0.580	0.421	0.842	0.788	0.336	0.551
scDD	561	1439	0.187	0.114	0.720	0.281	0.821	0.752	0.224	0.508
T-test	1144	856	0.381	0.068	0.428	0.572	0.864	0.827	0.458	0.623
Wilcox	1073	927	0.358	0.073	0.464	0.537	0.859	0.817	0.429	0.607
DEG = 3000										
BPSC	1478	1522	0.493	0.120	0.507	0.493	0.880	0.805	0.493	0.722
DECENT	1674	1326	0.558	0.105	0.442	0.558	0.895	0.830	0.558	0.857
DEGseq	1228	1772	0.409	0.140	0.591	0.409	0.860	0.773	0.409	0.585
DESeqNB	1653	1347	0.551	0.107	0.449	0.551	0.893	0.828	0.551	0.811
DESeqLRT	1247	1753	0.416	0.139	0.584	0.416	0.861	0.776	0.416	0.666
DEsingle	1428	1572	0.476	0.124	0.524	0.476	0.876	0.799	0.476	0.709
EBSeq	1110	1890	0.370	0.150	0.630	0.370	0.850	0.758	0.370	0.654
edgeRLRT	1537	1463	0.512	0.116	0.488	0.512	0.884	0.813	0.512	0.729
edgeRQLF	1506	1494	0.502	0.118	0.498	0.502	0.882	0.809	0.502	0.758
EMDomics	844	2156	0.281	0.171	0.719	0.281	0.829	0.724	0.281	0.594
LIMMA	1612	1388	0.537	0.110	0.463	0.537	0.890	0.822	0.537	0.768
MAST	1337	1663	0.446	0.132	0.554	0.446	0.868	0.787	0.446	0.685
Monocle	1454	1546	0.485	0.122	0.515	0.485	0.878	0.802	0.485	0.691
NBSeq	1497	1503	0.499	0.119	0.501	0.499	0.881	0.808	0.499	0.718
NODES	1173	1827	0.391	0.145	0.609	0.391	0.855	0.766	0.391	0.620
ROTS	1170	1830	0.390	0.145	0.610	0.390	0.855	0.766	0.390	0.618
scDD	697	2303	0.232	0.182	0.768	0.232	0.818	0.705	0.232	0.547
T-test	1501	1499	0.500	0.119	0.500	0.500	0.881	0.808	0.500	0.719
Wilcox	1413	1587	0.471	0.126	0.529	0.471	0.874	0.797	0.471	0.695

TP: True Positives; FP: False Positives; TN: True Negatives; FN: False Negatives; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F1 score; AUC: Area Under Receiver Operating Curve

For Islam data, the edgeRQLF and DESeqNB methods had the highest TP values with lower FP, followed by edgeRLRT, DECENT and EMDomics methods

(Figure 5.6). The performance of the methods such as NODES, LIMMA, and DEGseq was extremely poor in terms of lower TP and higher FP values (Figure 5.6). This finding indicated these methods detected fewer true DE genes with higher probability. Among the single-cell methods, DECENT's performance was found to be superior, followed by DEsingle and BPSC (Figure 5.6). Further, the median TPR value for DESeqNB and edgeRQLF was found to be highest, followed by edgeRLRT and DECENT (Figure 5.6). This observation indicated that these tested methods identified genes that are truly DE at higher rates compared to others. The lower values of FPR also indicated the better performance of edgeRQLF, DESeqNB, edgeRLRT, and DECENT over other methods. In other words, these methods had higher probabilities of detecting a lower number of FP genes. While LIMMA, NODES, scDD, and DEGseq had the highest numbers and rates of FP genes compared to other tested methods. Similar interpretations can be made for all the tested methods through other performance metrics, such as PPR, ACC, NPV, and F1 measures (Figure 5.6).

For other datasets, such as Tung, Chen, Savas, Soumillon1, Grun, Ziegenhain, Soumillon3, Gierahn, and Klein, a similar interpretation can be made for the tested methods (Figures 5.7 – 5.15). It can be observed that the performance of the tested methods varies differently across the datasets when assessed through each of the 10-performance metrics. For instance, EBSeq, edgeRQLF performed better for Tung data, while DECENT, EMDomics, provided better results for Soumillon3 data. In other words, the tested methods' performance was mostly data specific (no method best fit for all datasets) when assessed

through individual performance metrics. However, we found that bulk RNA-seq methods are quite competitive and even performed better (for some cases) than single-cell methods under the two cellular groups comparison.

Performance assessment based on ROC

Under this comparison setting, the performance of the DE methods was tested on multiple real datasets through AUROC, and the results are shown in Figures 5.5 – 5.15. For Soumillon2 data, DECENT provided the highest AUROC values, followed by DESeqNB, LIMMA, and edgeRQLF (Figure 5.5l), which indicated that the underlying model of DECENT, *i.e.,* ZINB fits well to the underlying data (Figure 5.5). For instance, for DE gene set size 3000, an AUROC value of 0.857 was observed for DECENT followed by DESeqNB (0.811), LIMMA (0.768), and edgeRQLF (0.758) (Table 5.7). In other words, the DECENT has higher sensitivity and specificity rates to detect true DE genes in real Soumillon2 data compared to other methods. The single-cell specific tools scDD and MAST performed worst in this comparison, while DEGSeq followed by EBSeq and DESeqLRT showed the overall poor performance among bulk RNA-seq methods along with EMDomics (Figure 5.5l, Table 5.7).

For Islam data, the DESeqNB method produced the highest AUROC value, followed by edgeRQLF, edgeRLRT, DECENT, and EMDomics (Figure 5.6). The lowest AUROC values were observed for NODES, LIMMA, scDD, and DEGseq with higher probabilities than others (Figure 5.6). Here, the bulk RNA-seq methods performed extremely well, even better than single-cell tools like DECENT, Monocle, MAST, *etc.* The simple methods, such as T-test, and Wilcox, performed

relatively well with moderately high sensitivities and specificities for detecting DE genes (Figure 5.6). Similar interpretations can be made for the other 9 datasets based on the AUROC from Figures 5.7 – 5.15. Through sensitivity-specificity analysis, it was observed that the performance of the tested methods varies differently across the considered real datasets. For instance, EMDomics and MAST performed very well in Klein data, while LIMMA and EBSeq were better suited to Chen data. However, scDD, NODES, and ROTS consistently performed worst across all the datasets, while methods such as DEGSeq, DESeqLRT, and MAST performed very badly for some of the datasets. Similar conclusions can be made for other datasets. Hence, it can be noted that a single method may not be chosen to provide the best results for DE analysis in scRNA-seq and mostly depend on the data and performance evaluation metric.

Performance assessment based on FDR rates

The results from the tested methods' performance assessed through FDR across the 11 datasets are shown in Figures 5.5G – 5.15G. For Soumillon2 data, DECENT's median FDR value was found to be lowest, followed by DESeqNB and edgeRQLF (Figure 5.5G). For instance, DE gene set 3000; the FDR value was observed to be 0.442 for DECENT and 0.448 for DESeqNB, whereas methods including scDD, EMDomics provided the highest FDR values (Table 5.7). This indicates that the UMI-based specialized DECENT tool's performance was superior and robust compared to count-based bulk RNA-seq tools (Figure 5.5G). Further, normalized data-based tools, *i.e.,* scDD, EMDomics and ROTS, performed not so well in terms of robustness for detecting true DE genes.

For Islam data, the findings indicated that the performance of DESeqNB, edgeRQLF, edgeRLRT, and DECENT was observed to be robust among the competitive methods (Figure 5.6G). Specifically, DECENT's performance was better and robust among the single-cell methods, followed by Monocle and MAST. However, bulk RNA-seq methods, such as DESeqNB, edgeRQLF, and edgeRLRT, performed better and robust even compared to single-cell methods for Islam data having fewer cells (Figure 5.6G). Further, among all methods, DEGSeq, LIMMA, and NODES performed worst in terms of robustness for detecting true DE genes. Similar interpretations can be made for other datasets through the computed FDR metric (Figures 5.7G -5.15G). Through such analysis, we observed that the tested methods' performance varied differently across the real datasets for detecting robust DE genes (Table 5.8). For instance, EBSeq performed well in terms of robustness for Savas, Soumillon1, Chen, Geinhein, and Tung data but performed poorly in the remaining datasets. While scDD, NODES, and ROTS consistently worst performed methods over the datasets (Table 5.8). Hence, we can infer that not a single method was found globally best for robust DE analysis scRNA-seq data, and mostly the performance is data specific.

Table 5.8. Ranking of DE methods across all datasets based on the FDR metric.

Methods	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	Rank Score
BPSC	15	12	11.5	1	14	16	12	12	6	16.5	17.5	4.55
DECENT	4	5	13	5	3	2	4	1	1	3	2	9.32
DEGSeq	16	18	16	14	11	4	14	15	18	13	5	4
DESeqLRT	6	11	9	13	19	15	18	18	13	18	9	3.74
DESeqNB	1	8.5	2	6	6	11	17	7	2	10	7	7.5
DEsingle	10	6.5	3	10	10	8	11	13	9	11	17.5	5.84
EBSeq	9	2	14	8	1	13	1	10	14	1	1	7.68
edgeRLRT	3	3	10	4	8	10	8.5	5	5	7	14.5	7.47
edgeRQLF	2	1	11.5	2	2	5	5	4	4	4.5	14.5	8.66

EMDomics	5	8.5	4	19	13	1	10	2	17	9	7	6.55
LIMMA	18	10	19	11	15	17	8.5	3	3	2	4	5.76
MAST	14	4	15	3	5	3	16	11	12	15	12	5.79
Monocle	12	16	7.5	15	7	9	2	14	11	6	3	6.18
NBSeq	11	15	7.5	18	4	6	3	8	8	12	7	6.34
NODES	19	13	6	16	16	18	13	16	15	14	11	3.32
ROTS	13	17	18	9	17	14	15	17	16	16.5	14.5	2.79
scDD	17	19	17	17	18	19	19	19	19	19	10	1.42
Ttest	8	14	5	12	9	12	7	6	7	8	14.5	6.18
Wilcox	7	6.5	1	7	12	7	6	9	10	4.5	19	6.89

D1: Islam; D2: Tung; D3: Zigenhein; D4: Grun; D5: Soumillion1; D6: Klein; D7: Savas; D8: Soumillion3; D9: Soumillion2; D10: Chen; D11: Geinhein

Performance assessment based on runtime

Under this setting, we evaluated the tested methods' performance based on runtime criterion, where the runtime refers to the computational time required to analyze the data. Through this, the method which requires less runtime was considered to be better and *vice-versa*. To measure this, we ran the code written in R (*v* 4.0.2) for each tested method by following the instructions and recommendations of their respective R software packages. The required average CPU time (over 50 runs for each program) was observed for each of the methods for analyzing a count data with 10000 genes with 500 cells (250 cells in each cell group). All these analyses were performed on a 16 GB RAM computer with Windows 10 OS and Intel(R) Core (TM) i3-6100U CPU clock rate as 2.93 GHz. It was found that DECENT is the slowest and more computationally intensive method followed by DEsingle due to the implementation of an iterative ECM algorithm to estimate the model parameters. For instance, the UMI data (10000 genes over 500 cells), DECENT took ~20 hours, followed by DEsingle (~12 hours) to detect the DE genes. Among the methods, T-test, Wilcox, and LIMMA are the fastest to run; MAST, edgeR, and DESeq are also relatively fast. Further, the methods, such as

EBSeq, ROTS, EMDomics, and NODES, are relatively computationally intensive due to the implementation of permutation and bootstrap procedures, which are usually time-consuming processes. The remaining methods do not include any heavily time-consuming steps, therefore considered as computationally efficient.

Performance assessment based on MCDM-TOPSIS analysis

We observed conflicts among the 13 criteria through which the tested methods' performance was assessed. For instance, DECENT performed better among the methods specially designed for single cell studies, when assessed through most of the performance metrics (Figure 5.5-5.15) but performed worst based on runtime criterion. Due to such conflicts in the performance evaluation, the TOPSIS approach was necessary to choose the best option over the available 19 options under the simultaneous consideration of 13 decision criteria (*i.e.,* MCDM). The results from the MCDM-TOPSIS analyses are shown in Figures 5.16 – 5.26.

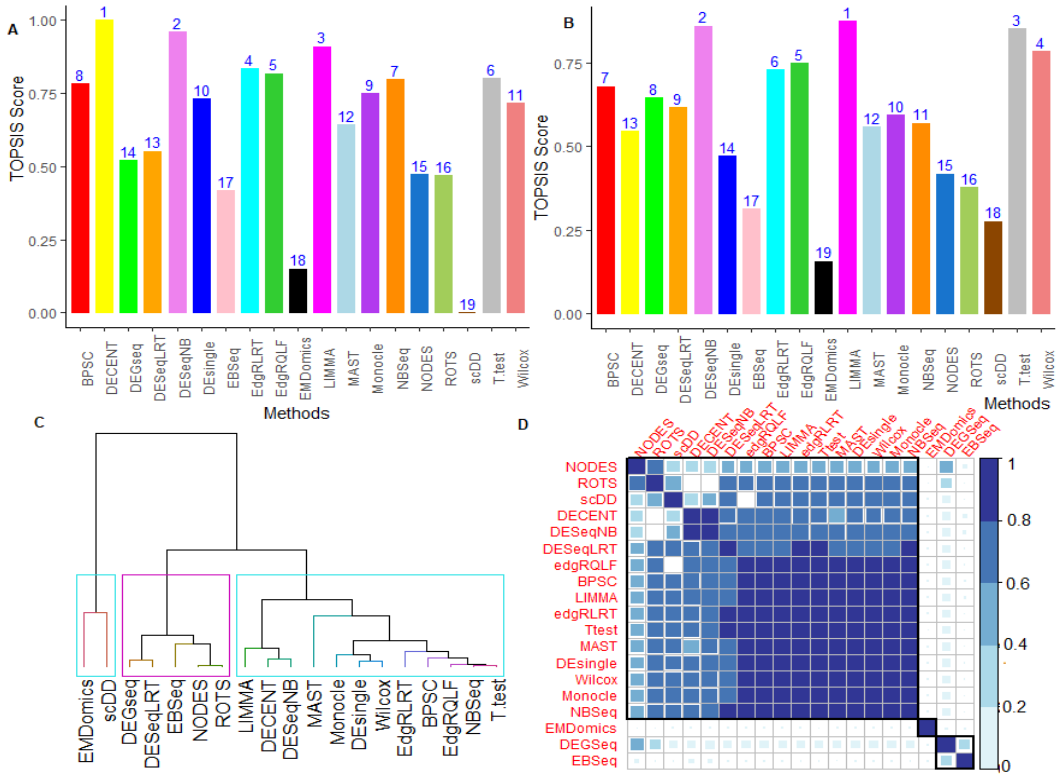


Figure 5.16. Performance evaluation of DE methods under MCDM setup for Soumillion2 data**.** The results are shown for (A) MCDM-TOPSIS analysis of the methods are shown for 13 performance metrics including runtime criterion; (C) Average similarities between the evaluated DE methods based on the 13-performance metrics. The dendrogram was obtained by average-linkage hierarchical clustering; (B) MCDM-TOPSIS analysis of the DE methods based on 12 performance metrics excluding runtime criterion; (D) Similarity analysis among methods based on their ability to detect common DE genes. The p-values for each comparison were computed through Binomial test (color shows value significant at 1% level of significance and white cells represents non-significant values).

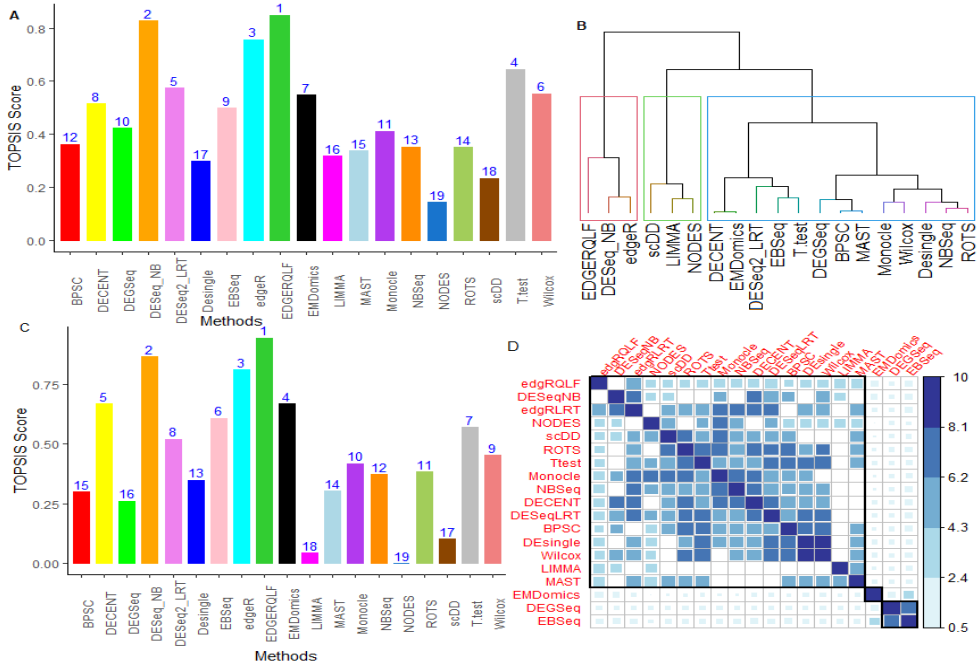


Figure 5.17. Performance evaluation of the methods under MCDM for Islam data.

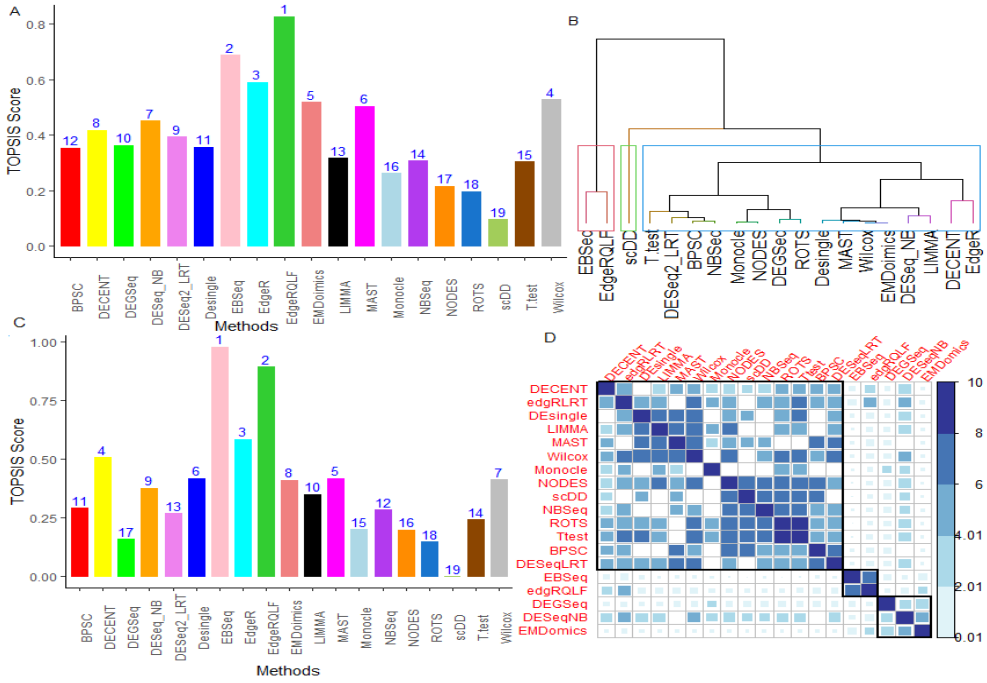


Figure 5.18. Performance evaluation of the methods under MCDM for Tung data.

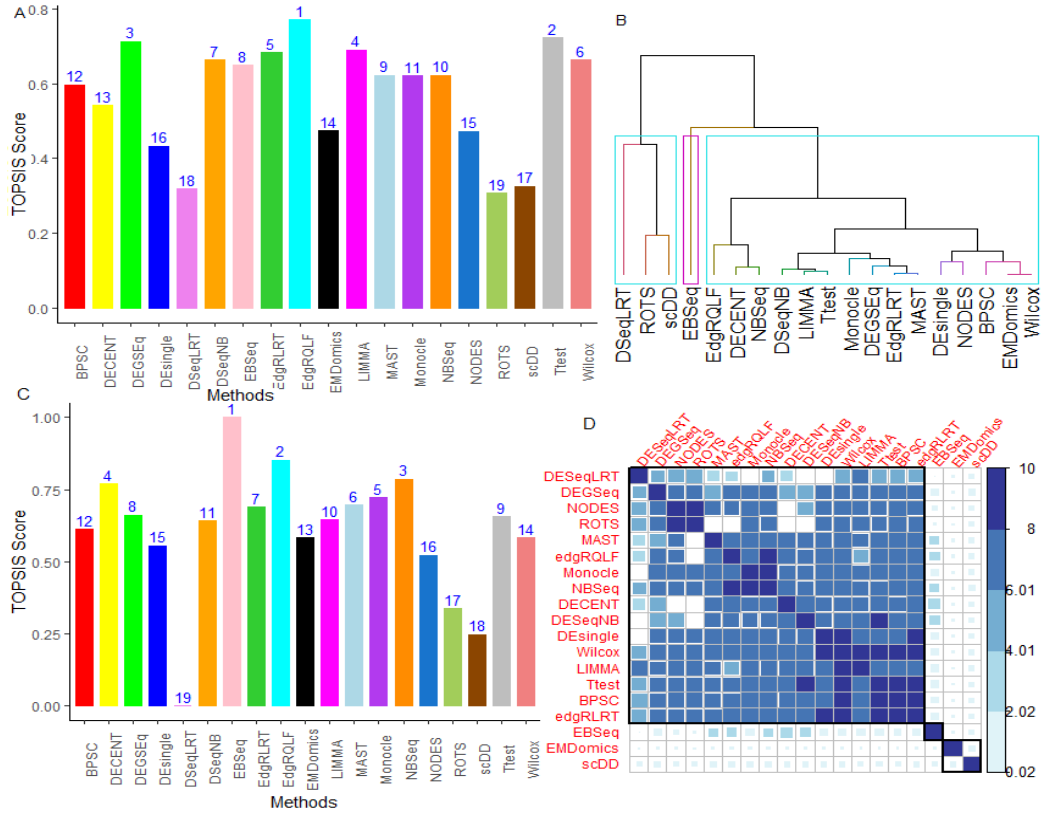


Figure 5.19. Performance evaluation of methods under MCDM setup for Soumillion1 data.

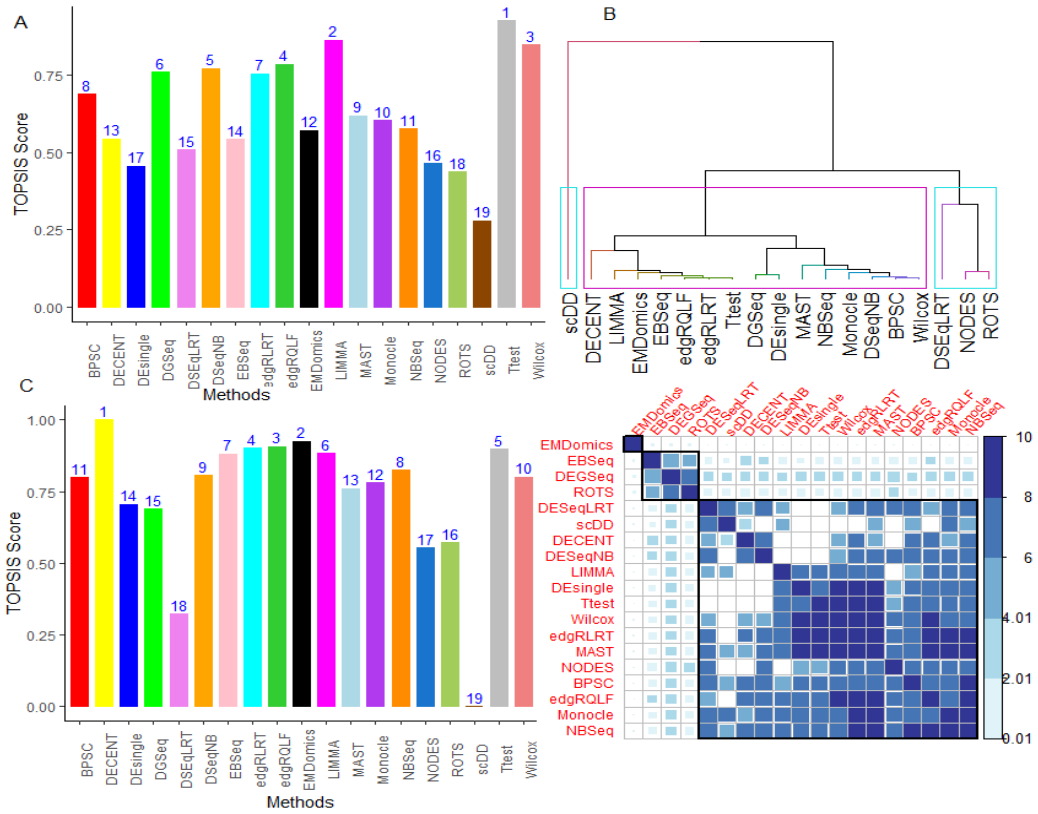


Figure 5.20. Performance evaluation of methods under MCDM setup for Soumillion3 data.

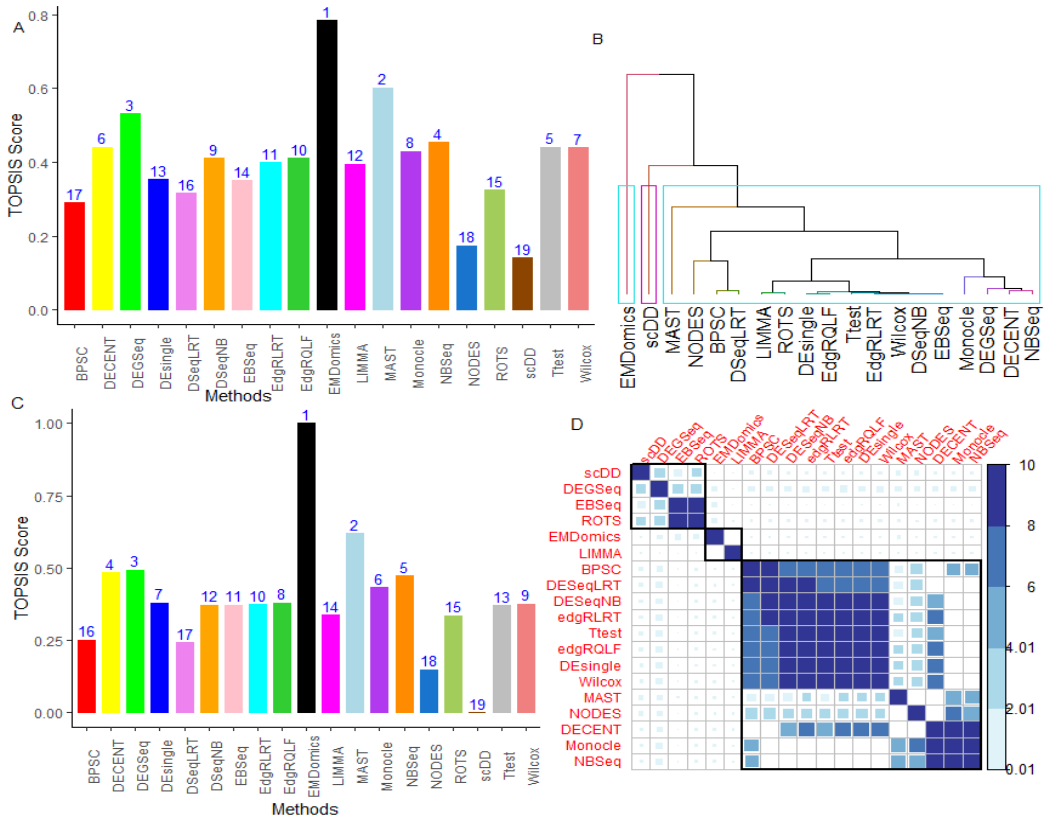


Figure 5.21. Performance evaluation of the methods MCDM setup for Klein data.

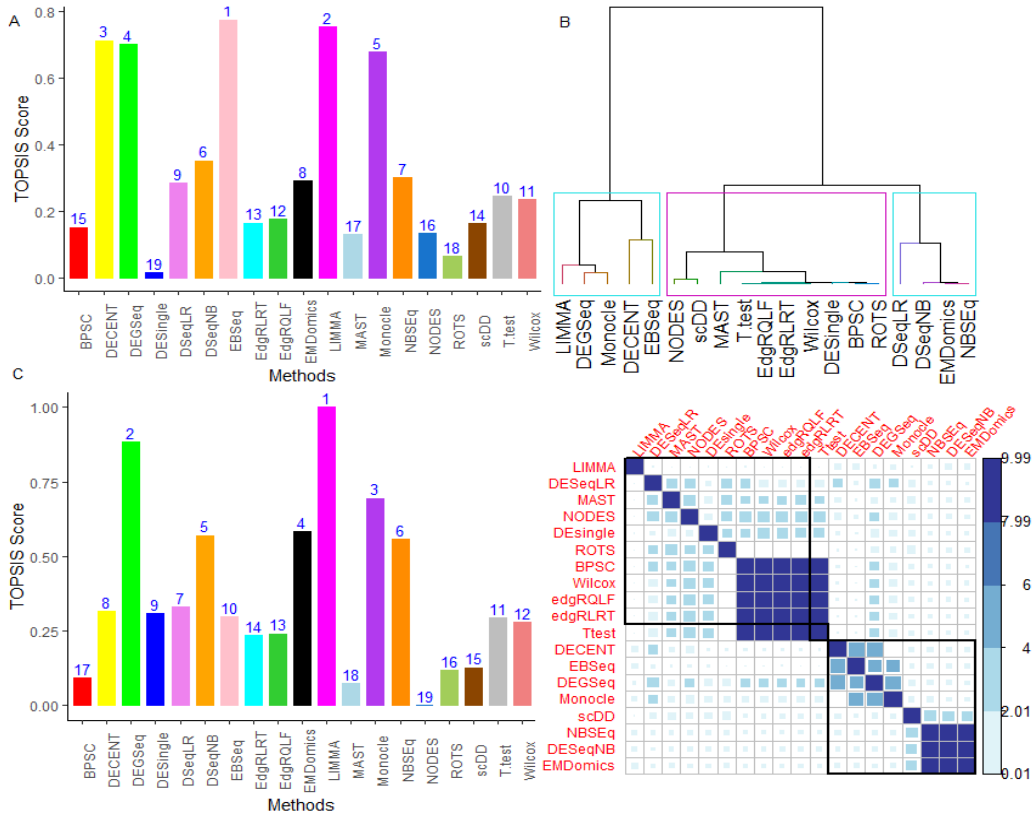


Figure 5.22. Performance evaluation of methods under MCDM for Gierahn data.

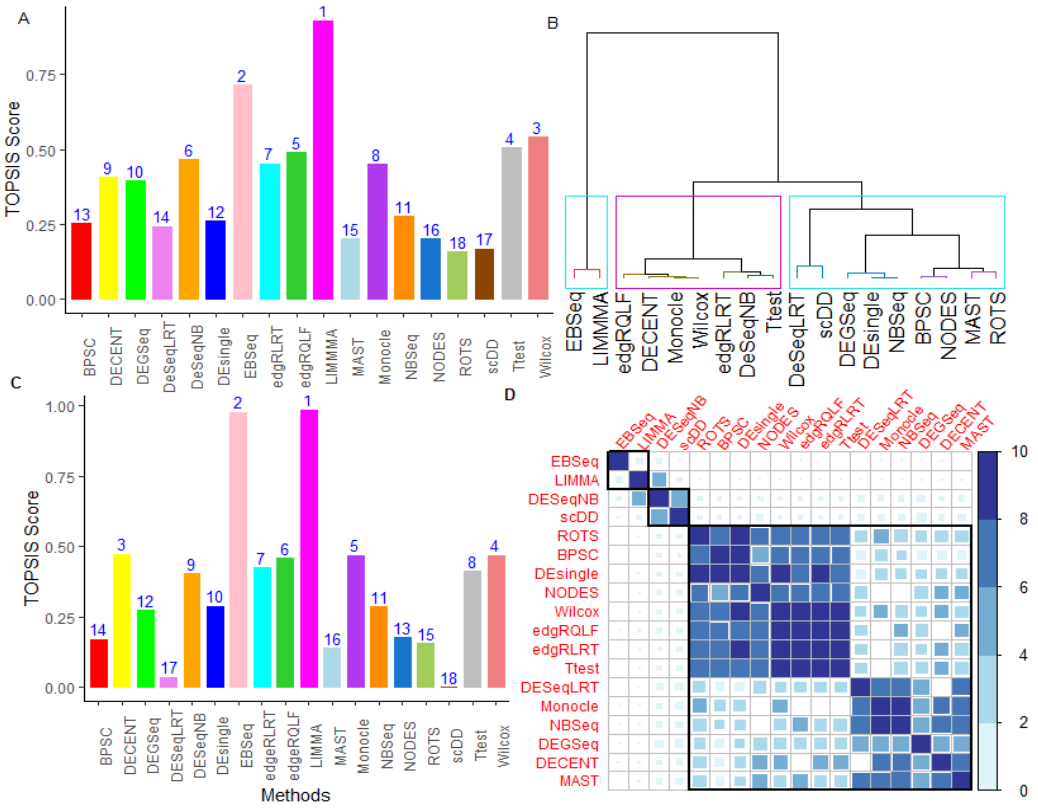


Figure 5.23. Performance evaluation of the methods under MCDM for Chen data.

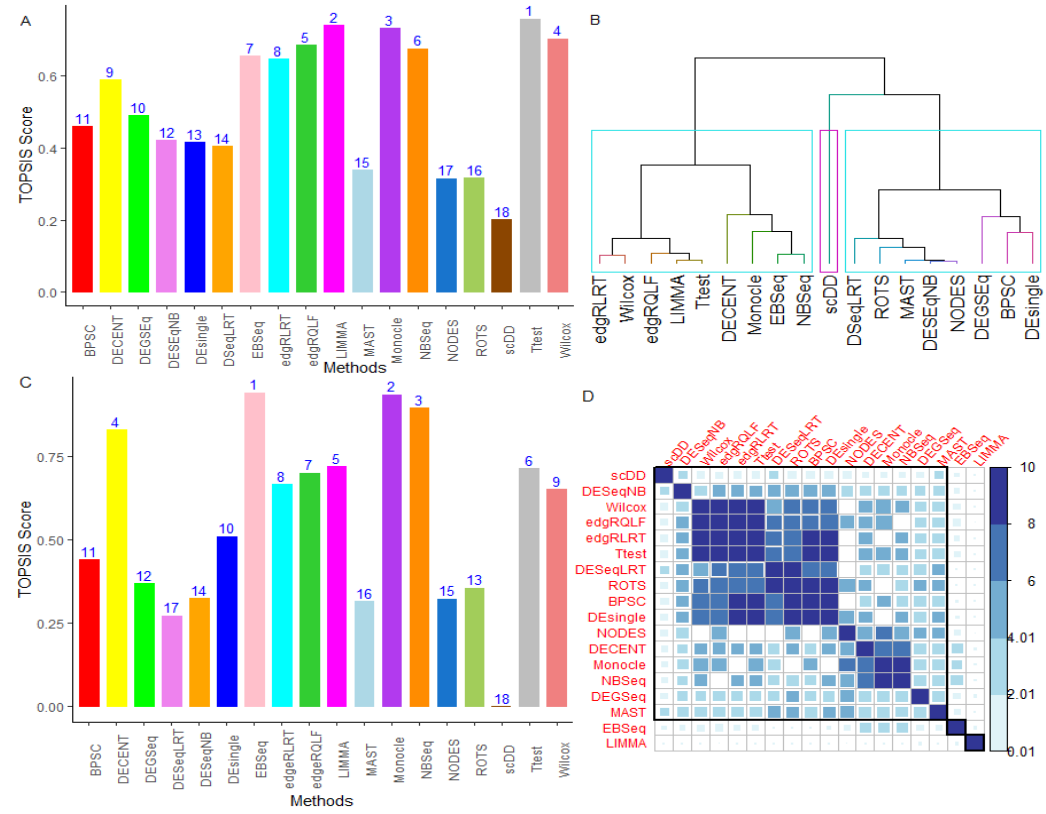


Figure 5.24. Performance evaluation of the methods under MCDM for Savas data.

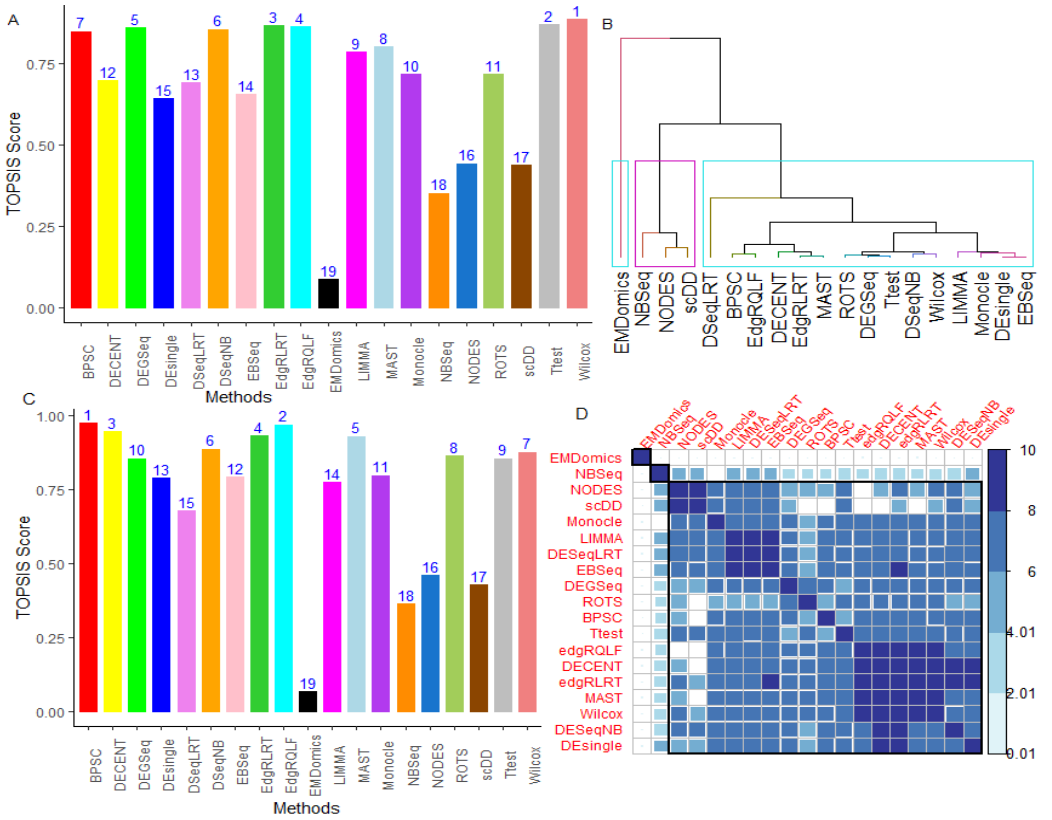


Figure 5.25. Performance evaluation of the methods under MCDM for Grun data.

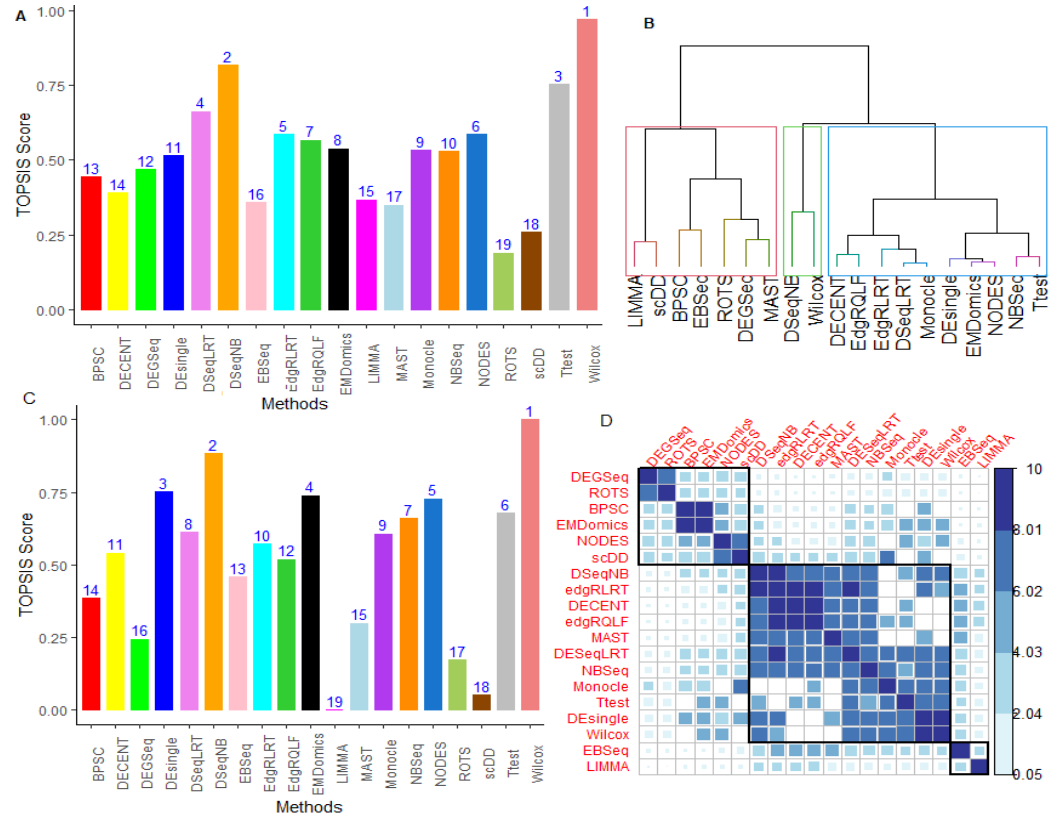


Figure 5.26. Performance evaluation of methods under MCDM for Zigenhein data.

For Soumillon2 data, DECENT provided the highest TOPSIS score followed by DESeqNB and LIMMA compared to other tested methods (Figure 5.16) under MCDM analysis without considering the runtime criterion. In contrast, we found the methods, including scDD, EMDomics, and EBSeq, performing worst among others under these multi-criteria setting. Further, when runtime criterion was included in MCDM-TOPSIS analysis, the tested methods' rankings were found to be significantly changed (Figure 5.16). For instance, the rank of DECENT slipped to 13 (Figure 5.16B) under runtime based MCDM analysis from rank 1 (without runtime-MCDM analysis) (Figure 5.16A). This indicated that the performance of the best methods (when assessed under MCDM analysis) was compromised when their runtime was integrated into the analysis. Here, it is interesting to note that the bulk RNA-seq methods such as LIMMA, DESeqNB, and edgeRQLF performed better under the MCDM settings (Figure 5.16).

Through MCDM analysis of tested methods on Islam data, edgeRQLF was found to be the best option to detect true DE genes, followed by DESeqNB and edgeRLRT (Figure 5.17). Surprisingly, general methods such as T-test and Wilcox methods ranked as 4 and 6, respectively, due to their lesser runtime (Figure 5.17). Among all the tested methods, NODES performed worst (rank:19) for this data, followed by scDD (rank:18), DEsingle (rank:17), and LIMMA (16). Under the MCDM-TOPSIS (without runtime) settings, the DECENT performed better (rank: 4) (computationally intensive), followed by Monocle, MAST, DEsingle, and BPSC in the single-cell categories. Similar interpretations can be made from the MCDM-TOPSIS analysis for the other 9 datasets (Figures 5.18 – 5.26).

Under this multi-criteria setting, it was found that the performance of tested methods varies differentially across the datasets and mostly depends on the data characteristics, such as the number of cells in the data and number of cells per group (Table 5.9). For instance, count-based NB tools, including edgeRQLF, DESeqNB, and edgeRLRT, performed better when the total number of cells in the data is relatively small and performed poorly for data with a large number of cells (Table 5.9). Further, the specially designed UMI-based DECENT performed better, particularly when there is a sufficient number of cells present in data (e.g., > 1000) (Table 5.9). However, the normalized data-based LIMMA performed exceptionally well for scRNA-seq data having many (*e.g.* >2000) cells but performed poorly under a small and medium number of cell situations.

Table 5.9. Effect of the number of cells on performance of the DE methods assessed through AUROC.

Methods	Small				Score	Medium				Score	Large			Score
BPSC	15	12	11.5	1	2.13	14	16	12	17.5	1.08	12	6	16.5	1.34
DECENT	4	5	13	5	2.79	3	2	4	2	3.63	1	1	3	2.89
DEGSeq	16	18	16	14	0.84	11	4	14	5	2.42	15	18	13	0.74
DESeqLRT	6	11	9	13	2.16	19	15	18	9	1.00	18	13	18	0.58
DESeqNB	1	8.5	2	6	3.29	6	11	17	7	2.05	7	2	10	2.16
DEsingle	10	6.5	3	10	2.66	10	8	11	17.5	1.76	13	9	11	1.42
EBSeq	9	2	14	8	2.47	1	13	1	1	3.37	10	14	1	1.84
edgeRLRT	3	3	10	4	3.16	8	10	8.5	14.5	2.05	5	5	7	2.26
edgeRQLF	2	1	11.5	2	3.34	2	5	5	14.5	2.82	4	4	4.5	2.50
EMDomics	5	8.5	4	19	2.29	13	1	10	7	2.58	2	17	9	1.68
LIMMA	18	10	19	11	1.16	15	17	8.5	4	1.87	3	3	2	2.74
MAST	14	4	15	3	2.32	5	3	16	12	2.32	11	12	15	1.16
Monocle	12	16	7.5	15	1.55	7	9	2	3	3.11	14	11	6	1.53
NBSeq	11	15	7.5	18	1.50	4	6	3	7	3.16	8	8	12	1.68
NODES	19	13	6	16	1.37	16	18	13	11	1.16	16	15	14	0.79
ROTS	13	17	18	9	1.21	17	14	15	14.5	1.03	17	16	16.5	0.55
scDD	17	19	17	17	0.53	18	19	19	10	0.74	19	19	19	0.16
T-test	8	14	5	12	2.16	9	12	7	14.5	1.97	6	7	8	2.05
Wilcox	7	6.5	1	7	3.08	12	7	6	19	1.89	9	10	4.5	1.92

Small number of cells (< 600) in scRNA-seq; Medium number of cells (1000 – 2000) scRNA-seq; Large number of cells (> 2000); score: Average of the rank scores computed through Eq. 32 based on the AUROC measure

Between-methods Similarity Analysis

The similarity analysis of the tested DE methods, based on the computed performance metrics, revealed the similarities in performance among the methods. We also compared the overlaps in terms of the detections of DE genes between any pair of methods (*i.e.*, degree of similarity) by extracting 1000 DE genes through each of them. For the Soumillon2 dataset, the results from such analysis are shown in Figure 5.5C and 5.5D. Here, we observed that bulk RNA-seq methods, *i.e.,* DESeqNB, edgeRQLF, NBPSeq, and LIMMA, are grouped together, have similar performance with single-cell methods, such as DECENT, DEsingle, BPSC, MAST, and Monocle along with general T-test and Wilcox methods (Figure 5.5C). Further, these methods shared a greater degree of similarity in terms of detecting more common DE genes compared to other methods (Figure 5.5D). This finding was well supported with higher correlations among themselves. In contrast, the methods which performed moderately well were clustered together, which are overrepresented by the methods, such as DEGSeq, ROTS, EBSeq, *etc*. Whereas the poorly performed methods (scDD and EMDomics), capable of dealing with the data's multi-modal nature, were grouped together and shared fewer common DE genes with other methods (Figure 5.5).

For Islam data, it was observed that the count-based bulk RNA-seq methods, such as edgeRQLF, DESeqNB, and edgeRLRT, are clustered together and have superior performance (Figure 5.6). While normalized data-based DE methods, *i.e.,* LIMMA, NODES, and scDD, are grouped together and found to have

lower TOPSIS scores indicating their poor performance (Figure 5.6A). Further, general methods (ROTS, Wilcox, EMDomics, and T-test) were clustered together, which had similar performance with the robust count based single-cell methods (BPSC, MAST, Monocle, DEsingle, and DECENT) and several bulk RNA-seq DE methods (DESeqLRT, EBSeq, and NBPSeq) (Figure 5.6A). It is interesting to note that the degree of similarity between DESeqNB and DESeqLRT was found to be low, indicating that the DE test statistic has a significant effect on the performance of the methods. The degree of similarity in terms of sharing common genes between any given pair of methods varied widely (Figure 5.6D) within and across datasets and mostly depends on the real data characteristics (*i.e.,* total number of cells and cells per group) (Figures 5.5-5.15, Table 5.9). These findings were in agreement with the previous studies [201,215]. Similar interpretations can be made for the other 9 datasets from Figures 5.16–5.26.

Combined-data Methods Analysis

The performance of the tested methods was found to be highly inconsistent across the datasets (Figures 5.16-5.26). Therefore, we performed a combined-data analysis of the assessed methods through the TOPSIS technique. For instance, edgeRQLF performed better in Islam data but not in Ziegenhain data when assessed through ACC metric (Figures 5.16, 5.26). However, their performance was somewhat positively associated with the total number of cells and the number of cells per group [216] (Table 5.9). To be more precise, on the selection of the best method across the multiple datasets, we performed TOPSIS analysis of the

methods based on the performance metrics, such as F1 score, FDR, TPR, FPR, and AUROC, and the results are shown in Figures 5.27-5.29.

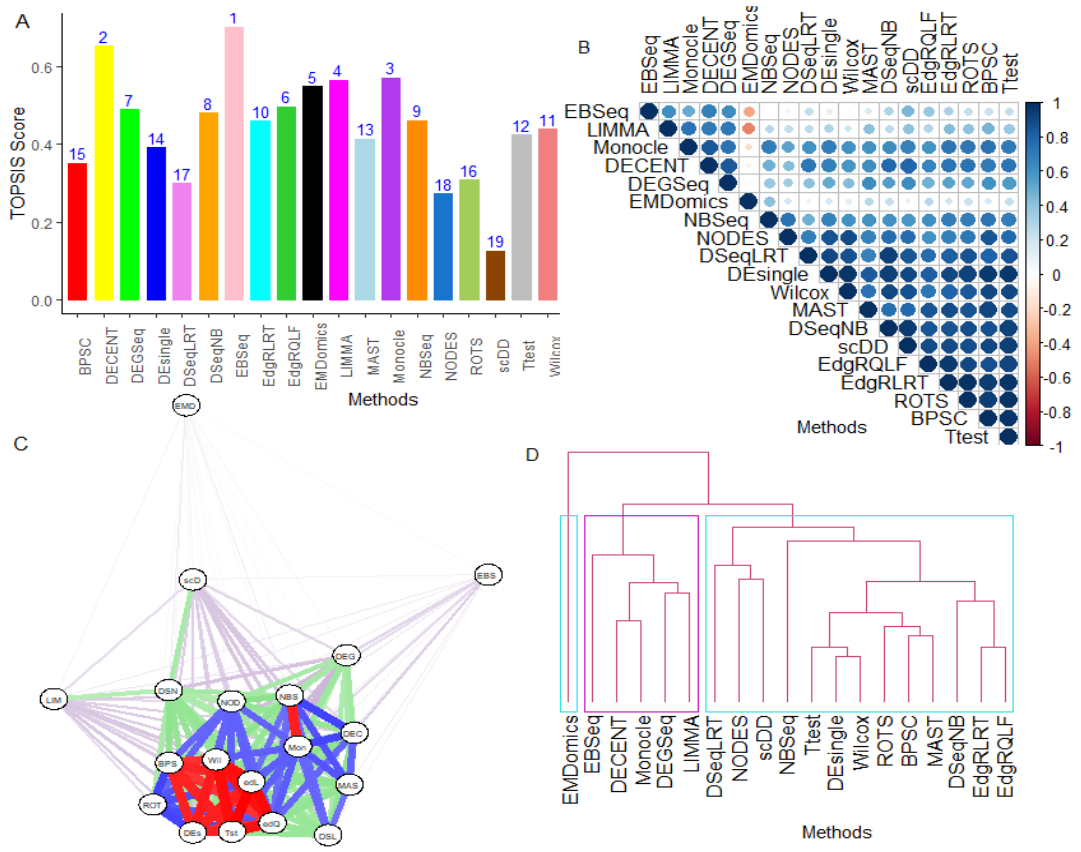
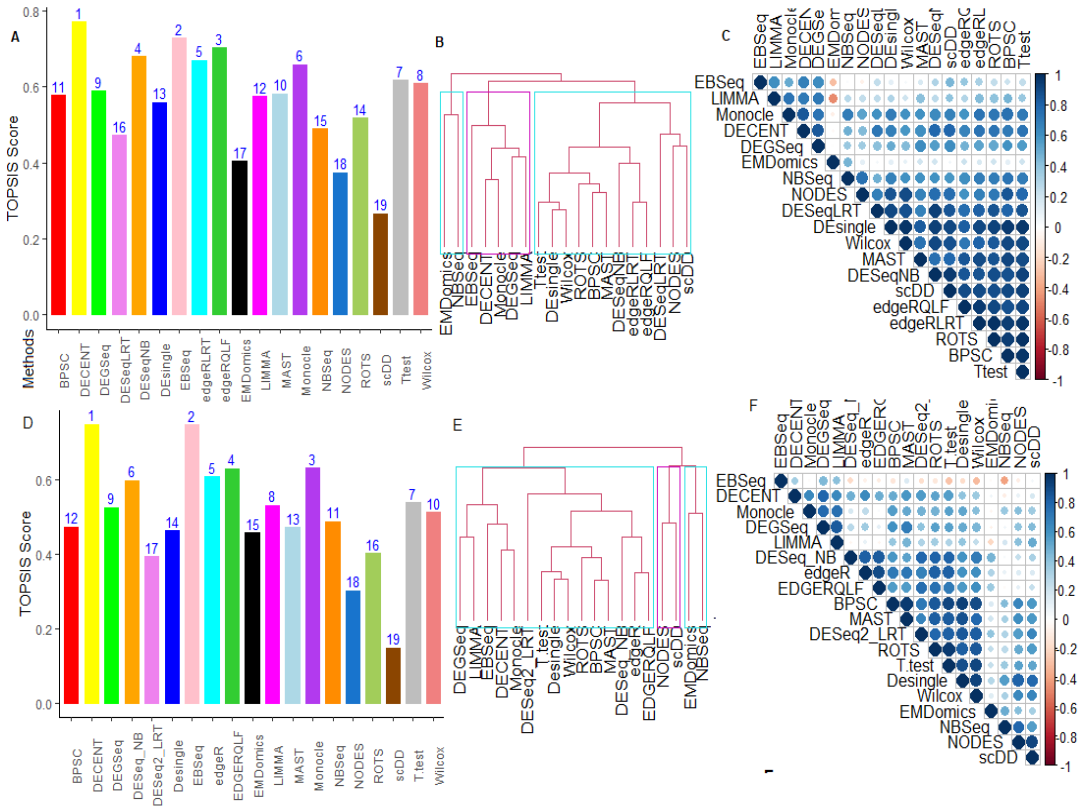


Figure 5.27. Combined data analysis of the DE methods based on F1 score through TOPSIS Approach. The comparative performance evaluation of the DE methods was performed based on F1 score through TOPSIS approach under multi-data setup. This analysis was performed on data matrix having F1 scores of the tested methods across the 11 considered datasets. (A) Shows results from TOPSIS analysis of the tested DE methods; (B) Correlation analysis of the evaluated DE methods through Rank correlation. The correlation plot was obtained by Spearman's rank correlation method based on the matrix of average (over DE gene sets) F1 scores across all data sets. (C) Weighted similarity analysis of the tested methods (Supp. Document S14) based on their ability to detect common genes. Here, the nodes represent the tested methods and edges represent the shared degree of similarity between the pair of methods. Further, the red color edges (with scores > 0.7) among the methods indicated highest similarity, blue color edges indicate higher similarity ([0.5, 0.7]), green color edges represent with low similarity ([0.2, 0.5]) and magenta color edges represent lowest degree of similarity ([0, 0.2]) among the methods. The nodes in the network abbreviated as, EMD: EMDomics; LIM: LIMMA; EBS: EBSeq; scD: scDD; DEG: DEGSeq; DSN: DESeqNB; NOD: NODES; BPS: BPSC; NBS: NBSeq; Wil: Wilcox; Mon: Monocle; DEC: DECENT; MAS: MAST; Tst: T-test; DEs: DESingle; ROT: ROTS; DSL: DESeqLRT; edQ: edgeRQLF; edL: edgeRLRT (D) Similarity analysis of the evaluated DE methods through clustering. The dendrogram was obtained by average-linkage hierarchical clustering based on the matrix of average (over DE gene sets) F1 scores across all data sets.

Through F1 based TOPSIS analysis of methods, it was found that the TOPSIS score of EBSeq and DECENT was highest, followed by edgeRQLF compared to others (Figure 5.27). This indicates that both are better options for DE analysis of scRNA-seq data over others across all datasets, but highly computationally intensive and required several hours even for a small dataset (Figure 5.27). Further, through F1-based similarity analysis, the UMI based parametric methods, such as DECENT, EBSeq, Monocle DEGSeq, are grouped together and have similar performance with LIMMA across the datasets (Figure 5.27C). However, EMDomics is the only (NP) method clustered separately, as it is a general-purpose method (*i.e.,* origin is out of RNA-seq context). Also, its performance was negative correlated with other methods (Figure 5.27). Moreover, count-based bulk RNA-seq methods, such as DESeqNB, DESeqLRT, edgeRQLF, edgeRLRT, and NBSeq are clustered together and were found to be similar with single-cell methods (*i.e.,* BPSC, scDD, NODES, MAST and DEsingle) and general-purpose methods (T-test, Wilcox, and ROTS) (Figure 5.27).

Figure 5.28. Combined data analysis of the DE methods based on FDR and Accuracy metrics through TOPSIS Approach. The comparative performance evaluation of the DE methods was performed based on FDR and Accuracy metrics through TOPSIS approach under multi-data setup. This analysis was performed on data matrix having FDR and Accuracy scores of the tested methods across the 11 considered datasets. (A) Shows results from TOPSIS analysis of the tested DE methods based on FDR; (B) Similarity analysis of the evaluated DE methods based on FDR through clustering. The dendrogram was obtained by average-linkage hierarchical clustering based on the matrix of average (over DE gene sets) FDR scores across all data sets. (C) Similarity analysis of the evaluated DE methods based on FDR through correlation. The correlation plot was obtained by Spearman's rank correlation method based on the matrix of average (over DE gene sets) FDR scores across all data sets; (D) Shows results from TOPSIS analysis of the tested DE methods based on Accuracy; (E) Similarity analysis of the evaluated DE methods based on Accuracy through clustering. The dendrogram was obtained by average-linkage hierarchical clustering based on the matrix of average (over DE gene sets) Accuracy across all data sets. (F) Similarity analysis of the evaluated DE methods based on Accuracy metrics through correlation. The correlation plot was obtained by Spearman's rank correlation method based on the matrix of average (over DE gene sets) Accuracy scores across all data sets.



FDR and ACC-based TOPSIS analysis of the 19 methods over the 11 real datasets was also performed, and the results are shown in Figure 5.28. It was found that the TOPSIS score (for both FDR and ACC) of DECENT was highest, followed by EBSeq compared to others (Figure 5.28). Further, averaging ACC measure across the DE gene sets and ranking of the methods revealed that DECENT's performance is somewhat consistently better, followed by edgeRQLF and EBSeq (Figure 5.5F – 5.15F). DECENT considers the cell capture rates, cell auxiliaries, and employs an efficient ECM algorithm for parameter estimation, hence capable of detecting more true and robust DE genes. Further, it is also well equipped to handle the molecular capture process, cell sizes, extra zero inflation, and overdispersion present in the data. DECENT uses the ZINB model to fit the (scRNA-seq) UMI data, which accurately estimates the mean and dispersion

parameters unlike NB based tools, thus having more statistical accuracy in detecting DE genes (Table 5.4-5.6, Figure 5.4). In this comparison, the performance of NP methods, such as NODES, EMDomics, ROTS, and scDD, was found to be consistently bad across the datasets. Similar interpretations can be made for the other performance metrics, such as TPR, FPR, FDR, AUROC, given in Figures 5.16-5.26.

Among all the tested methods, the single-cell methods, such as scDD, NODES, and DEsingle, performed poor along with the general methods, *i.e.,* EMDomics ad ROTS, and bulk RNA-seq methods (DESeqLRT, NBSeq, and LIMMA) in terms of accuracy and robustness in detecting true DE genes (Figure 5.28). Under this setting, simple methods, such as T-test and Wilcox, performed reasonably well with the least computational times (Figures 5.27, 5.28) to get better and robust results. Expressly, these methods, along with DEsingle, NODES, and ROTS, are limited to only two cell groups comparisons and cannot accommodate cell capture rates, cell covariates, *etc.* whereas EMDomics can perform a limited number of analysis types. The remaining methods implement statistical frameworks that can accommodate more complex (fixed effect) designs, including comparisons across multiple groups and adjustments for batch effects and cell-level covariates. Further, there are specific methods, including Monocle, and LIMMA, which accurately detected the true DE genes but are prone to higher error rates (Figure 5.28).

Through Sensitivity-Specificity-based TOPSIS analysis of tested methods across the datasets indicated that DECENT and EBSeq performed exceptionally

well with higher rates of sensitivities and specificities (higher AUROC) for detecting true DE genes (Figure 5.29).

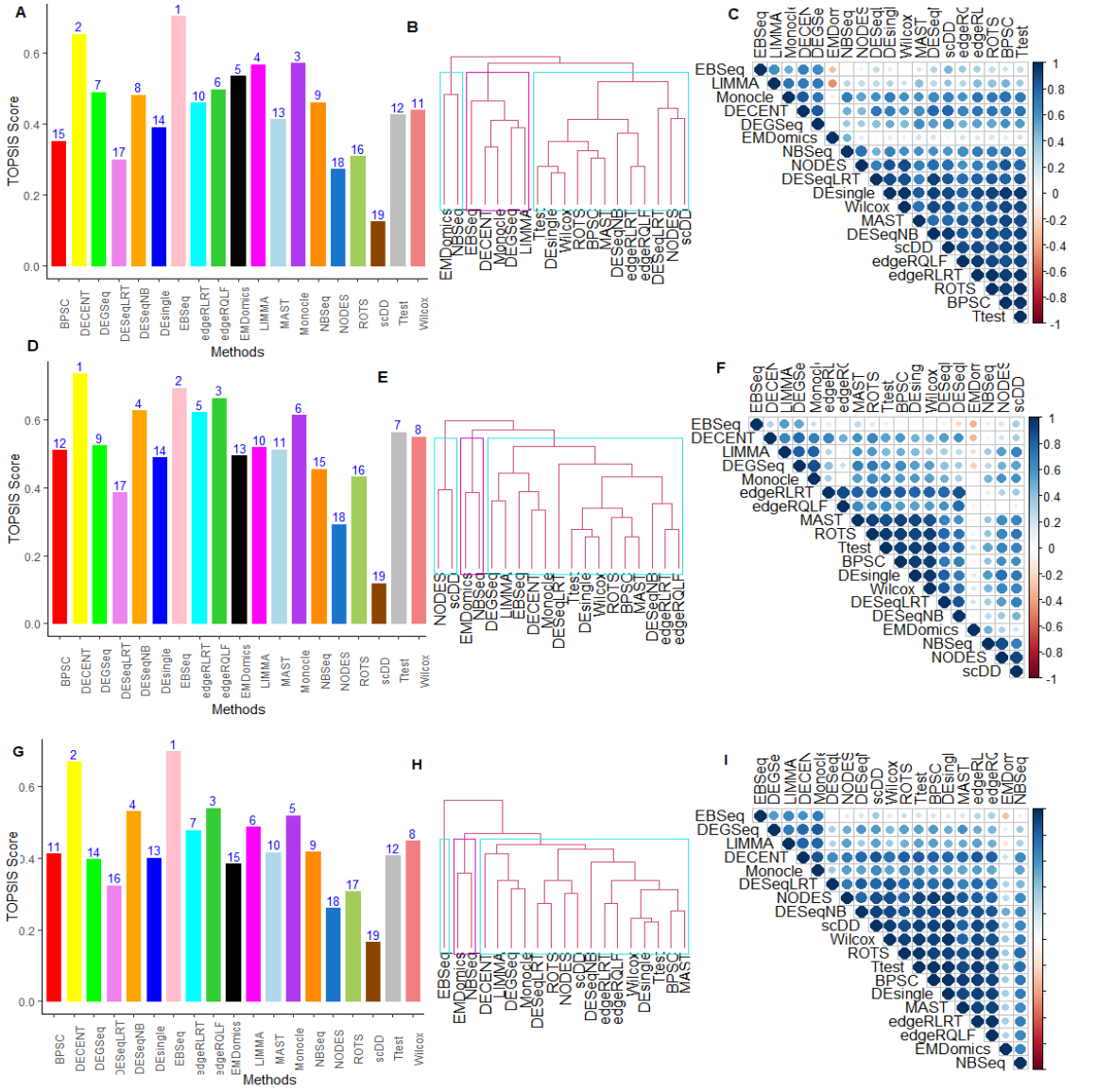


Figure 5.29. Combined data analysis of the methods based on Sensitivity-Specificity through TOPSIS Approach. (A) Shows results from TOPSIS analysis of the tested DE methods based on TPR; (B) Similarity analysis of the evaluated DE methods based on TPR through clustering. (C) Similarity analysis of the evaluated DE methods based on TPR through correlation. (D) Shows results from TOPSIS analysis of the tested DE methods based on FPR; (E) Similarity analysis of the evaluated DE methods based on FPR through clustering; (F) Similarity analysis of the evaluated DE methods based on FPR metrics through correlation; (G) Shows results from TOPSIS analysis of the tested DE methods based on AUROC; (H) Similarity analysis of the evaluated DE methods based on AUROC through clustering. The dendrogram was obtained by average-linkage hierarchical clustering based on the matrix of average (over DE gene sets) AUROC scores across all data sets. (I) Similarity analysis of the evaluated DE methods based on AUROC through correlation.

The group of methods, including Monocle, LIMMA, EMDomics, were observed to have higher rates of sensitivities, but compromised the specificities for detecting DE genes and have similar performance with NBSeq and DEGSeq (Figure 5.29). Further, the count-based bulk RNA-seq methods (*i.e.,* edgeRQLF, edgeRLRT, and DESeqNB) had higher specificities along with general DE methods (*i.e.,* T-test and Wilcox). Still, they compromised the sensitivities for detecting true DE genes in scRNA-seq data (Figure 5.29). Surprisingly, the specially designed single-cell methods, such as BPSC, MAST, DEsingle, and NODES, were found to have lower sensitivities and specificities for DE analysis and have similar performance with ROTS. These findings were also supplemented by AUROC-based TOPSIS analysis (Figure 5.29G).

FDR-based similarity analysis indicated that the EMDomics and NBSeq were clustered together with similar performance, whereas the bulk and single-cell methods (*i.e.,* EBSeq, LIMMA, DECENT, and Monocle) were clustered together. The count-based bulk RNA-seq methods (*i.e.,* edgeRQLF, edgeRLRT, DESeqNB, and DESeqLRT) were clustered together with poorly performed UMI-based methods (BPSC, DEsingle, MAST, NODES, scDD), which had similar performance with the general two-class methods (Wilcox, T-test, ROTS) (Figure 5.28). However, the EMDomics was negatively correlated with some of the methods, as it is a general-purpose NP genomic method and does not consider the RNA-seq data features (Figure 5.28C). Similar interpretations can be made for ACC, TPR, FPR, and AROC based similarity analysis (Figures 5.28, 5.29).

The combined analysis through the TOPSIS method allowed us to select the best option for DE analysis over the tested methods, simultaneously considering multiple real datasets. Through such analysis, DECENT, EBSeq, and edgeRQLF were consistently performed better, whereas the group methods, such as scDD, NODES, ROTS, EMDomics, and DEsingle always performed extremely poor over the datasets irrespective of the performance criteria used. The remaining tested methods' performance varied differently across the datasets under different performance metrics (Figures 5.15 – 5.25). It is mostly positively related to the total number of cells present in the data (Table 5.9). Interestingly, the performance of popular count-bulk RNA-seq methods, such as edgeRQLF and DESeqNB, were found to be consistently better over edgeRLRT and DESeqLRT, respectively (Figures 5.27-5.29). This observation indicated that the performance of the DE methods, such as edgeR and DESeq2, mostly depends on the test statistic(s) they use to perform DE analysis of genes. Moreover, in agreement with previous comparative studies [192,193,201,215], DE methods developed bulk RNA-seq analysis did not perform worse than those specially designed for scRNA-seq data, even performed better for some cases.

“The purpose of statistical models is not to fit the data but to sharpen the question..”

Samuel Karlin

CHAPTER 6

AN IMPROVED STATISTICAL APPROACH FOR DIFFERENTIAL EXPRESSION ANALYSIS OF SINGLE-CELL RNA-SEQ DATA

Background

Advent of scRNA-seq technologies have revolutionized transcriptomics through generating gene expression data at the single cell resolution level [195,272]. It has numerous advantages over bulk RNA-seq technology, which only characterize the global expression dynamics of genes in a tissue sample, while ignoring the inherent cell-cell heterogeneity [273,274]. Thus, it is pertinent to assess the variability that exists among the cells in a tissue sample as this is crucial to understand the complexity and dynamicity of biological processes such as embryogenesis [195,256], cancer [275], *etc*. Through scRNA-seq technology, expression is quantified by mapping reads to reference genome followed by counting the number of reads mapped to each gene [195]. Here, individual transcript molecules are attached with an UMI tag; subsequently, counting the UMIs usually yield the number of transcripts for each gene in a cell [196]. Further, huge amount of UMI count data are generated for several thousand(s) of genes across thousand(s) of cells and subsequently deposited in public domain databases by researchers across the globe. Hence, it is necessary to develop new,

and innovative statistical approaches and tools for such data analysis to harness the potential of this new technology.

Small amounts of the mRNA molecules and imperfect procedures for capturing them in individual cells, lead to dropout events, *i.e.* genes show zero or very low expression, though they are expressed in cells [202,209]. Further, it is well established, that the capture rates vary between cells for a given scRNA-seq protocol, showing a major source of unwanted technical variation that adds to the dropout events [276,277]. Addition of UMIs during the library preparation step reduces the amplification bias but has no effects on dropout events [197]. Further, the dropout events add more zeros to the output data, and can be attributed as: true/biological zeros (gene is not expressed in the cell); or false/technical zeros (gene is expressed but not detected) [203]. The presence of higher proportions of zeros and technical noise in scRNA-seq data can severely affect the performance of downstream DE analysis.

Bulk RNA-seq DE methods such as edgeR [204], and DESeq2 [205,206] were used extensively for DE analysis of scRNA-seq data. These models use the NB model to capture the distributional nature of read counts under a GLM framework. Further, Limma-Voom considers linear models for log-transformed counts data and observation-level weights to account for the dispersion nature of the transformed data [161,228], while DEGseq assumes the Poisson distribution of the read counts [218]. The utility of such approaches raises serious concerns about their validity due to high dropout events [203], transcriptional bursting [207], lower molecular capturing in cells [208,209], and higher dispersion, *etc*. Therefore,

dedicated scRNA-seq DE methods have been developed which use different strategies to cope with the above concerns [202,203,208,209,211–213]. For instance, SCDE uses a mixture model (*i.e.* Poisson for dropout part and NB for amplification part) to capture the observed abundance of a given transcript in each cell [214]. SCDE always assumes that the observed zero-count belongs to the dropout events with certainty. Further, MAST uses a hurdle model, *i.e.* logistic regression for the level of gene expression and a Gaussian linear model for rate of expression conditioned on expression levels [202]. However, SCDE and MAST do not differentiate between biological and technical zeros during the model building. BPSC approach [221] was developed for performing DE analysis of scRNA-seq data through integrating Beta-Poisson model in the GLM framework. The BPSC approach does not consider the count nature of UMI data, and severely affected by the dropout events [215]. These methods specifically consider the bi-modal distributional nature of the scRNA-seq data. Hence, a class of methods including D3E [237] and scDD [222] was developed to address the multimodal distributions of transformed scRNA-seq data, but they failed to consider the UMI count nature of the data and excluded the dropout events. Further, methods such as Monocle [212], Monocle2 [251], and NBID [244] were designed to handle the unique features of UMI in scRNA-seq experiments. They fit NB models directly to the observed UMI count data without any explicit focus on dropout events. Next, another class of specialized methods, such as ZINB-WaVE [203,242], DEsingle [211], and DECENT [209], were developed to handle excess zero inflation in scRNA-seq data. These methods are based on fitting of ZINB models to the UMI

counts data. To be more specific, ZINB-WaVE [203] estimates observational level weights through EM algorithm for adjusting bulk DE methods, *i.e.* edgeR [204], DESeq2 [206]. The DEsingle [211] approach assumes ZINB models for observed scRNA-seq UMI count data to estimate the parameters through MLE method for two cellular populations separately. However, DECENT [209] assumes ZINB model for observed UMI count data and considers a Beta-Binomial model for the molecule capturing process with the option to use cell-level auxiliary information. These methods ignore multimodal distributions of the observed expression data, estimate the DE parameters under parametric model assumptions, and are mostly focused on two-groups comparisons. Further, there is another class of DE methods which explicitly considers technical variation and molecular capturing processes, based on external spike-ins data. This class includes methods such as TASC [253], BASiCs [243], DECENT [209], and DESCEND [208]. Moreover, several comprehensive reviews and comparative analysis of DE methods covering all the above classes can be found in literature [192,193,201,215–217].

It is evident that cells in scRNA-seq data behave differentially and tend to be in different cell clusters [278], due to cell-cell heterogeneity. Biologically, these cell clusters are often different cell-types (*e.g.* neurons and glia in brain sample) and correspond to different active states of cell types. Hence, descriptive data mining strategies (*i.e.* clustering) have been adapted for scRNA-seq data analytics. In this study, we argue that the underlying cell clusters may have a significant effect on the means of non-zero counts of genes, and subsequently may influence the power of detection of DE genes. Further, there are limited methods available to

date which consider the molecular capturing process, cell cluster information, and other cell-level auxiliary information for DE analysis. The incorporation of these data into the DE methods, may enhance the performance. This process requires building specific statistical models in order to perform statistical tests reliably.

In this chapter, we therefore, propose a novel statistical approach, *i.e.* SwarnSeq, for the DE analysis of UMI count scRNA-seq data. Here, we integrate the parametric ZINB model with binomial molecular capture process in the presence of cell-level data. This allows us to detect DE genes and Differential Zero-Inflated (DZI) genes under a GLM framework. SwarnSeq can also classify the influential genes from scRNA-seq study into various groups. SwarnSeq can use external RNA spike-in data to adjust the distribution of the observed UMI counts with capture rates; however, it also works without spike-ins. In this Chapter, we describe SwarnSeq approach and benchmark it against 11 other existing methods, *i.e.,* DEGseq [218], edgeR [204], DEseq2 [205,206], LIMMA [161], Monocle2 [213], MAST [202], BPSC [221], SCDD [222], DEsingle [211], DECENT [209], and NODES [223], using 10 real scRNA-seq datasets. The detail descriptions about these methods are given in Chapter 5. Our analytic results indicate that the SwarnSeq approach outperformed the competing existing methods on multiple real datasets, when assessed under 3 comparative settings.

Material and Methods

Motivational data example

In scRNA-seq DE analysis, the cells are clustered, and these cell clusters are further divided into two groups (for example: group 1 has cluster *M* and group 2

has remaining clusters). In existing analyses, this cell cluster information is kept out of the model building, and this may have a significant influence on the mean of non-zero counts. To validate this claim, we took a subset of scRNA-seq dataset having 200 genes and 150 cells (Group 1: 50; Group 2: 100) from publicly available human preimplantation embryonic cells data [279]. Then, we model the mean of non-zero counts under a GLM framework by providing group and cell cluster information as auxiliaries. The results are shown in Table 6.1.

Table 6.1. Effect of cell clusters and groups on mean of non-zero scRNA-seq counts.

	#cells	Max	Min	#Zeros	Avg. Exp.	Co-efficient	Z-value	Sig.[a]
Intercept	-	-			-	7.35	11.81	***
Group 1	50					Ref.	Ref.	Ref.
Group 2	100	-			-	-2.8016	1.703	*
Cell Clust. 1	16	179	0	8	27.9375	-2.1	-1.875	*
Cell Clust. 2	47	437	0	30	23.70213	-3.55	-4.277	***
Cell Clust. 3	3	45	0	2	15	-2.8527	-1.626	NS
Cell Clust. 4	8	145	0	6	23	-2.97	-2.78	**
Cell Clust. 5	6	3496	466	0	1557.167	-2.18	-1.356	NS
Cell Clust. 6	14	13	0	11	1.857	-3.005	-3.84	***
Cell Clust. 7	32	497	0	23	24.375	-4.67	-3.7	***
Cell Clust. 8	16	308	0	6	49.937	-3.041	-3.89	***
Cell Cluster 9	8	20	0	5	3.125	-3.477	-2.735	**
log(theta)	-					-0.841	-2.706	**

Cell Clust.: Cell cluster; Max: Maximum read count; Min: Minimum read count; Avg. Exp.: Average expression values; a: comparing group 1 vs. remaining Cell clusters (e.g. Cell cluster 1 vs. Cell clusters 2-9) *, **, *** represents values significant at 5%, 1% and 0.1 % level of significance; (.,.): number of cells in [a]:

Our preliminary analysis indicates that cellular group 2 has significant effects on the mean counts of the gene, which means that the gene is expressed differentially with respect to group 1 (Table 6.1). Further, most of the cell clusters have significant effects on the mean read count the gene. Specifically, all the cell clusters except cell clusters 3 and 5 have significant effects on mean count (Table 6.1). The Table 6.1 describes the analysis for a single gene, if there are *N* genes

N such tables can be formulated. Also, if there are *K* cell clusters, there will be *K(K-1*)/2 grouping combinations and there will be *KN*(*K*-1)/2 such tables. Therefore, we can hypothesize that the cell clusters may influence the DE analysis of genes in scRNA-seq data. This toy data example motivates us to develop statistical approach and tool for DE analysis of scRNA-seq data by integrating cell cluster information other cell-level auxiliaries, and cell capture rates into the model building process under a GLM framework.

Single cell RNA-seq datasets

Our comprehensive analysis includes benchmarking of the proposed SwarnSeq method against 11 competitive existing methods (given in Chapter 5) on multiple real scRNA-seq datasets. This process starts with collection of publicly available scRNA-seq datasets from the GEO NCBI database (https://www.ncbi.nlm.nih.gov/geo). In our comparative analysis, we included the 10 UMI count gene expression datasets derived from 8 independent scRNA-seq studies. Further, the selected datasets were generated for lung cancer cells, pluripotent stem cells, liver cells, adipose stem/stromal cells, HEK cells from human, and embryonic stem cells, blood cells, and cells from mice. There are limited studies, where transcript molecular concentration and external spike-in data are publicly available. Hence, we used molecular concentration and ERCC spike-in data available from Tung et al.'s experiment to estimate the cell capture rates, while for other data cases, cell capture rates are estimated from the data *per se*. The selected datasets are briefly described as follows.

Lung cancer data (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111108)

This dataset is publicly available from GEO repository with accession GSE111108 [275]. The ScRNA-seq data are generated for an equal mixture of cells from the 3 Human Lung Adenocarcinoma cell lines (H2228, NCI-H1975 and HCC827) through 10X Genomics protocol and sequenced with Illumina NextSeq 500. This data consist expression counts of 33456 genes over 4000 cells. At preliminary stage, we removed the cells whose library size is less than 1800 and also further removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 17326 genes over 2126 cells for further analysis. Further, we used the *OptimCluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters. For this purpose, we set the seed value as 1712 and, found that the 2126 cells are clustered into 8 optimum cell clusters (Figure 6.1). For DE analysis, we took cell cluster 3 (987 cells) as group 1 and remaining cell clusters as cell group 2 (1139).

Pluripotent stem cell data

This dataset is publicly available in GEO database with accession id GSE77288 [197]. We downloaded the filtered UMI count matrix from their GitHub repository (https://github.com/jdblischak/singleCellSeq). The full dataset contains three Yoruba induced pluripotent stem cell lines, with three 96-well plates per individual. Here, we used the ERCC spike-ins, UMI and molecular concentration data were used. We only used data of two individual cell lines NA19101 (288 cells) and NA19239 (288 cells) (for 18938 genes) for further analyses. Here, we have not removed any cells from analysis. also further removed the genes which have non-

zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 15955 genes over 576 cells for further analysis. Further, we used the *OptimCluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters. For this purpose, we set the seed value at 108 and, found that the 576 cells are clustered into 10 optimum cell clusters (Figure 6.1). Further, for DE analysis, we took the two given cell lines (*i.e.* NA19101 and NA19239) as two cellular groups.

Mouse blood cell data

This dataset is publicly available in NCBI GEO database with accession GSE109999 [275]. For this experiment, we downloaded the count expression data as it has undergone rigorous preprocessing by the authors of the original publication. Here, the blood cells are derived from B lymphocytes, erythroblasts, granulocytes, high-end progenitor/stem and T cells from the bone marrow of a C57BL/6 10-13-week-old female. Here, expression counts of 19903 genes over 383 cells were generated through a modified CEL-seq2 protocol. Here, we have not removed any cells from analysis, but removed the genes which have non-zero expressions in ≤5 cells. Through this, we selected expression counts of 13055 genes over 383 cells for further analysis. Further, we found that the 383 cells are clustered into 9 optimum cell clusters through SwarnSeq R package (seed=110). For DE analysis, we took cell cluster 2 (180 cells) as one group and remaining cell clusters (203 cells) as other group.

Liver cell data

We downloaded counts expression dataset from the NCBI GEO database with id GSE115469 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115469) [280]. We directly took the count data, as it has undergone rigorous pre-processing, mapping, and other data analysis procedures by the authors of the original publication. In this study, the fractionated fragile, fresh hepatic tissue from human livers was obtained to get viable parenchymal and non-parenchymal cells. Then, expression profiling of cells is done by high throughput sequencing. The data consists the counts expression data of 20007 genes over 8444 cells. To reduce the size of the data, we removed the cells whose sizes are less than 1500 and genes which have non-zero counts in 5 cells. Through this, the expression data of 17316 genes over 5466 cells retained for further analysis. For fitting SwarnSeq model, the 5466 cells are clustered into 16 optimum cell clusters through executing the *OptimCluster* function (seed=222) implemented in SwarnSeq R package. Further, for DE analysis, we took cell cluster 3 (1852 cells) as one group and remaining cell clusters (3614 cells) as other group.

Mouse cell data

The dataset is publicly available in NCBI GEO database with accession GSE29087 [196] (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29087) and widely used for benchmarking of scRNA-seq DE tools. scRNA-Seq expression profiles were generated for 22928 genes over 96 cells, (48 mouse ES cells, 44 mouse embryonic fibroblasts and 4 negative controls) were analyzed by single cell tagged reverse transcription. Negative control cells are removed from the further analysis. For DE analysis, ES and MEF cell lines are considered as two cellular groups.

Here, we have not removed any cells from analysis, and further removed the genes which have non-zero expressions in ≤ 5 cells. Through this, we selected counts of 11436 genes over 92 cells with 8 optimum cell clusters.

Adipose stem/stromal cell data

This dataset is publicly available from GEO database with Accession GSE53638 [255] (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53638). Here, cells are collected during directed differentiation of human adipose-derived stem/stromal cells and further, 11,116 cells are profiled by the authors of the original publication. Here, the cells were collected at different stages and different time points (day 0, day 3 and day 7) and sequenced using the SCRB-seq protocol with UMI. To study the performance of scRNA-seq DE tools, we used two group comparison settings based on different time points, i) Data 1 (Day 0 (1245 cells, baseline) vs. Day 3 (590 cells), (ii) Data 2 (Day 0 (1245 cells) vs. Day 7 (1023 cells), and (iii) Data 3 (Day 7 (1023 cells) vs. Day 3 (590 cells). Here, we have not removed any cells from analysis, and further removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected UMI counts of 14863, 15637, and 15015 genes (from 23895 genes) over 1835, 2268, 1613 cells for Data 1, Data 2, and Data 3, respectively. The optimum number of cell clusters was determined as 11, 10 and 8 for each dataset through executing the *OptimCluster* function.

Mouse embryonic cell data

This data is publicly available in NCBI GEO database with accession id GSE65525 [256]. Here, the mouse embryonic stem cells expressions were profiled through high throughput sequencing using a droplet-microfluidic approach. The UMI count

data with 24174 genes over 1481 cells was used in this study. Further, at the preliminary stage, we removed the cells whose library size is less than 1500 and also removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 23971 genes over 1481 cells for further analysis. For fitting SwarnSeq model, the 1481 cells are clustered into 11 optimum cell clusters (seed =108) (Figure 6.1). Further, for DE analysis, we took day 4 (cells 683) as group 1 and day 7 (798 cells) as group 2.

HEK cell data (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92495)

This dataset is publicly available in NCBI GEO database with accession id GSE92495 [257]. Here, we considered count dataset for HEK cells, as it has expressions of 24176 genes over 1453 cells. Further, at the preliminary stage, we removed the cells whose library size is less than 1500 and also removed the genes which have non-zero expressions in ≤5 cells. Through this process, we selected expression counts of 15524 genes over 1453 cells for further analysis. Here, 1453 cells are clustered into 8 optimum cell clusters (seed=208) (Figure 6.1). Further, for DE analysis, we took cell cluster 8 (537 cells) as group 1 and remaining cell clusters (916 cells) as group 2.

Model formulations

Notations: Let, Z_{ijk}: *rv* representing the true (unknown) read (UMI) counts of k^{th} (k = 1, 2, …, K) gene of j^{th} (j = 1, 2, …, M_i) cell in i^{th} (i = 1, 2, …, N) cell cluster/cell type; K: total number of genes; M_i: number of cells in i^{th} cell cluster; $M(=\sum_{i=1}^{N} M_i)$: total number of cells; N: number of cell clusters; μ_{ijk}: mean of k^{th} gene of j^{th} cell in i^{th} cell cluster for NB distribution; θ_{ijk}: size (=1/dispersion) parameter of k^{th} gene of

j^{th} cell in i^{th} cell cluster for NB distribution; π_{ijk}: mixture probability (*i.e.* the probability for a count to be an excess zero in a cell) parameter for k^{th} gene of j^{th} cell in i^{th} cell cluster (usually cell clusters are detected through clustering).

In bulk RNA-seq, the counts are usually modelled by using a NB distribution. Further, the PMF of the NB distribution is expressed as:

$$f_{NB}(z) = P[Z_{ijk} = z] = \frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z} \quad \forall\, z = 0, 1, 2, \ldots \quad (6.1)$$

where, $\mu_{ijk} \geq 0$; $\theta_{ijk} > 0$ are the parameters of NB distribution, *G*(.): Gamma function. The NB distribution becomes Poisson, when $\theta_{ijk} \to \infty$.

For any $\pi_{ijk} \in [0, 1]$, the true read counts in scRNA-seq study is assumed to follow a ZINB distribution [203,209,211]. Now, the PMF of the ZINB distribution to model the read counts from scRNA-seq data can be given as:

$$P[Z_{ijk} = z] = \begin{cases} \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}} & when\ z = 0 \\ (1-\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}; & z > 0 \end{cases} \quad (6.2)$$

Now, $Z_{ijk} \sim ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$, then the expected value and variance of Z_{ijk} can be obtained as (proof given in Appendix III):

$$E(Z_{ijk}) = (1-\pi_{ijk})\mu_{ijk} \text{ and } V(Z_{ijk}) = (1-\pi_{ijk})\mu_{ijk}\left(1 + \pi_{ijk}\mu_{ijk} + \frac{\mu_{ijk}}{\theta_{ijk}}\right) \quad (6.3)$$

If $\pi_{ijk} = 0$; $ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk}) \to NB(\mu_{ijk}, \theta_{ijk})$

If $\theta_{ijk} \to \infty\ (No\ dispersion)$; $ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk}) \to ZIP(\pi_{ijk}, \mu_{ijk})$

Proposed SwarnSeq Method

Model adjustment for cell capture rates

Theorem: Let, Y_{ijk} be the *rv* for observed (known) read (UMI) counts of k^{th} gene of j^{th} cell in i^{th} cell cluster and ρ_{ijk} be the transcriptional capture rate *rv* for k^{th} gene of j^{th} cell in i^{th} cell cluster. If Z_{ijk} follows a $ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$ distribution, and ρ_{ijk} follows a binomial model with parameter p_{ijk} $(0 \leq p_{ijk} \leq 1)$, then Y_{ijk} will also follow $ZINB$ distribution with parameters $(\pi_{ijk}, \mu_{ijk}p_{ijk}, \theta_{ijk})$.

Proof: Given that, $Z_{ijk} \sim ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$ and $\rho_{ijk} = (Y_{ijk}|Z_{ijk} = z) \sim B(z, p_{ijk})$

Now, the PMF of Z_{ijk} is given in Eq. 6.2 and the PMF of ρ_{ijk} can be expressed in Eq. 6.4:

$$P[Y_{ijk} = y|Z_{ijk} = z] = \binom{z}{y} {p_{ijk}}^{y}(1 - p_{ijk})^{z-y} \quad (6.4)$$

The joint probability distribution of Y_{ijk} and Z_{ijk} can be written as:

$$P[Y_{ijk} = y, Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}]$$

$$= P[Y_{ijk} = y|Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}] \quad (6.5)$$

Now, the marginal probability distribution of Y_{ijk} can be given as:

$$P[Y_{ijk} = y|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] = \sum_z P[Y_{ijk} = y|Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}] \quad (6.6)$$

<u>Case-1</u>: For zero count $(Y_{ijk} = 0)$

$$P[Y_{ijk} = 0|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] = P[Y_{ijk} = 0|Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = 0|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}] + \sum_{z=1}^{\infty} P[Y_{ijk} = 0|Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}]$$

$$= \pi_{ijk} + (1 - \pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}p_{ijk}}\right)^{\theta_{ijk}}$$

$$= \pi_{ijk} + (1 - \pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{\theta_{ijk}} \quad (\mu_{ijk}p_{ijk} = \mu'_{ijk}\,(say)) \quad (6.7)$$

<u>Case-2</u>: For non-zero counts, *i.e.* $Y_{ijk}(>0) = t = 1, 2, 3, \ldots$

$$P\left[Y_{ijk} = t \middle| \pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}\right]$$

$$= \sum_{z \geq t} P\left[Y_{ijk} = t \middle| Z_{ijk} = z, p_{ijk}\right] P\left[Z_{ijk} = z \middle| \pi_{ijk}, \mu_{ijk}, \theta_{ijk}\right]$$

$$= (1 - \pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \sum_{z \geq t} \binom{z}{t} {p_{ijk}}^{t} (1 - p_{ijk})^{z-t} \frac{G(z + \theta_{ijk})}{G(z+1)G(\theta_{ijk})} \left(\frac{\mu_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{z}$$

$$= (1 - \pi_{ijk}) \frac{G(t + \theta_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{\theta_{ijk}} \left(\frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{t}$$

$$= (1 - \pi_{ijk}) \frac{G(t + \theta_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{\theta_{ijk}} \left(\frac{\mu'_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{t} \quad (6.8)$$

Now, Eq. 6.7 and 6.8 are in the form of Eq. 6.2, which indicates the distribution of Y_{ijk} is also $ZINB\left(\pi_{ijk}, \mu'_{ijk}, \theta_{ijk}\right)$. Kindly see Appendix III for proof. When $p_{ijk} = 1$ (under full capture rates), then $ZINB\left(\pi_{ijk}, \mu'_{ijk}, \theta_{ijk}\right) \rightarrow ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$.

GLM Framework in presence of Cell Capture Rates

Here, we estimate the parameters of ZINB model from the observed scRNA-seq count data under a GLM framework. We have shown that the observed scRNA-seq data, Y_{ijk} (for *k*th gene) as a ZINB *rv* with parameters $\boldsymbol{\mu}'_k = (\mu'_{11k}, \ldots, \mu'_{1M_1k}, \ldots \mu'_{N1k}, \ldots, \mu'_{NM_Nk})$; $\boldsymbol{\pi}_k = (\pi_{11k}, \ldots, \pi_{1M_1k}, \ldots, \pi_{N1k}, \ldots, \mu_{NM_Nk})$; $\boldsymbol{\theta}_k = (\theta_{11k}, \ldots, \theta_{1M_1k}, \ldots, \theta_{N1k}, \ldots, \theta_{NM_Nk})$ and further following GLMs (Eq. 6.9 – 6.11) are considered to model these parameters.

$$\boldsymbol{\alpha_k} = \log \boldsymbol{\mu}'_k = \boldsymbol{X\gamma}_k + \boldsymbol{Rw}_k + \boldsymbol{Cs}_k + \boldsymbol{O}_{\boldsymbol{\mu}} \quad (6.9)$$

$$\boldsymbol{\tau_k} = \text{logit}\, \boldsymbol{\pi}_k = = \boldsymbol{X\beta}_k + \boldsymbol{Ru}_k + \boldsymbol{Cv}_k + \boldsymbol{O}_{\boldsymbol{\pi}} \quad (6.10)$$

$$\boldsymbol{\varphi}_k = \log \boldsymbol{\theta}_k \quad (6.11)$$

where, $logit(\pi) = log\left(\frac{\pi}{1-\pi}\right)$; $\boldsymbol{\alpha_k}$, $\boldsymbol{\tau_k}$ and $\boldsymbol{\varphi_k}$**:** *M*×1 vector of parameters for *k*th gene; ***X*****:** *M*×*G* design matrix providing group information (first column consists of 1's to include intercept term); *G*: number of cellular groups (cell clusters are divided into *G* groups, if group is unknown); ***R***: *M*×*N* design matrix providing cell cluster information; ***C***: *M*×*C* design matrix providing cell level auxiliary information; $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta_k}$**:** *G*×*1* vectors of cellular groups effects for *k*th gene; $\boldsymbol{w}_k$ and $\boldsymbol{u}_k$: *N*×*1* vectors of cell cluster effects for *k*th gene; $\boldsymbol{s}_k$ and $\boldsymbol{v}_k$: *C*×*1* vectors of effects for cell level co-variates like cell cycle, cell phase, *etc.* for *k*th gene; *C*: Levels of cell level auxiliaries. $\boldsymbol{O}_{\boldsymbol{\mu}}, \boldsymbol{O}_{\boldsymbol{\pi}}$: offsets for $\boldsymbol{\mu}'_k$ and $\boldsymbol{\pi}_k$respectively.

Estimation of Model Parameters with EM Algorithm

The parameters in Eq. 6.9 – 6.11 for *k*th gene, *i.e.* $\boldsymbol{\Omega}_k = \{\boldsymbol{\alpha}_k, \boldsymbol{\tau}_k, \boldsymbol{\varphi}_k\}$ can be estimated by using the MLE Method. However, no closed form solutions exist for the resulting log-likelihood equation in Eq. 6.12 Hence, we developed an EM algorithm to estimate the parameters for the given observed scRNA-seq count data, *i.e.* $Y_{ijk} = y_{ijk}$. Now, the incomplete data likelihood function for *k*th gene can be expressed as:

$$L\left(\boldsymbol{\Omega}_k; Y_{ijk} = y_{ijk}\right) = \prod_{i=1}^{N} \prod_{j=1}^{M_i} \left\{\pi_{ijk} \delta_0\left(y_{ijk}\right) + \left(1 - \pi_{ijk}\right) f_{NB}\left(y_{ijk}\right)\right\} \quad (6.12)$$

Further, the EM algorithm recasts the ZINB model into a missing data problem by introducing a latent *rv*, V_{ijk}. Now, the V_{ijk} can be defined as:

$$V_{ijk} = \begin{cases} 1 & \text{if the observed count data comes from the zero componet} \\ 0 & \text{if the observed count data comes from the count component} \end{cases}$$

Now, the joint likelihood function for complete data, *i.e.* (Y_{ijk}, V_{ijk}) can be expressed in Eq. 6.13, as:

$$L(\boldsymbol{\Omega}_k; Y_{ijk}, V_{ijk}) = \prod_{i=1}^{N}\prod_{j=1}^{M_i}\left[\left\{\pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right\}^{V_{ijk}}\left\{(1-\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{y_{ijk}}\right\}^{1-V_{ijk}}\right] \quad (6.13)$$

Then, the log-likelihood function in Eq. 6.13 becomes:

$$l(\boldsymbol{\Omega}_k; Y_{ijk}, V_{ijk}) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} V_{ijk} log\left\{\pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right\} + \sum_{i=1}^{N}\sum_{j=1}^{M_i}(1-V_{ijk}) log\left\{(1-\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{y_{ijk}}\right\} \quad (6.14)$$

$$= l_1(\boldsymbol{\Omega}_k; V_{ijk}) + l_2(\boldsymbol{\Omega}_k; Y_{ijk}, V_{ijk}) \quad (6.15)$$

where, $l_1(.)$: log-likelihood due to the zero-component of the model and $l_2(.)$: log-likelihood due to the count-component of the model. Now, the expected value of the log-likelihood function, Eq. 6.14, can be expressed as:

$$Q = E\left[l(\boldsymbol{\Omega}_k; Y_{ijk}, V_{ijk})\right] = \sum_{i=1}^{N}\sum_{j=1}^{M_i} E(V_{ijk}|Y_{ijk}, \boldsymbol{\Omega}_k) log\left\{\pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right\} + \sum_{i=1}^{N}\sum_{j=1}^{M_i} w_{ijk} log\left\{(1-\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{y_{ijk}}\right\} \quad (6.16)$$

Further, the posterior probabilities in Eq. 6.16 for the observations originate from the count component of the model can be given as:

$$w_{ijk} = P\left[V_{ijk} = 0 | Y_{ijk}, \boldsymbol{\Omega}_k\right] = \frac{(1-\pi_{ijk}) f_{NB}(y_{ijk};\mu_{ijk},\theta_{ijk})}{\pi_{ijk}\delta_0(y_{ijk}) + (1-\pi_{ijk}) f_{NB}(y_{ijk};\mu_{ijk},\theta_{ijk})} \quad (6.17)$$

where, $f_{NB}(.)$ is the PMF of NB distribution given in Eq. 6.1.

A. E-step: The E-step in the EM algorithm involves evaluating the expected value of the log-likelihood of the complete data (Eq. 6.16), given the observed data with the current estimates of the parameters. In our proposed approach, for each gene, given the observed data and a current estimate of the ZINB parameters, the expected value of the log-likelihood is calculated. Let, $\widehat{\boldsymbol{\Omega}}_k^{\,c} = \{\widehat{\boldsymbol{\alpha}}_k^{\,c}, \widehat{\boldsymbol{\tau}}_k^{\,c}, \widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}}^{\,c}\}$ be the given current estimate of the parameters, then the expected value of log likelihood in Eq. 6.16 at step (c + 1), *i.e.* Q^{c+1} is calculated. The conditional expectation, *i.e.* $E(V_{ijk}|Y_{ijk}, \widehat{\boldsymbol{\Omega}}_k^{\,c})$ in Eq. 6.16 can be given as:

$$E(V_{ijk}|Y_{ijk}, \widehat{\boldsymbol{\Omega}}_k^{\,c}) = \frac{\hat{\pi}_{ijk} + (1-\hat{\pi}_{ijk})\left(\frac{\hat{\theta}_{ijk}}{\hat{\theta}_{ijk}+\hat{\mu}_{ijk}}\right)^{\hat{\theta}_{ijk}}}{\hat{\pi}_{ijk}\delta_0(y_{ijk}) + (1-\hat{\pi}_{ijk}) f_{NB}(y_{ijk}|\hat{\mu}_{ijk},\hat{\theta}_{ijk})} \quad (6.18)$$

***B. M*-step:** Maximize Q^{c+1} to update the parameter estimates

i. The parameters from the count component of the model, $\{\widehat{\boldsymbol{\mu}}'_k, \widehat{\boldsymbol{\theta}}_k\}$ are updated within the GLM framework, and can be expressed as:

$$\log \boldsymbol{\mu}'_k = \boldsymbol{X\gamma}_k + \boldsymbol{Rw}_k + \boldsymbol{Cs}_k + \boldsymbol{O}_{\boldsymbol{\mu}} \quad (6.19)$$

The updated value of the estimates of parameters at step (c + 1) is obtained by providing the observation wise weights, $\widehat{w}_{ijk}{}^{(c)}_{ijk}$, given in Eq. 6.17 and parameters estimates at c-step. For this purpose, the *glm.nb* function in MASS R package is executed.

ii. The zero-inflation probability, $\hat{\pi}_{ijk}$, is updated with the logistic regression, can be expressed as:

$$logit(\boldsymbol{\pi}_k) = \boldsymbol{X\beta}_k + \boldsymbol{Ru}_k + \boldsymbol{Cv}_k + \boldsymbol{O}_{\boldsymbol{\pi}} \quad (6.20)$$

The updated value of $\hat{\pi}_{ijk}$ at step (c + 1) is obtained by incorporating the observation level weights, $\hat{w}_{ijk}{}_{ijk}^{(c)}$given in Eq. 6.17 and the parameters estimate at c-step. For this, *glm*(…, family= 'binomial') function in stat R package is executed.

C. Starting values for EM algorithm

The success of an iterative algorithm, *e.g.* EM, depends on the provision of supplying initial values for the parameters. In our SwarnSeq method, we provide the initial values for the estimators for each gene by estimating through Generalized Linear (GL) Poisson and GL Binomial models for non-zero and zero counts, respectively. For this purpose, the *glm* function implemented in stats package is executed.

D. Assessing convergence

The EM algorithm iterates over an Expectation (E) step and Maximization (M) step for each gene until convergence achieved [203,210,281]. Let, $\widehat{\boldsymbol{\Omega}}_k{}^c = \{\widehat{\boldsymbol{\alpha}}_k{}^c, \widehat{\boldsymbol{\tau}}_k{}^c, \widehat{\boldsymbol{\varphi}}_{\boldsymbol{k}}{}^c\}$ be the vector parameter estimates for *k*th gene. The criteria for convergence can be expressed as:

$$\left|Q\left(\widehat{\boldsymbol{\Omega}}_k{}^{c+1}; Y_{ijk}, V_{ijk}\right) - Q\left(\widehat{\boldsymbol{\Omega}}_k{}^c; Y_{ijk}, V_{ijk}\right)\right| < \epsilon \quad (6.21)$$

where, ϵ is the threshold for convergence (*e.g.* in SwarnSeq R package, the default for $\epsilon = 10^{-10}$ and maximum iteration is 10^3). It is worthy to note that for some genes the EM algorithm may fail to converge or may be not successful; therefore, we used Nelder's optimization algorithm [282] implemented in *optim* function of stats R package to estimate the MLE of parameters.

Differential expression analysis

The gene-wise mean parameter depends on the cellular groups through the model given in Eq. 6.9. Further, the factors such as cell clusters and cell co-variates are included in the model to remove the unwanted effects. For DE analysis, two group comparisons are made and the model in Eq. 6.9 can be expanded as:

$$log(\mu_{ijk}) = \gamma_{0k} + \gamma_{1k}x_{ijk} + w_{1k}r_{1jk} + \cdots + w_{Nk}r_{Njk} + s_{1k}c_{1jk} + \cdots + s_{mk}c_{mjk} + O_{\mu_k} \quad (6.22)$$

where, x_{ijk}: binary indicator for cellular group membership, γ_{0k}: (intercept term) logarithm of mean parameter for gene *k* in the reference cellular group, γ_{1k}: log FC parameter for gene *k*, w_{ik}: regression co-efficient for *i*th cell cluster for *k*th gene, r_{ijk}: indicator variable for cell cluster membership of *j*th cell in *i*th cluster for *k*th gene, s_{mk}: regression co-efficient for *m*th cell co-variates of *k*th gene, c_{mjk}: indicator variable for *m*th co-variates of *j*th cell for *k*th gene and O_{μ_k}: offset term.

To decide whether, the *k*th gene is DE or not, the following hypotheses are tested.

$$H_0: \gamma_{1k} = 0 \; vs. \; H_1: \gamma_{1k} \neq 0$$

The above test can be performed by using LRT statistic, and can be expressed as:

$$DS_k = -2\{l(\boldsymbol{\Omega}_k = \widehat{\boldsymbol{\Omega}}_{k0}) - \; l(\boldsymbol{\Omega}_k = \widehat{\boldsymbol{\Omega}}_k)\} \quad (6.23)$$

where, $\widehat{\boldsymbol{\Omega}}_{k0}$: MLE of $\boldsymbol{\Omega}_k$ for *k*th gene under the constraint of H_0 and $\widehat{\boldsymbol{\Omega}}_k$: unconstrained MLE of $\boldsymbol{\Omega}_k$ for *k*th gene. The test statistic, DS_k, follows a Chi-square distribution with 1 degree of freedom (for 2 groups) under H_0. Further, based on the distribution of DS_k, the *p-value*, adjusted *p-value* and FDR for *k*th gene can be computed after adjustment for multiple hypothesis testing. Then, the DE genes were detected through the computed FDR or adjusted *p-values* of the genes.

Testing for differential zero inflation

Through LRT statistic(s), we have shown that genes in scRNA-seq data are highly zero inflated (Figure 5.3 in Chapter 5). Therefore, to facilitate DZI analysis in the SwarnSeq method, the gene-wise zero inflation parameter depends on the cellular groups through the model given in Eq. 6.10. Further, factors such as cell clusters and other cell-level auxiliaries are included in the model to remove the unwanted effects. For DZI analysis, two group comparisons are made and the model in Eq. 6.10 can be written as:

$$logit(\pi_{ijk}) = \beta_{0k} + \beta_{1k}x_{ijk} + u_{1k}r_{1jk} + \cdots + u_{Nk}r_{Njk} + v_{1k}c_{1jk} + \cdots + v_{mk}c_{mjk} + O_{\pi_k} \quad (6.24)$$

where, x_{ijk}: binary indicator for cellular group membership, β_{0k}: logarithm of mean parameter for gene *k* in the reference cellular group, β_{1k} is the *logFC* parameter for gene *k*, u_{ik}: the regression co-efficient for i^{th} cell cluster for k^{th} gene, r_{ijk}: indicator variable for cell cluster membership of j^{th} cell in i^{th} cluster for k^{th} gene, v_{mk}: regression co-efficient for m^{th} cell co-variates of k^{th} gene, c_{mjk}: indicator variable for m^{th} co-variates of j^{th} cell and O_{π_k}: offset term.

To decide whether, k^{th} gene is DZI or not, the following hypotheses are tested.

$$H_{10}: \beta_{1k} = 0 \; vs. \; H_1: \beta_{1k} \neq 0$$

A similar test statistic to that given in Eq. 6.23 can also be developed for testing of DZI of genes.

Classification of influential genes

The SwarnSeq method can divide the detected influential genes into different classes, as shown in Table 3. For instance, the $H_0: \gamma_{1k} = 0$ detects all the genes

that are DE across two cellular groups, while $H_{10}: \beta_{1k} = 0$ detects the DZI genes. Further, it detects the first type of influential genes with both H_0 and H_{10} rejected, which indicates there is a significant difference in the number of cells with zero values for genes across the cellular groups, but the (non-zero counts) expressions in the remaining cells also show significant differences. We call such a group of influential genes as 'DEZIG'.

Table 6.2. Classification of influential genes using SwarnSeq method.

		Differentially Expressed	
Differentially Zero Inflated		Yes	No
	Yes	DEZIG	DZIG
	No	DEG	None

DEZIG: Differentially Expressed and Differentially Zero Inflated Genes;
DZIG: Differentially Zero Inflated Genes; DEG: Differentially Expressed Genes

The second type of genes are those for which H_0 is rejected, but H_{10} is not. This means that there is no significant difference in the number of cells whose expressions are zeros across cellular conditions for genes, but they are expressed differentially. We call this type of genes as 'DEG'. Further, the third type (DZIG) of genes is that for which H_{10} is rejected, but H_0 is not. It includes genes for which, there is a significant difference in the number of cells with real zero values across the two cellular conditions, but the expression in the remaining cells shows no significant difference.

Estimation of capture rates parameter

The distribution of the observed scRNA-seq read counts (UMI) depends on the cell specific transcriptional efficiency parameter, p_{ijk}. For computational simplicity, we assume $p_{ij1} = p_{ij2} = \cdots = p_{ijK} = p_{ij}$, *i.e.,* the cell specific efficiency parameters

remain same across all the genes. The proposed procedure for estimation of cell capture rate parameters is described as follows.

Case 1: When RNA spike-ins are available

Suppose *n* RNA spike-ins are added to each cell's lysate and spike-in transcripts are processed in parallel, which will result a set of read (UMI) read counts for spike-in transcripts. Let, $C_1, C_2, \ldots, C_n$ be the respective *m*RNA concentrations of *n* spike-in transcripts added to j^{th} (j=1, 2, …, M_i) cell of i^{th} (i=1, 2, …, N) cell cluster and $R_{ij1}, R_{ij2}, \ldots, R_{ijn}$ be the observed UMI counts of the *n* RNA spike-in transcripts for j^{th} cell. Now, the transcriptional capture rate for j^{th} cell in i^{th} cell cluster can be estimated through a linear regression equation, given in Eq. 6.25.

$$R_{ijk} = p_{ij0} + p_{ij1} C_k + \epsilon_k \tag{6.25}$$

Here, $\hat{p}_{ij1}$, regression co-efficient, is the estimate of the capture rate for j^{th} cell in i^{th} cell cluster.

Case 2*: *When RNA spike-ins are not available

Transcriptional capture efficiency parameters of cells are the key factors for variation in the observed cell specific library sizes [260]. Hence, the observed cell library sizes can be used to empirically compute the cell specific capture rate, which is given as:

Let, $(\rho_1,\ \rho_2)$ be the range of capture rates and S_{ij} be the library size of j^{th} cell in i^{th} cell cluster and , $L_{ij} = log_{10}(S_{ij})\ \forall\ i, j$

Now, $L_{min} = \min_j L_{ij}$ and $L_{max} = \max_j L_{ij}$

$$\hat{p}_{ij} = \rho_1 + (\rho_2 - \rho_1)\frac{L_{ij} - L_{min}}{L_{max} - L_{min}} \tag{6.26}$$

Determination optimal number of cell clusters from scRNA-seq data

Let, $Y_{ij.}$: mean expression value of j^{th} cell in i^{th} cell cluster, $Y_{i..}$: mean expression value of i^{th} cell cluster, and $Y_{...}$ be the over-all mean. Then, the Total Sum of Squares (TSS) can be expressed as:

$$TSS = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(Y_{ij.} - Y_{...}\right)^2$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(Y_{ij.} - Y_{i..}\right)^2 + \sum_{i=1}^{N} M_i \left(Y_{i..} - Y_{...}\right)^2$$

$$= WSS + BSS \tag{6.27}$$

where, WSS: Within cluster Sum of Squares, BSS: Between cluster Sum of Squares. Now, the proposed index can be given as:

$$r_h = \frac{WSS}{BSS} \tag{6.28}$$

For different values of number of clusters (*h*) in the scRNA-seq data, the *r*-measure is computed through Eq. 6.28. The *h* value which provides the maximum value for *r*, can be chosen as the estimator for optimum cell clusters for the observed scRNA-seq data. The optimum number of cell clusters for all the 10 datasets are determined through the above technique, and the results are shown in Figure 6.1.

Performance evaluation metrics

The performance of SwarnSeq method for identifying genuine DE genes was evaluated with respect to 11 existing competitive methods using the AUROC (*i.e.* TPR *vs.* FPR), and other performance metrics on 10 real scRNA-seq datasets (see Materials section). These metrics include TP, FP, TN, FN, TPR, FPR, PPR, FDR, NPV, ACC, and F1, which are defined in Eq. 5.43 – 5.50 of the Chapter 5.

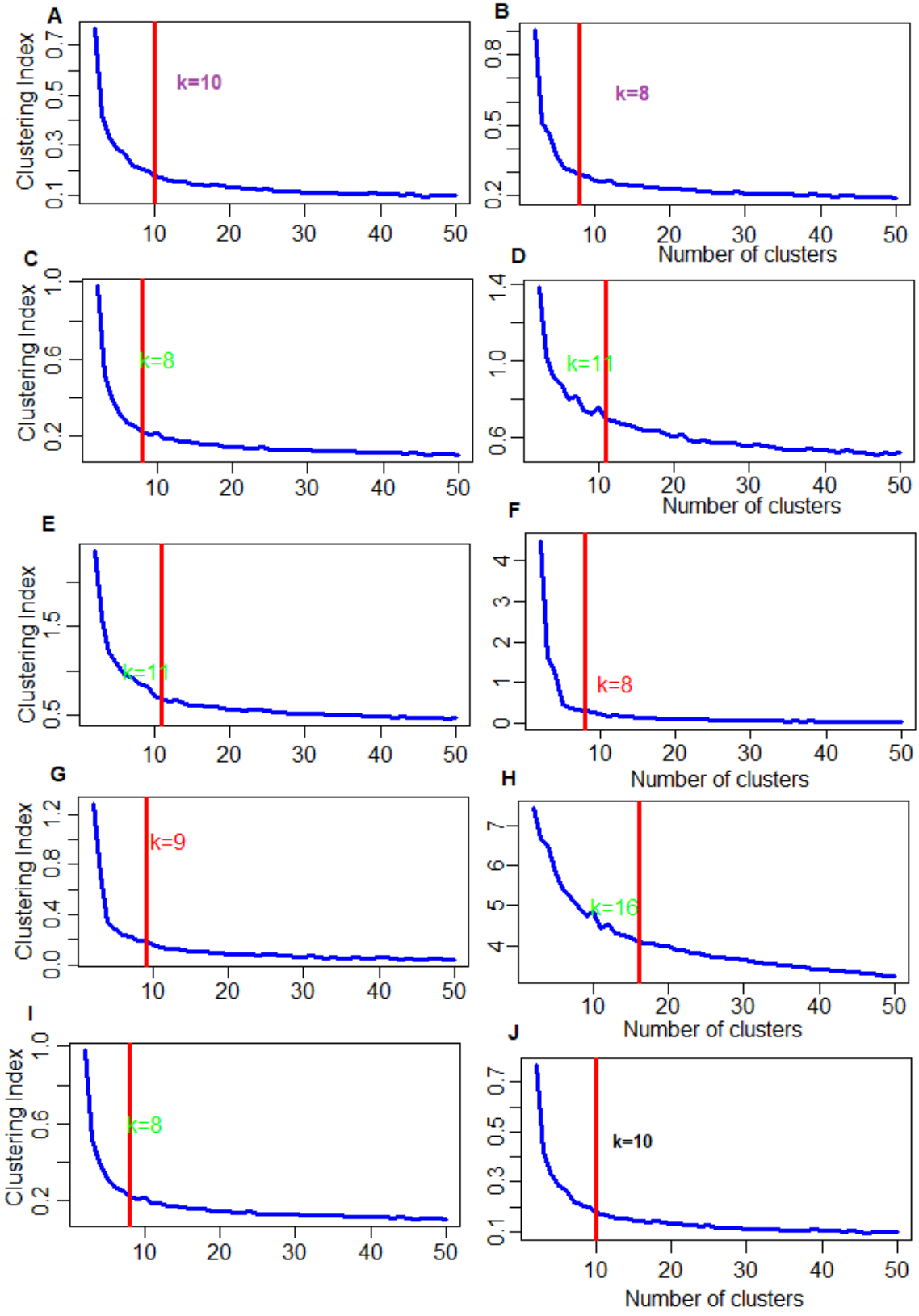


Figure 6.1. Cluster analysis and determination of optium number of cell clusters

for real scRNA-seq datasets. Clustering analysis is performed on the scRNA-seq datasets through k-means algorithm. The optimum number of cell clusters is determined for each scRNA-seq datasets through above proposed method (Eq. 6.28) implemented in *OptimCluster* function implemented in SwarnSeq R package. X-axis represents number of cell clusters. Y-axis represents clustering index. Here, the number of cell clusters is kept in the range of [2, 50]. Vertical lines represent the optimum number of cell clusters with k value. The graphs are shown for for (A) GSE53638 (Data 1); (B) GSE77728; (C) GSE53638 (Data 3); (D) GSE53638 (Data 2); (E) GSE29087; (F) GSE65525 (G) GSE111108; (H) GSE92495; (I) GSE115469; (J) GSE109999.

Results

Preliminary analytical results

In the Chapter 5, we considered two publicly available zero inflated and overdispersed datasets to show the suitability and goodness of fit of different count data models, *viz.* NB, ZINB, PD, HD and ZIPD [270,271]. Besides, in this chapter we used one real scRNA-seq data to fit the above models and the results are shown in Table 6.3.

Table 6.3. Fitting of well-known discrete models to scRNA-seq read count data.

Reads	Obs. Freq.	Pred. Freq. NB	Pred. Freq. PD	Pred. Freq. HD	Pred. Freq. ZINB	Pred. Freq. ZIP
0	115	108.05	0.09	4.82	126.82	115
1	84	57.92	0.78	3.34	56.96	0.06
2	45	42.58	3.37	20.33	39.79	0.36
3	33	34.13	9.73	13.54	31.11	1.34
4	31	28.48	21.03	42.8	25.61	3.73
5	18	24.34	36.39	27.48	21.73	8.32
6	12	21.13	52.46	59.94	18.79	15.44
…	…	…	…	…	…	…
19	7	5.41	0.47	2.74	5.2	3.71
20	5	4.95	0.2	2.19	4.8	2.06
21	5	4.53	0.08	1.11	4.43	1.09
22	4	4.15	0.03	0.83	4.1	0.55
23	5	3.8	0.01	0.41	3.8	0.27
24	6	3.49	0	0.29	3.53	0.12
25	4	3.21	0	0.14	3.27	0.06
26	7	2.95	0	0.09	3.04	0.02
…	…	…	…	…	…	..
29	3	2.29	0	0.01	2.45	0
42	3	0.81	0	0	1.01	0

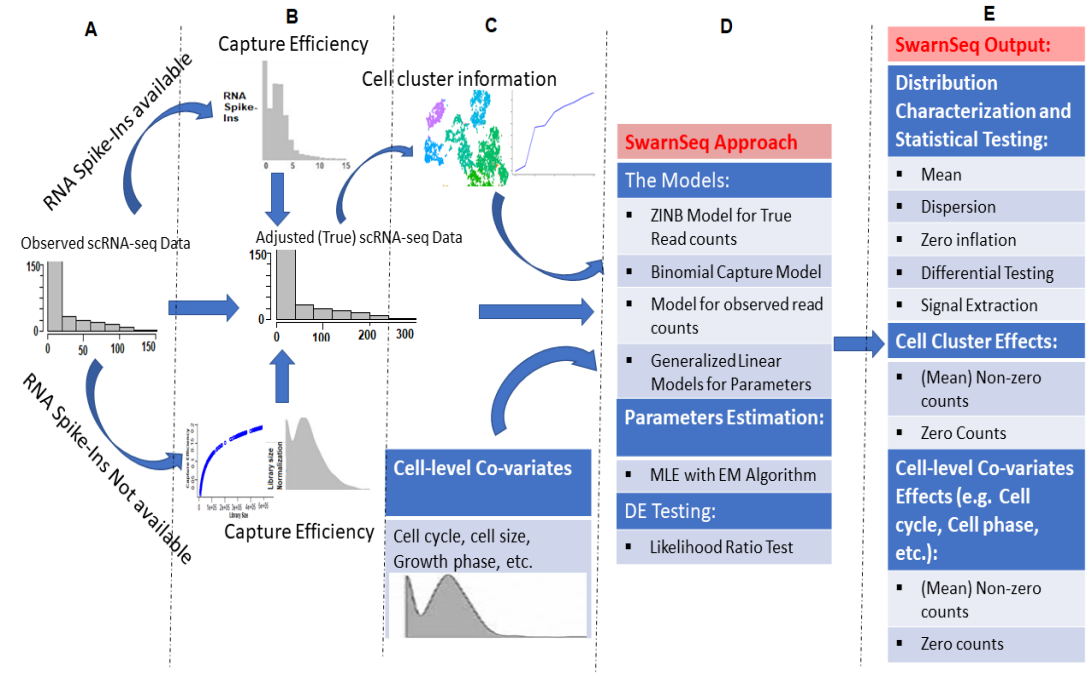
43	1	0.75	0	0	0.94	0
46	3	0.59	0	0	0.78	0
47	1	0.55	0	0	0.73	0
49	2	0.47	0	0	0.64	0
50	2	0.44	0	0	0.6	0
Parameters (MLE)		μ=8.14 θ=0.574	μ=8.65	μ=8.651 φ=1.92	μ=8.652 θ=0.47377 π=1.17e-05	μ=11.1373 π=0.224

Our analytical results indicated that the expected frequencies computed from ZINB are much closer to their observed counter parts as compared to other models (Table 6.3). At this preliminary stage, we can infer that ZINB model best suits to the zero inflated and overdispersed scRNA-seq data as compared to NB model extensively used in RNA-seq data analysis.

Proposed Model overview

Figure 6.2 gives an overview of the SwarnSeq method framework. The observed counts in the scRNA-seq are noisy reflection of the true expression of genes due to low transcriptional capturing (Figures 5.1, 6.2). We modelled the observed read counts, Y_{ijk} of k^{th} gene in j^{th} cell in i^{th} cluster, as the joint distribution of k^{th} gene's true expression Z_{ijk} and transcriptional capture rate (p_{ijk}) of j^{th} cell in i^{th} cluster. In other words, after incorporating the transcriptional capturing procedure in the modeling process, the mean of non-zero counts in ZINB distribution depends on cell capturing rate parameter.

Figure 6.2. Illustration of the operational framework of the SwarnSeq method. (A) cross-cell distribution of observed scRNA-seq counts; (B) cross-cell distribution of true/adjusted scRNA-seq counts with capture efficiency with respect to spike-ins information; (C) Auxiliary information such as cell cluster and cell level co-variates as inputs to the SwarnSeq; (D) Details of SwarnSeq method fitted for each gene; (E) For each gene, the output of SwarnSeq includes the distribution characterization (*i.e.* mean, dispersion and zero inflation) over cell populations, differential expression testing between two cell populations, differential zero inflation testing between two cell populations, effects of cell clusters on zero-inflation parameter and mean of non-zero counts, effects of cell level auxiliary information on zero-inflation parameter and mean of non-zero counts.



The relation between the capture efficiency with the distribution of the observed read counts is shown in Figure 6.3.

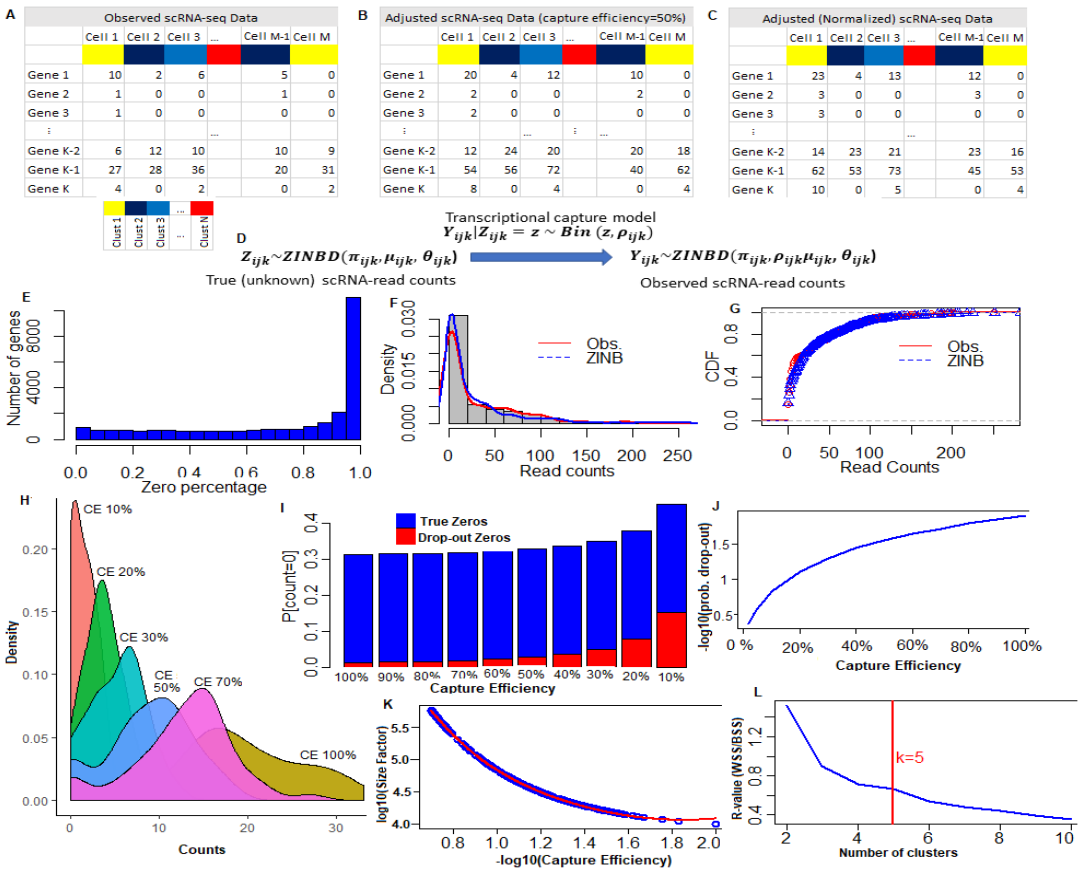


Figure 6.3. Data structures, Models, and Distributions used in the SwarnSeq method. (A) Structure of the observed scRNA-seq data; (B) Input structure of the true scRNA-seq data adjusted with capture efficiency; (C) Input structure of the normalized true scRNA-seq data; (D) ZINB and transcriptional capture models used in SwarnSeq approach; (E) Histogram of zero percentages of all expressed genes in a real scRNA-seq dataset; (F) An example of ZINB model fitting for scRNA-seq data. The fitting of observed and theoretical ZINB models are shown for real scRNA-seq data for a gene; (G) Cumulative distribution function fitting for observed and theoretical ZINB models; (H) Theoretical ZINB distribution of observed scRNA-seq counts of a gene with different random capture efficiency. The distributions are shown for capture efficiencies 100%, 70%, 50%, 30%, 20% and 10%. Here, the 100% capture efficiency represents the distribution of true scRNA-seq counts; (I) The histograms of zero probabilities for different capture efficiencies are shown. The red color bars represent the probability density of real true zero expressions. The blue bars represent the probability density of the NB part of the ZINB model. (J) The plot shows the relation between the probability of drop-out events and capture efficiencies of cells. (K) The relation between the library sizes and the capture efficiencies of the cells is shown. (L) Deciding the number of optimum cell clusters for a real scRNA-seq data. CE: Capture Efficiency.

The relation among means of count part in ZINB model before and after incorporation of the transcriptional capturing procedure is found to be $\mu_{ijk} > \mu'_{ijk}$ (Eq. 6.29). In other words, the distribution of observed scRNA-seq read counts shift more towards zeros after incorporation of the transcriptional capturing process (Figure 6.3). This means that more zeros are found in observed data and will be from the count part of the model. Further, the expected value and variance of observed read counts of genes depends (*i.e.*, directly proportional) on the cell capture rate (Appendix IV) and can be expressed in Eq. 6.29 and 6.30. Here, when p_{ijk} becomes smaller both mean and variance of Y_{ijk} also becomes smaller.

$$E(Y_{ijk}) = (1 - \pi_{ijk})\mu_{ijk}p_{ijk} \tag{6.29}$$

$$V(Y_{ijk}) = (1 - \pi_{ijk})\mu_{ijk}p_{ijk}\left(1 + \pi_{ijk}\mu_{ijk}p_{ijk} + \frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}}\right) \tag{6.30}$$

The relations between the expected value and variance of the observed read counts, the estimated parameters of the SwarnSeq Model, Eq. 6.5 – 6.8, and the cell capture parameters (Eq. 6.25, 6.26) are shown in Figures 6.4 – 6.6.

Figure 6.4. Relation among expected value, variance, and co-efficient variation of the SwarnSeq model. (A) Expected value vs. Variance plot. The relation between the expected value and variance of the SwarnSeq model given in Eq. 6.29, 6.30 is shown. X-axis represents the log transformed expected value and Y-axis represents the log transformed variance. (B) Expected

value vs. Co-efficient of Variation (CV) plot. X-axis represents the log transformed expected value and Y-axis represents the log transformed value of CV. (C) CV vs. Zero inflation plot. The relation between the CV and Zero inflation probability parameter of the SwarnSeq model is shown. X-axis represents the log transformed CV and Y-axis represents the log transformed value of Zero inflation probability. (D) CV vs. Dispersion plot. X-axis represents the log transformed CV and Y-axis represents the log transformed value of Dispersion. (E) Variance vs. Zero inflation plot. X-axis represents the log transformed Variance and Y-axis represents the log transformed value of Zero inflation probability. (F) Variance vs. Dispersion plot. X-axis represents the log transformed Variance and Y-axis represents the log transformed value of the dispersion

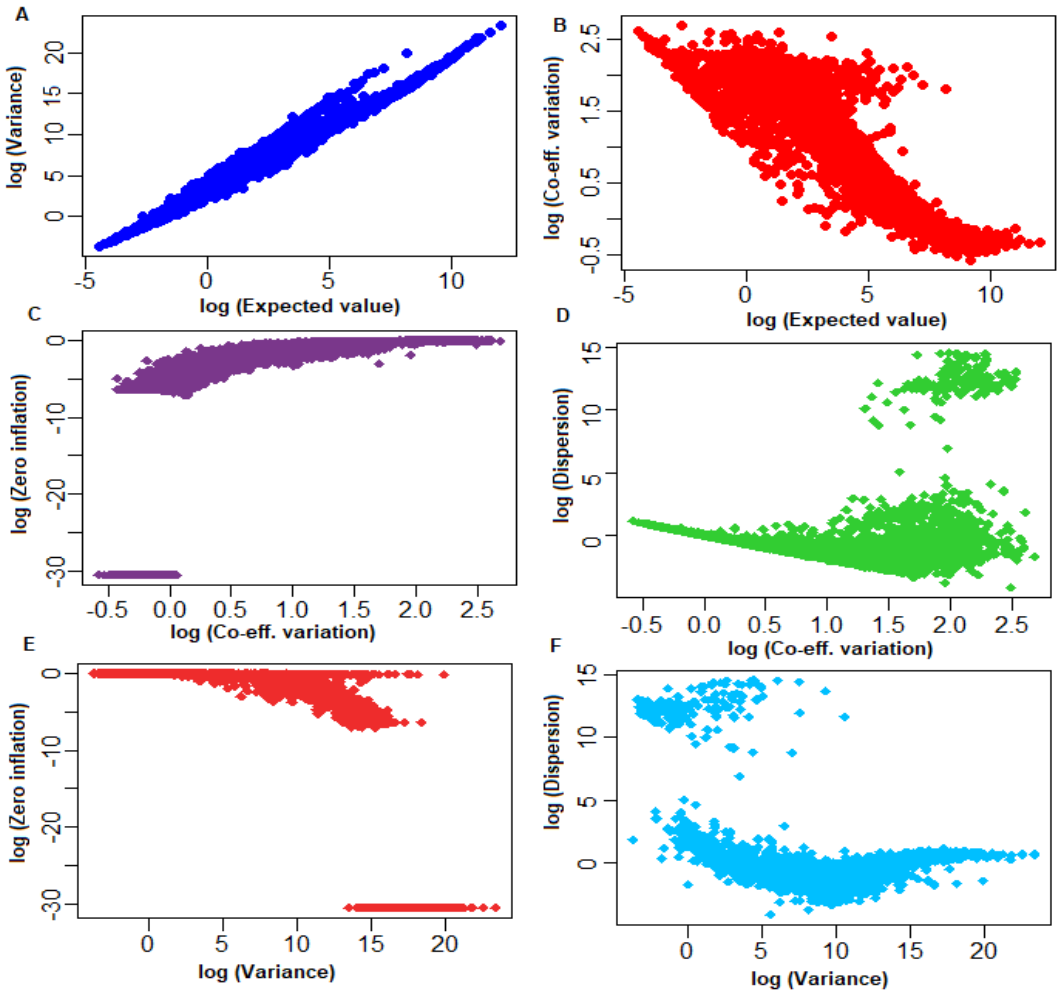


Figure 6.5. Relation among gene parameter estimates of SwarnSeq model. Here, the parameters of the ZINB model shown in Eq. are estimated through MLE method. (A) Mean of count parts vs. Dispersion plot. X-axis represents the log transformed value of mean count parts and Y-axis represents the log transformed value of the estimated dispersion parameter. (B) Mean of count parts vs. Zero inflation probability plot. The relation between the mean count parts (NB part) with the zero inflation probability parameters of the SwarnSeq model is shown. (C) Dispersion vs. Zero inflation parameter plot. The relation between the dispersion parameter from the NB part of the ZINB model with zero inflation probability parameters of the SwarnSeq model is shown. (D) Observed zero proportions vs. Estimated zero inflation probability plot. The relation between the Zero inflation probability parameter of the SwarnSeq model with the observed zero proportions present in the scRNA-seq data is shown (E) Observed zero proportions vs. (Observed zero - zero inflation) probability plot. The relation between the observed zero proportions present in the scRNA-

seq data and the (Observed zero - zero inflation) probability is shown (F) Dropout zeros vs. True Zeros plot. The relation between the Dropout zeros, i.e. excess zeros present in scRNA-seq data modeled through Dirac's delta function in Eq. 5.3, and the True zeros from the NB model is shown.

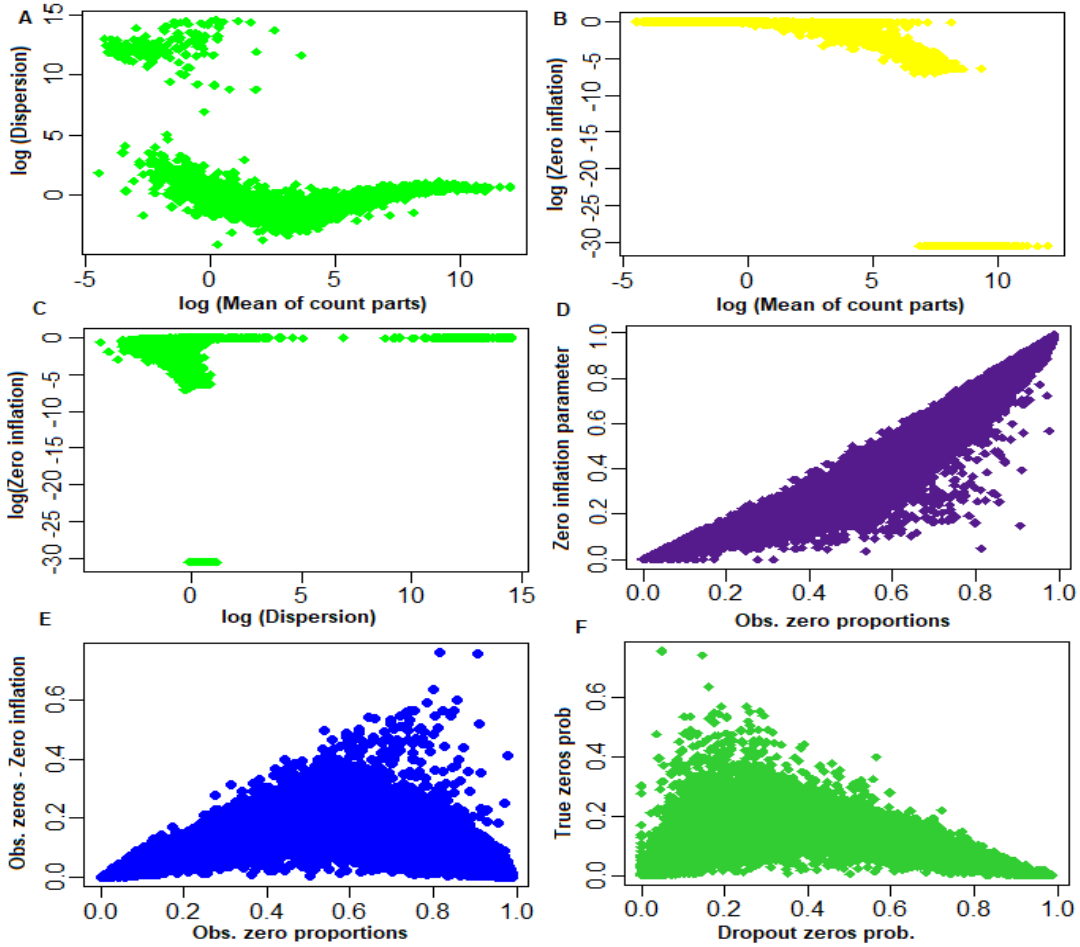
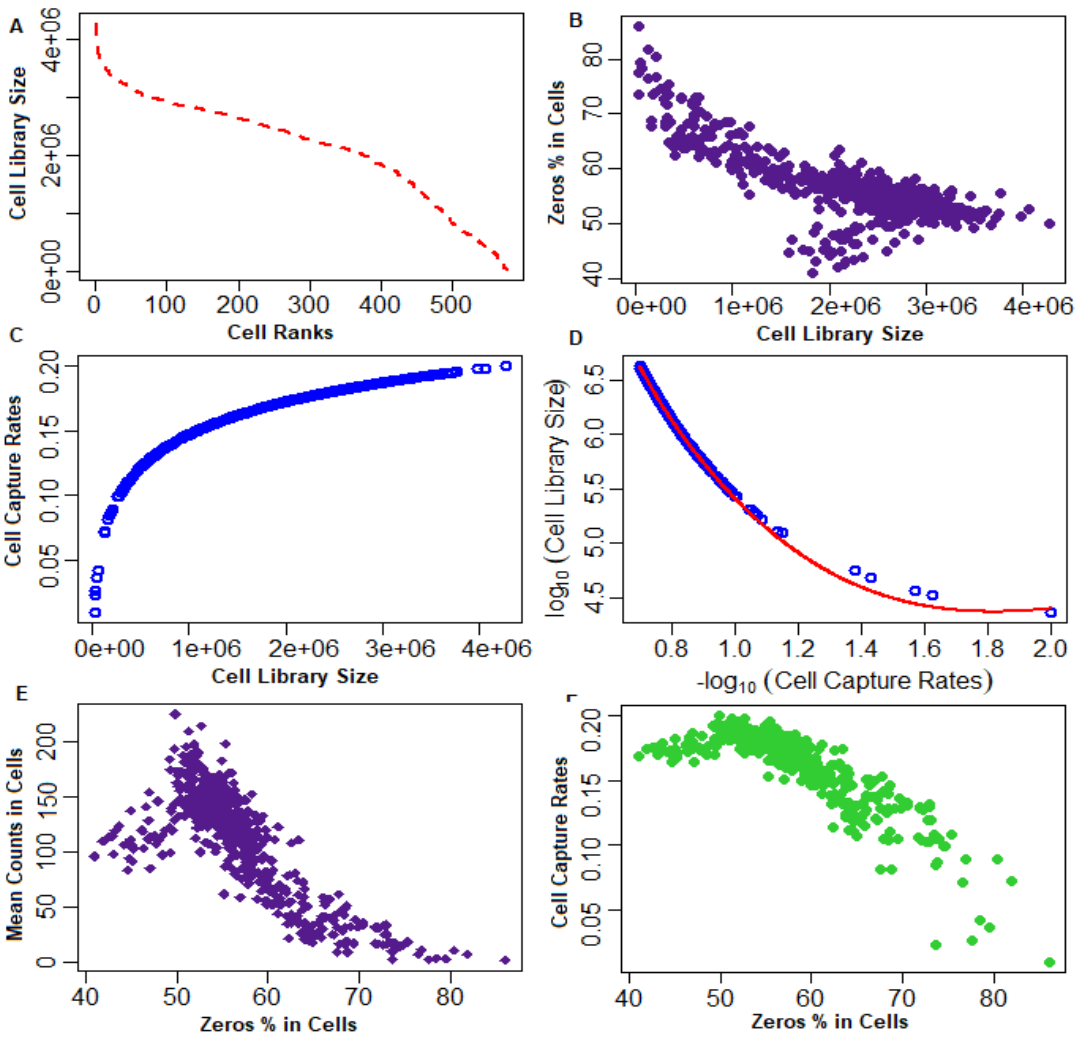


Figure 6.6. Relation among cell specific parameters estimated through the SwarnSeq model. (A) Cell library size distribution over cells. In X-axis, the ranks of the cells are shown, and Y-axis shows the library sizes of the cells. Here, the underlying distribution is S-shaped. (B) Zero inflation vs. Cell library sizes plot. Relationship between the cell library sizes and the zero counts in cell are shown. Here, X-axis represents the cell library sizes and Y-axis represents the zero counts percentages in the cells. It can be shown that every cell has higher zero counts (>40%) as expression of genes, due to the availability of lower concentration of mRNA molecules. Further, the cell library sizes are inversely proportional to the zero percentage in cells. In other words, the cells with higher library sizes contain lesser percentages of zeros and *vice-versa.* (C) Cell library size vs. Cell capture rates plot. The relation between the library sizes and the capture efficiencies of the cells estimated from SwarnSeq model is shown. The library sizes and mRNA molecules capture rates of the cells are represented in X-axis and Y-axis, respectively. Here, the library sizes are directly proportional to the capture rate of cells (D) The log-transformation of cell library sizes is plotted with log transformation of cell capture rates and a curve is fitted shown in red color. (E) Mean non-zero counts of cells *vs.* zeros percentage in cells plot. The relation between the mean non-zero counts and zero percentages in cells is shown. Here, the relation between the zero percentages and mean non-zero counts in cells is reciprocal. (F) Zeros percentage in cell vs. Cell capture rates plot. The relation between the zeros present in the cell with the cell capture rates is shown. X-axis represents the Zeros percentage in cell and Y-axis represents the cell capture rates.

Here, the relation is inversely proportional, means cells with higher capture rates have lesser zeros as counts expression the cell.



The mixture probability and dispersion parameters (Eq. 6.29 and 6.30) for the observed read counts remain unchanged after the incorporation of transcriptional capture efficiency parameter in the modelling process. For instance, when $p_{ijk} = 1$ (100% capture), the genes in a cell will have zero counts which are not truly expressed (*i.e.* biological zeros); this is expected under a perfect deep sequencing scenario. In other words, observed read counts are the true expected counts of

genes in a cell under a perfect deep sequencing scenario. When $p_{ijk} < 1\ (real\ case)$, the zeros in the observed scRNA-seq read counts are the mixture of drop-out and true zeros. It may be noted that π_{ijk} remain unaffected by the capture rate parameter, hence, the $\hat{\pi}_{ijk}$ from observed data can be used to measure the proportions of true zeros of genes in the data.

SwarnSeq allows the modeling of the effects of cellular groups, cell clusters and other cell-level covariates on both the zero-inflation probability and mean of non-zero read counts. When cell level auxiliary information is specified, SwarnSeq uses a log-linear model for the covariate effect on mean and a logit model for the covariate effect on zero-inflation in a GLM framework. Further, SwarnSeq performs for DE analysis of genes for a given two groups situation and can be generalized to multiple groups situation. The DZI analysis of genes of scRNA-seq data is allowed in the SwarnSeq method, which leads to the identification of severely zero-inflated genes over the cellular populations. Additionally, genes in scRNA-seq data are classified into different gene types based on DE and DZI analysis (Table 6.2).

SwarnSeq as Differential Expression tool

We benchmarked the proposed SwarnSeq method against 11 existing methods for DE analysis (described in Chapter 5) on a wide range of real scRNA-seq datasets. The problem in benchmarking of scRNA-seq DE methods on real scRNA-seq datasets is the unavailability of reference genes. Hence, to obtain a credible list of reference genes, we used FC criterion (*i.e.* ratios of mean expressions of genes over the two groups). For each of the 10 datasets, we selected the top 3000 genes based on the FC criterion as reference gene lists.

The 12 scRNA-seq DE methods, including SwarnSeq, were benchmarked using all the 10 datasets, and the SwarnSeq method was also applied to data from Tung et al., where ERCC spike-ins are available. We used the processed UMI count data for these scRNA-seq studies as these datasets have gone through careful quality control steps by the authors of the original publications.

Benchmarking based on Receiver Operating Characteristic

This comparison setting used the experimental designs and the count datasets for performance analysis of DE methods. For instance, Mouse cell data (GSE29087) [196] was used to detect DE genes between 48 mouse embryonic stem cells and 44 mouse embryonic fibroblast cells. Then, the 12 competitive methods, including SwarnSeq, were compared in terms of their AUC using the identified reference gene lists. Basically, through each of the method, DE gene sets of size 3000 are selected for each of the datasets. Then, the AUC values were computed by executing *proc* function implemented in pROC R package [283] using the output (*i.e. p-values* or adjusted *p-values*) of each method as predictor, and a binary vector, indicating whether a gene belongs to the reference gene list, as response. The ROC curves of different methods are shown in Figures 6.7, 6.8 along with corresponding AUC values.

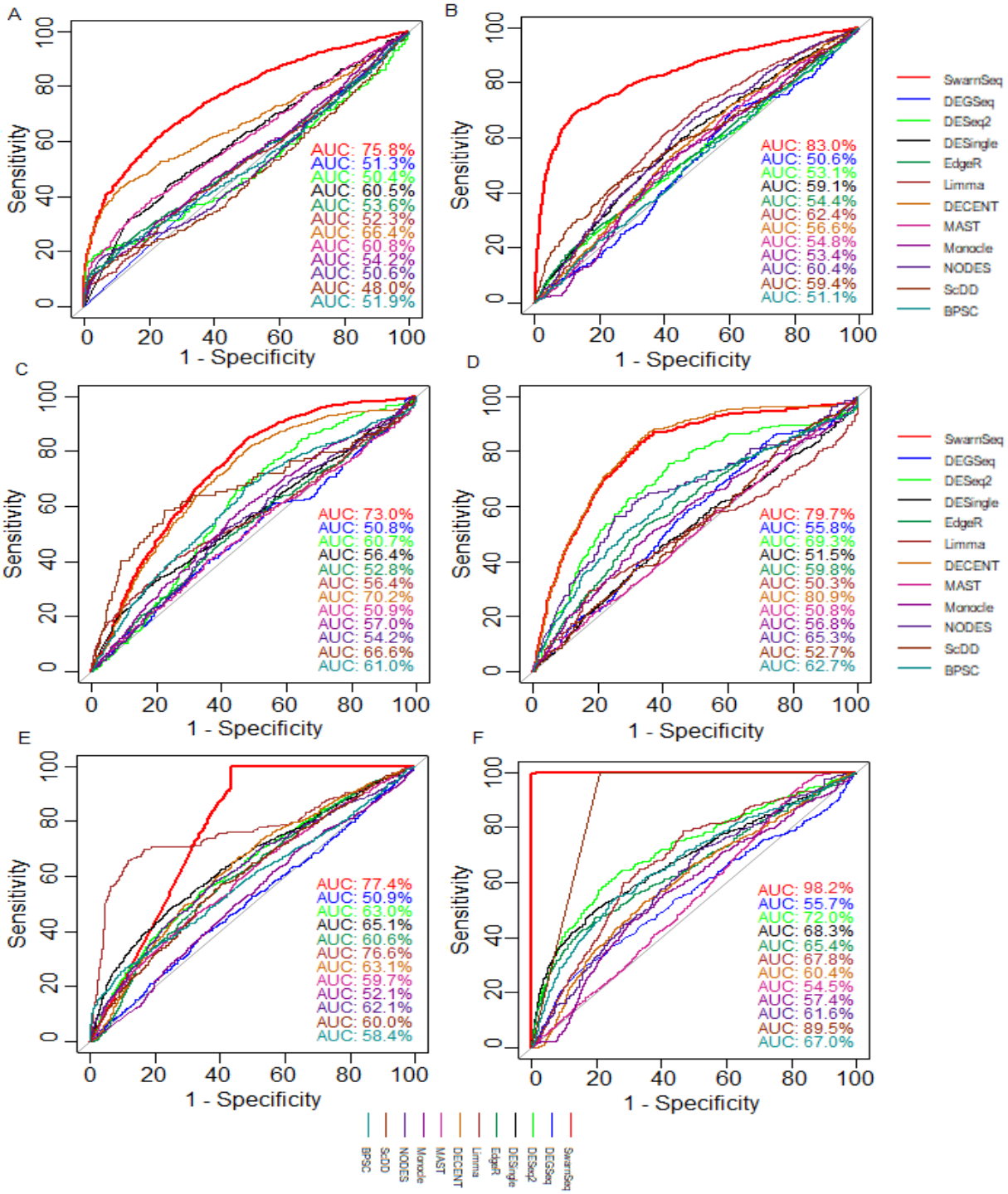


Figure 6.7. Differential expression analysis of real scRNA-seq data (Part I). Receiver Operating Characteristic curves for differential expression methods on different real scRNA-seq data. Evaluation of the performance of different methods based on AUROC is shown for (A) GSE53638 (Data 1); (B) GSE77728; (C) GSE53638 (Data 3); (D) GSE53638 (Data 2); (E) GSE29087; (F) GSE65525. Different goldstandard gene lists are prepared based on the FC values for benchmarking various differential expression analysis methods on different real scRNA-seq datasets.

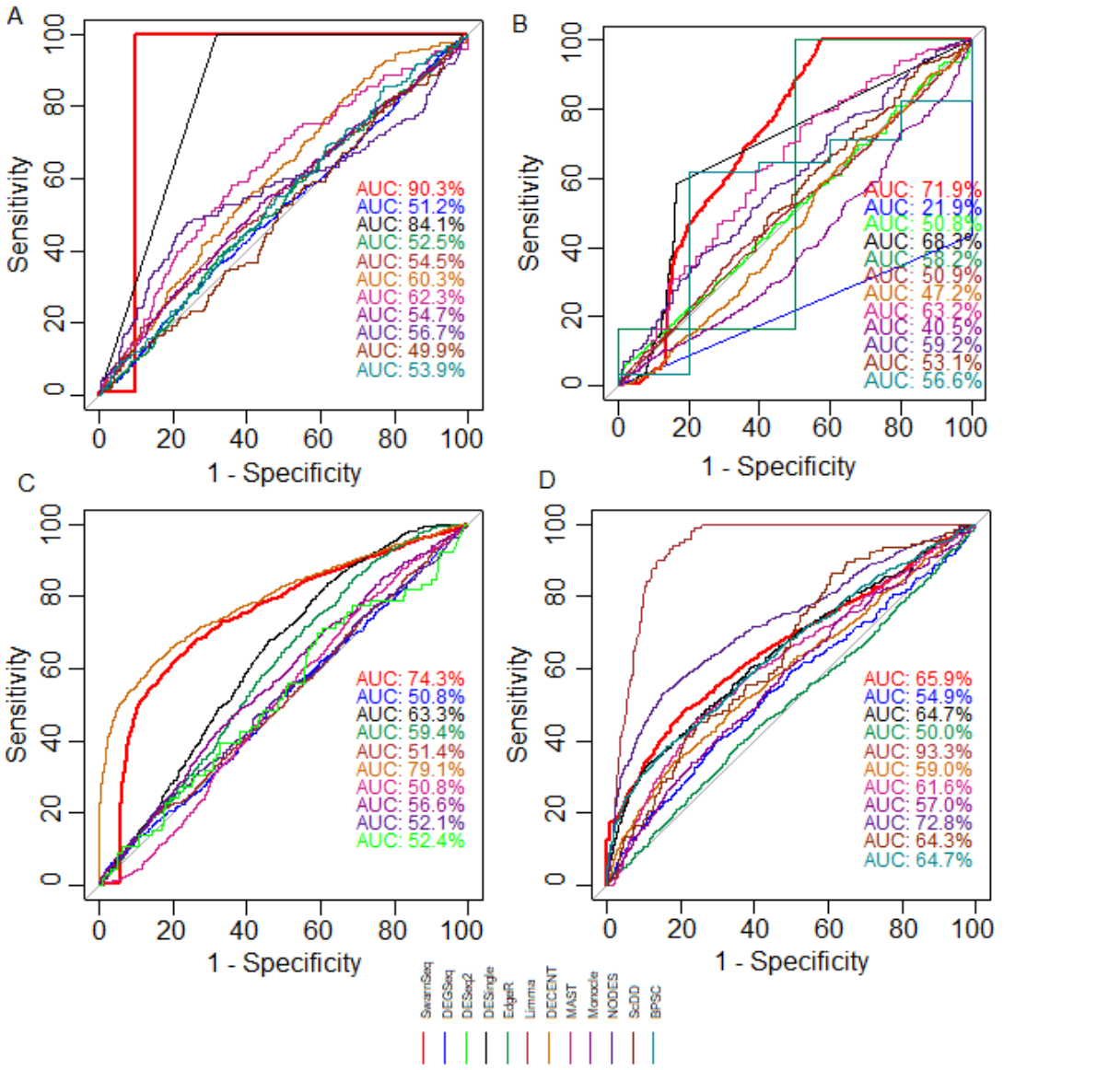


Figure 6.8. Differential expression (DE) analysis of real scRNA-seq data (Part II). Receiver operating characteristic curves for differential expression methods on different real scRNA-seq data. Evaluation of the performance of different methods based on Area Under Receiver Operating Characteristic Curves (AUC) is shown for (A) GSE111108; (B) GSE92495; (C) GSE115469; (D) GSE109999. Different goldstandard gene lists are prepared based on the fold change values for benchmarking different differential expression analysis methods on different real scRNA-seq datasets. Swarnseq achieves competitive and better accuarcy for identifying genuine differential gene lists in all four different real datasets. DE methods are denoted by different colors.

In this comparison setting for GSE53638 (data 1), the SwarnSeq (0.76) produced highest AUC values followed by DECENT (0.66), MAST (0.61), DESingle (0.61), Monocle (0.54), and BPSC (0.52) among single cell

specific tools (Figure 6.7A). The scDD performed the worst among the scRNA-seq DE tools for this data. Further, edgeR (0.54) had higher AUC values followed by Limma (0.52), DEGseq (0.51) and DESeq2 (0.48) in the bulk RNA-seq tool category (Figure 6.7A). Importantly, it was found that the SwarnSeq performed better than other methods of both bulk and scRNA-seq DE tools. For GSE53638 (data 3) data, SwarnSeq (0.73) had the highest AUC values followed by DECENT (0.70) and performed best among bulk and scRNA-seq DE tools (Figure 6.7C). Moreover, among bulk RNA-seq DE tools, edgeR (0.54) had higher AUC followed by Limma (0.52), DEGseq (0.51) for the GSE53638 (3) data (Figure 6.7C). Similarly, for GSE29087 data, the AUC for SwarnSeq method was highest (0.83) among other competitive bulk and single cell RNA-seq DE tools (Figure 6.7B). Among the bulk RNA-seq DE tools, Limma had higher AUC (0.62), when applied to GSE29087 scRNA-seq data. Similar interpretations can be made for other datasets, as shown in Figures 6.7 and 6.8. Our analysis indicated that under AUROC settings, our SwarnSeq method performed better in 8 datasets (with rank 1) and competitive with other methods in remaining 2 datasets (rank 2 and 3) (Figures 6.7 and 6.8). In other words, the performance of SwarnSeq method is consistently better than other competitive methods on real scRNA-seq datasets.

Benchmarking based on FDR

The second comparison setting included assessment of the 12 methods through computation of FDRs for different DE gene sets on the 10 different real scRNA-seq datasets. For this purpose, different DE gene sets of sizes 100, 200, 300, …, 3000 were selected based on the *p-values*/adjusted *p-values* computed through

each of the 12 methods. Then, the selected DE genes were compared with respect to the reference gene list to compute FDRs for each of the 10 datasets. The results are shown in Figures 6.9 and 6.10.

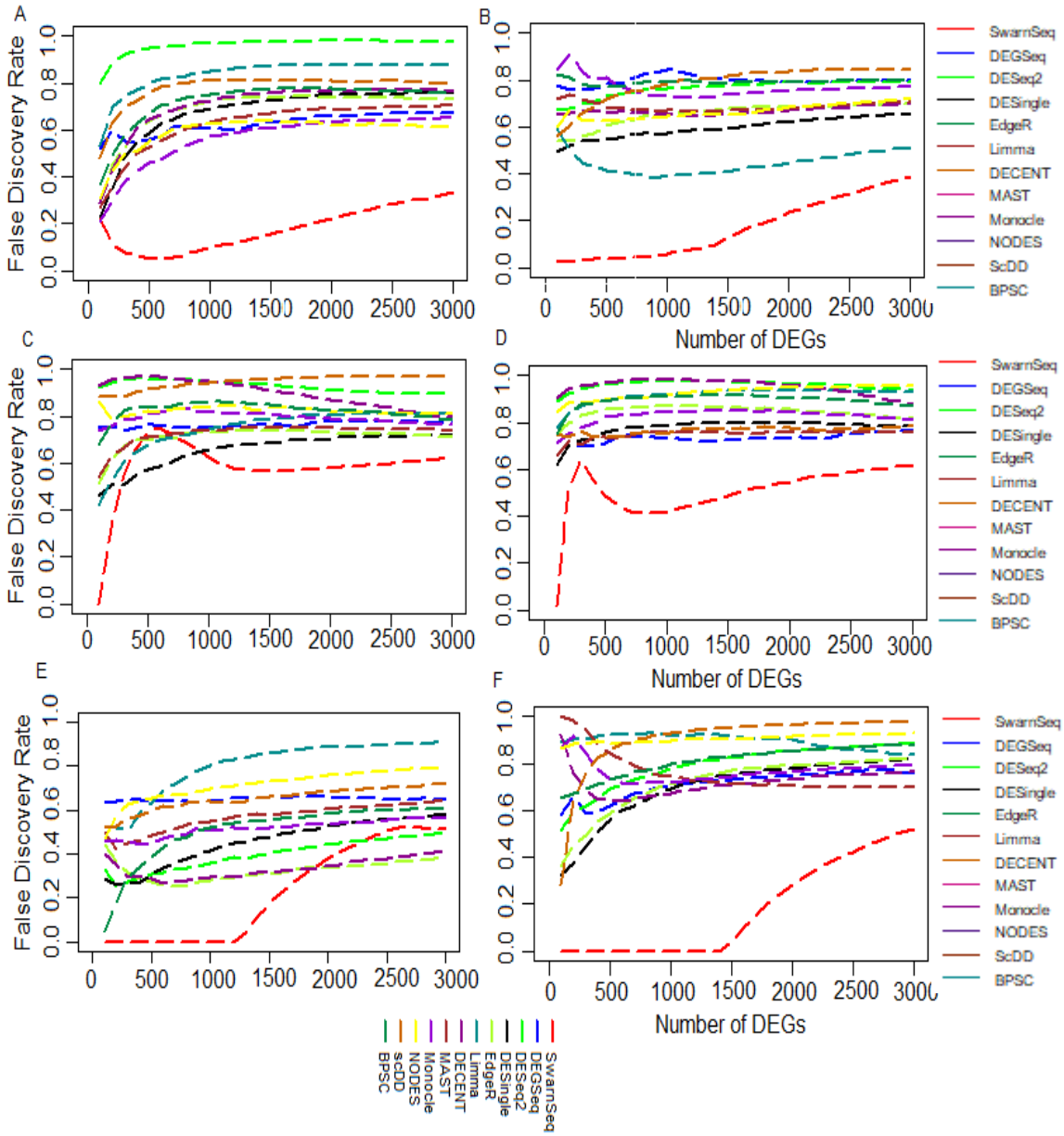


Figure 6.9. FDR based Performance analysis of DE methods on real scRNA-seq data. FDR curves for differential expression methods on different real scRNA-seq data are shown. Evaluation of the performance of different methods based on false discovery rate is shown for (A) GSE53638 (Data 1); (B) GSE77728; (C) GSE53638 (Data 3); (D) GSE53638 (Data 2); (E) GSE29087; (F) GSE65525. Different reference gene lists are prepared based on the FC values for benchmarking different differential expression analysis methods on different real scRNA-seq

datasets. SwarnSeq achieves competitive and better accuarcy for identifying genuine differential gene lists in all four different real datasets. DE methods are denoted by different colors.

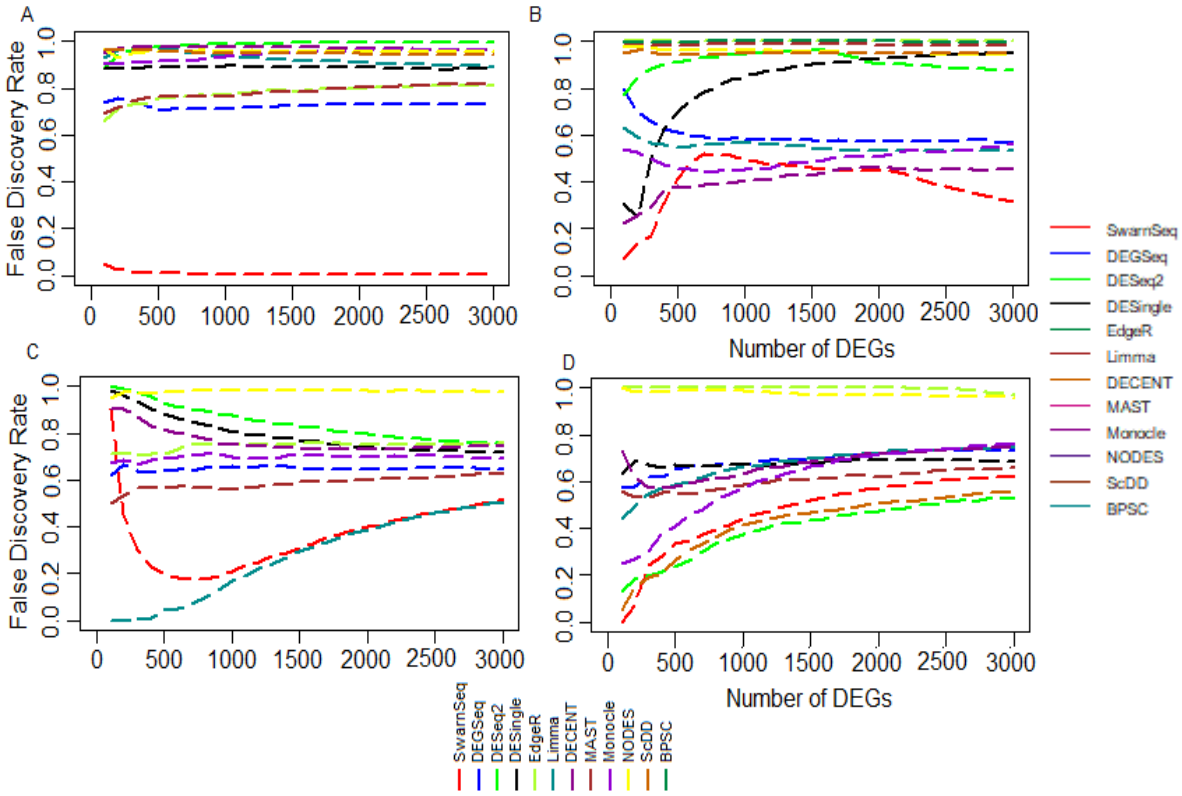


Figure 6.10. FDR based Performance analysis of DE methods on real scRNA-seq data (part II). FDR curves for differential expression methods on different real scRNA-seq data are shown. Evaluation of the performance of different methods based on false discovery rate is shown for (A) GSE111108; (B) GSE92495; (C) GSE115469; (D) GSE109999.

In this comparison setting, it was found that the FDR computed for the SwarnSeq method was found to be lower as compared to other competitive methods for GSE53638 (data 1) (Figure 6.9A). Similar findings were observed across all the selected DE gene sets for the same data (Figure 6.9). This indicates that the proposed SwarnSeq performed better to detect DE genes as compared to other competitive methods. Also, its performance was found to be robust compared to methods across all DE gene sets. Similar interpretations can be made for other remaining datasets (Figures 6.9, 6.10). Under this FDR based comparison setting on multiple real scRNA-seq datasets, we demonstrated our

SwarnSeq method was consistently better and more robust to detect the DE genes of various sizes with respect to other bulk and scRNA-seq DE tools.

Benchmarking based on other performance metrics

This comparison setting included the performance evaluation of the 12 scRNA-seq DE tools based on performance metrics, *viz.* TP, TN, FN, FP, FPR, NPV, F1, and ACC on the10 scRNA-seq datasets. For this purpose, the DE methods were applied to each dataset following the instructions and recommendations of their respective software packages. Genes were declared as DE based on their computed *p-values*/adjusted *p-values* and subsequently DE gene sets of sizes 500, 1000, 1500, …, 3000 were selected for each of the datasets. Then, the performance metrics were computed for the DEGs from different datasets and the results are given in Tables 6.4, 6.5 – 6.10.

In this comparison setting, for a DE gene set of size 500, the SwarnSeq method identified more TP genes, followed by DECENT as compared to other competitive methods in GSE29087 data (Table 6.4). Further, the value of FP, FN, and FPR for the SwarnSeq was observed to be lower than other competitive methods. Moreover, the values of TPR, NPV, ACC, and F1 for SwarnSeq method were found to be higher than from other methods (Table 6.4).

Table 6.4. Performance evaluation metrics for GSE29087 scRNA-seq data.

	NDEG = 500									
Methods	**TP**	**FP**	**TN**	**FN**	**TPR**	**FPR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	500	0	8436	2500	0.167	0.000	1.000	0.771	0.781	0.286
DEGSeq	181	319	8140	2819	0.060	0.038	0.362	0.743	0.726	0.103
DESeq2	346	154	8282	2654	0.115	0.018	0.692	0.757	0.754	0.198
DESingle	344	156	8280	2656	0.115	0.018	0.688	0.757	0.754	0.197
EdgeR	364	136	8302	2636	0.121	0.016	0.728	0.759	0.758	0.208
Limma	182	318	8188	2818	0.061	0.037	0.364	0.744	0.727	0.104

DECENT	355	145	8291	2645	0.118	0.017	0.710	0.758	0.756	0.203
MAST	258	242	8195	2742	0.086	0.029	0.516	0.749	0.739	0.147
Monocle	275	225	8212	2725	0.092	0.027	0.550	0.751	0.742	0.157
NODES	174	326	8110	2826	0.058	0.039	0.348	0.742	0.724	0.099
scDD	202	298	8138	2798	0.067	0.035	0.404	0.744	0.729	0.115
BPSC	308	192	8244	2692	0.103	0.023	0.616	0.754	0.748	0.176
NDEG = 1000										
SwarnSeq	1000	0	8436	2000	0.333	0.000	1.000	0.808	0.825	0.500
DEGSeq	357	643	7846	2643	0.119	0.076	0.357	0.748	0.714	0.179
DESeq2	641	359	8077	2359	0.214	0.043	0.641	0.774	0.762	0.321
DESingle	585	415	8021	2415	0.195	0.049	0.585	0.769	0.753	0.293
EdgeR	718	282	8164	2282	0.239	0.033	0.718	0.782	0.776	0.359
Limma	198	802	8126	2802	0.066	0.090	0.198	0.744	0.698	0.099
DECENT	706	294	8142	2294	0.235	0.035	0.706	0.780	0.774	0.353
MAST	449	551	7894	2551	0.150	0.065	0.449	0.756	0.729	0.225
Monocle	495	505	7934	2505	0.165	0.060	0.495	0.760	0.737	0.248
NODES	301	699	7737	2699	0.100	0.083	0.301	0.741	0.703	0.151
scDD	362	638	7798	2638	0.121	0.076	0.362	0.747	0.714	0.181
BPSC	481	519	7917	2519	0.160	0.062	0.481	0.759	0.734	0.241
NDEG = 1500										
SwarnSeq	1242	258	8178	1758	0.414	0.031	0.828	0.823	0.824	0.552
DEGSeq	510	990	7539	2490	0.170	0.116	0.340	0.752	0.698	0.227
DESeq2	886	614	7822	2114	0.295	0.073	0.591	0.787	0.761	0.394
DESingle	782	718	7718	2218	0.261	0.085	0.521	0.777	0.743	0.348
EdgeR	1037	463	7997	1963	0.346	0.055	0.691	0.803	0.788	0.461
Limma	212	1288	8052	2788	0.071	0.138	0.141	0.743	0.670	0.094
DECENT	1025	475	7961	1975	0.342	0.056	0.683	0.801	0.786	0.456
MAST	630	870	7589	2370	0.210	0.103	0.420	0.762	0.717	0.280
Monocle	720	780	7663	2280	0.240	0.092	0.480	0.771	0.733	0.320
NODES	403	1097	7339	2597	0.134	0.130	0.269	0.739	0.677	0.179
scDD	513	987	7449	2487	0.171	0.117	0.342	0.750	0.696	0.228
BPSC	671	829	7607	2329	0.224	0.098	0.447	0.766	0.724	0.298
NDEG = 2000										
SwarnSeq	1320	680	7803	1680	0.440	0.080	0.660	0.823	0.794	0.528
DEGSeq	682	1318	7238	2318	0.227	0.154	0.341	0.757	0.685	0.273
DESeq2	1117	883	7553	1883	0.372	0.105	0.559	0.800	0.758	0.447
DESingle	946	1054	7382	2054	0.315	0.125	0.473	0.782	0.728	0.378
EdgeR	1242	758	7678	1758	0.414	0.090	0.621	0.814	0.780	0.497
Limma	228	1772	7978	2772	0.076	0.182	0.114	0.742	0.644	0.091
DECENT	1314	686	7757	1686	0.438	0.081	0.657	0.821	0.793	0.526
MAST	795	1205	7289	2205	0.265	0.142	0.398	0.768	0.703	0.318
Monocle	925	1075	7376	2075	0.308	0.127	0.463	0.780	0.725	0.370
NODES	485	1515	6925	2515	0.162	0.180	0.243	0.734	0.648	0.194
scDD	633	1367	7069	2367	0.211	0.162	0.317	0.749	0.673	0.253

BPSC	832	1168	7268	2168	0.277	0.138	0.416	0.770	0.708	0.333
NDEG = 2500										
SwarnSeq	1601	899	7612	1399	0.534	0.106	0.640	0.845	0.800	0.582
DEGSeq	874	1626	6966	2126	0.291	0.189	0.350	0.766	0.676	0.318
DESeq2	1327	1173	7263	1673	0.442	0.139	0.531	0.813	0.751	0.483
DESingle	1103	1397	7039	1897	0.368	0.166	0.441	0.788	0.712	0.401
EdgeR	1242	1258	7178	1758	0.414	0.149	0.497	0.803	0.736	0.452
Limma	255	2245	7915	2745	0.085	0.221	0.102	0.742	0.621	0.093
DECENT	1548	952	7499	1452	0.516	0.113	0.619	0.838	0.790	0.563
MAST	945	1555	6973	2055	0.315	0.182	0.378	0.772	0.687	0.344
Monocle	1114	1386	7071	1886	0.371	0.164	0.446	0.789	0.714	0.405
NODES	556	1944	6517	2444	0.185	0.230	0.222	0.727	0.617	0.202
scDD	744	1756	6680	2256	0.248	0.208	0.298	0.748	0.649	0.271
BPSC	999	1501	6935	2001	0.333	0.178	0.400	0.776	0.694	0.363
NDEG = 3000										
SwarnSeq	1837	1163	7386	1163	0.612	0.136	0.612	0.864	0.799	0.612
DEGSeq	1055	1945	6681	1945	0.352	0.225	0.352	0.775	0.665	0.352
DESeq2	1502	1498	6938	1498	0.501	0.178	0.501	0.822	0.738	0.501
DESingle	1249	1751	6685	1751	0.416	0.208	0.416	0.792	0.694	0.416
EdgeR	1450	1550	6886	1550	0.483	0.184	0.483	0.816	0.729	0.483
Limma	279	2721	7813	2721	0.093	0.258	0.093	0.742	0.598	0.093
DECENT	1754	1246	7217	1246	0.585	0.147	0.585	0.853	0.783	0.585
MAST	1074	1926	6651	1926	0.358	0.225	0.358	0.775	0.667	0.358
Monocle	1303	1697	6769	1697	0.434	0.200	0.434	0.800	0.704	0.434
NODES	633	2367	6106	2367	0.211	0.279	0.211	0.721	0.587	0.211
scDD	845	2155	6281	2155	0.282	0.255	0.282	0.745	0.623	0.282
BPSC	1181	1819	6617	1819	0.394	0.216	0.394	0.784	0.682	0.394

TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

This finding indicates the better performance of our proposed method in terms of various computed metrics for the GSE29087 dataset. Further, we demonstrated consistently similar findings for our method over other DE gene sets of sizes 500, 1000, 1500, …, 3000 (Table 6.4). Similar interpretations can be made for other datasets, as shown in Table 6.5 – 6.10. The comparative analysis under this setting gave us confidence that our SwarnSeq method can detect the genes, which are truly DE in wide range of real datasets. Furthermore, its performance

was consistently better over the considered competitive scRNA-seq DE methods, when assessed through various performance metrics.

Table 6.5. Performance evaluation metrics for GSE92495 scRNA-seq data.

						NDEG = 500				
Methods	**TP**	**FP**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	292	208	2708	0.097	0.017	0.416	0.584	0.820	0.812	0.167
DEGSeq	193	307	2807	0.064	0.025	0.614	0.386	0.813	0.799	0.110
DESeq2	45	455	2955	0.015	0.036	0.910	0.090	0.803	0.780	0.026
DESingle	150	350	2850	0.050	0.028	0.700	0.300	0.810	0.794	0.086
EdgeR	1	499	2999	0.000	0.040	0.998	0.002	0.800	0.775	0.001
Limma	226	274	2774	0.075	0.022	0.548	0.452	0.818	0.806	0.129
DECENT	310	190	2690	0.103	0.015	0.380	0.620	0.821	0.815	0.177
MAST	9	491	2991	0.003	0.039	0.982	0.018	0.801	0.776	0.005
Monocle	273	227	2727	0.091	0.018	0.454	0.546	0.819	0.810	0.156
NODES	18	482	2982	0.006	0.038	0.964	0.036	0.802	0.777	0.010
ScDD	29	471	2971	0.010	0.038	0.942	0.058	0.802	0.778	0.017
BPSC	1	499	2999	0.000	0.040	0.998	0.002	0.800	0.775	0.001
						NDEG=1000				
SwarnSeq	506	494	2494	0.169	0.039	0.494	0.506	0.828	0.808	0.253
DEGSeq	423	577	2577	0.141	0.046	0.577	0.423	0.823	0.797	0.212
DESeq2	55	945	2945	0.018	0.075	0.945	0.055	0.797	0.749	0.028
DESingle	150	850	2850	0.050	0.068	0.850	0.150	0.804	0.762	0.075
EdgeR	1	999	2999	0.000	0.080	0.999	0.001	0.794	0.742	0.001
Limma	435	565	2565	0.145	0.044	0.565	0.435	0.827	0.802	0.218
DECENT	595	405	2405	0.198	0.032	0.405	0.595	0.835	0.819	0.298
MAST	12	988	2988	0.004	0.079	0.988	0.012	0.794	0.744	0.006
Monocle	550	450	2450	0.183	0.036	0.450	0.550	0.831	0.813	0.275
NODES	37	963	2963	0.012	0.077	0.963	0.037	0.796	0.747	0.019
ScDD	52	948	2948	0.017	0.076	0.948	0.052	0.797	0.749	0.026
BPSC	1	999	2999	0.000	0.080	0.999	0.001	0.794	0.742	0.001
						NDEG=1500				
SwarnSeq	813	687	2187	0.271	0.055	0.458	0.542	0.844	0.815	0.361
DEGSeq	631	869	2369	0.210	0.069	0.579	0.421	0.831	0.792	0.280
DESeq2	61	1439	2939	0.020	0.115	0.959	0.041	0.790	0.718	0.027
DESingle	150	1350	2850	0.050	0.108	0.900	0.100	0.797	0.729	0.067
EdgeR	1	1499	2999	0.000	0.120	0.999	0.001	0.786	0.710	0.000
Limma	678	822	2322	0.226	0.064	0.548	0.452	0.839	0.803	0.301
DECENT	862	638	2138	0.287	0.051	0.425	0.575	0.848	0.821	0.383
MAST	21	1479	2979	0.007	0.118	0.986	0.014	0.788	0.713	0.009
Monocle	775	725	2225	0.258	0.058	0.483	0.517	0.842	0.810	0.344
NODES	68	1432	2932	0.023	0.114	0.955	0.045	0.791	0.719	0.030

ScDD	74	1426	2926	0.025	0.114	0.951	0.049	0.791	0.720	0.033
BPSC	1	1499	2999	0.000	0.120	0.999	0.001	0.786	0.710	0.000
						NDEG = 2000				
SwarnSeq	1113	887	1887	0.371	0.071	0.444	0.557	0.860	0.821	0.445
DEGSeq	846	1154	2154	0.282	0.092	0.577	0.423	0.841	0.787	0.338
DESeq2	189	1811	2811	0.063	0.144	0.906	0.095	0.793	0.703	0.076
DESingle	150	1850	2850	0.050	0.148	0.925	0.075	0.789	0.697	0.060
EdgeR	1	1999	2999	0.000	0.160	1.000	0.001	0.778	0.678	0.000
Limma	936	1064	2064	0.312	0.082	0.532	0.468	0.852	0.804	0.374
DECENT	1085	915	1915	0.362	0.073	0.458	0.543	0.859	0.819	0.434
MAST	30	1970	2970	0.010	0.157	0.985	0.015	0.780	0.682	0.012
Monocle	988	1012	2012	0.329	0.081	0.506	0.494	0.852	0.806	0.395
NODES	102	1898	2898	0.034	0.152	0.949	0.051	0.786	0.691	0.041
ScDD	99	1901	2901	0.033	0.152	0.951	0.050	0.785	0.691	0.040
BPSC	3	1997	2997	0.001	0.159	0.999	0.002	0.778	0.678	0.001
						NDEG = 2500				
SwarnSeq	1550	950	1450	0.517	0.076	0.380	0.620	0.889	0.845	0.564
DEGSeq	1066	1434	1934	0.355	0.114	0.574	0.426	0.852	0.783	0.388
DESeq2	275	2225	2725	0.092	0.177	0.890	0.110	0.792	0.682	0.100
DESingle	150	2350	2850	0.050	0.188	0.940	0.060	0.781	0.665	0.055
EdgeR	1	2499	2999	0.000	0.200	1.000	0.000	0.770	0.646	0.000
Limma	1174	1326	1826	0.391	0.102	0.530	0.470	0.865	0.803	0.427
DECENT	1367	1133	1633	0.456	0.090	0.453	0.547	0.875	0.823	0.497
MAST	43	2457	2957	0.014	0.196	0.983	0.017	0.773	0.651	0.016
Monocle	1171	1329	1829	0.390	0.106	0.532	0.468	0.860	0.797	0.426
NODES	133	2367	2867	0.044	0.189	0.947	0.053	0.780	0.663	0.048
ScDD	129	2371	2871	0.043	0.189	0.948	0.052	0.780	0.662	0.047
BPSC	5	2495	2995	0.002	0.199	0.998	0.002	0.770	0.646	0.002
						NDEG = 3000				
SwarnSeq	2050	950	950	0.683	0.076	0.317	0.683	0.924	0.878	0.683
DEGSeq	1277	1723	1723	0.426	0.137	0.574	0.426	0.863	0.778	0.426
DESeq2	371	2629	2629	0.124	0.209	0.876	0.124	0.791	0.663	0.124
DESingle	150	2850	2850	0.050	0.228	0.950	0.050	0.772	0.633	0.050
EdgeR	2	2998	2998	0.001	0.239	0.999	0.001	0.761	0.614	0.001
Limma	1395	1605	1605	0.465	0.123	0.535	0.465	0.877	0.799	0.465
DECENT	1641	1359	1359	0.547	0.108	0.453	0.547	0.892	0.826	0.547
MAST	64	2936	2936	0.021	0.234	0.979	0.021	0.766	0.622	0.021
Monocle	1328	1672	1672	0.443	0.133	0.557	0.443	0.867	0.786	0.443
NODES	178	2822	2822	0.059	0.225	0.941	0.059	0.775	0.636	0.059
ScDD	162	2838	2838	0.054	0.227	0.946	0.054	0.773	0.634	0.054
BPSC	5	2995	2995	0.002	0.239	0.998	0.002	0.761	0.614	0.002

Table 6.6. Performance evaluation metrics for GSE53638 (Data 1).

Methods	TP	FP	FN	TPR	FPR	FDR	PPR	NPV	ACC	F1
NDEG = 500										
SwarnSeq	475	25	3525	0.119	0.002	0.050	0.950	0.769	0.775	0.211
DEGSeq	220	280	3780	0.055	0.024	0.560	0.440	0.752	0.742	0.098
DESeq2	26	474	3974	0.007	0.040	0.948	0.052	0.739	0.717	0.012
DESingle	207	293	3793	0.052	0.025	0.586	0.414	0.751	0.740	0.092
EdgeR	181	319	3819	0.045	0.027	0.638	0.362	0.749	0.737	0.080
Limma	102	398	3898	0.026	0.034	0.796	0.204	0.744	0.727	0.045
DECENT	177	323	3823	0.044	0.028	0.646	0.354	0.749	0.737	0.079
MAST	236	264	3764	0.059	0.022	0.528	0.472	0.753	0.744	0.105
Monocle	271	229	3729	0.068	0.020	0.458	0.542	0.755	0.749	0.120
NODES	225	275	3775	0.056	0.023	0.550	0.450	0.752	0.743	0.100
scDD	122	378	3878	0.031	0.032	0.756	0.244	0.746	0.730	0.054
BPSC	164	336	3836	0.041	0.029	0.672	0.328	0.748	0.735	0.073
NDEG = 1000										
SwarnSeq	907	93	3093	0.227	0.008	0.093	0.907	0.790	0.798	0.363
DEGSeq	396	604	3604	0.099	0.051	0.604	0.396	0.755	0.733	0.158
DESeq2	33	967	3967	0.008	0.082	0.967	0.033	0.731	0.687	0.013
DESingle	307	693	3693	0.077	0.059	0.693	0.307	0.749	0.721	0.123
EdgeR	275	725	3725	0.069	0.062	0.725	0.275	0.747	0.717	0.110
Limma	150	850	3850	0.038	0.072	0.850	0.150	0.739	0.701	0.060
DECENT	274	726	3726	0.069	0.062	0.726	0.274	0.747	0.717	0.110
MAST	366	634	3634	0.092	0.054	0.634	0.366	0.753	0.729	0.146
Monocle	429	571	3571	0.107	0.049	0.571	0.429	0.758	0.737	0.172
NODES	373	627	3627	0.093	0.053	0.627	0.373	0.754	0.730	0.149
scDD	189	811	3811	0.047	0.069	0.811	0.189	0.741	0.706	0.076
BPSC	249	751	3751	0.062	0.064	0.751	0.249	0.746	0.714	0.100
NDEG = 1500										
SwarnSeq	1264	236	2736	0.316	0.020	0.157	0.843	0.808	0.811	0.460
DEGSeq	556	944	3444	0.139	0.080	0.629	0.371	0.758	0.721	0.202
DESeq2	37	1463	3963	0.009	0.125	0.975	0.025	0.722	0.655	0.013
DESingle	410	1090	3590	0.103	0.093	0.727	0.273	0.748	0.703	0.149
EdgeR	390	1110	3610	0.098	0.095	0.740	0.260	0.747	0.700	0.142
Limma	195	1305	3805	0.049	0.111	0.870	0.130	0.733	0.675	0.071
DECENT	364	1136	3636	0.091	0.097	0.757	0.243	0.745	0.697	0.132
MAST	493	1007	3507	0.123	0.086	0.671	0.329	0.754	0.713	0.179
Monocle	593	907	3407	0.148	0.077	0.605	0.395	0.761	0.726	0.216
NODES	540	960	3460	0.135	0.082	0.640	0.360	0.757	0.719	0.196
scDD	287	1213	3713	0.072	0.103	0.809	0.191	0.739	0.687	0.104
BPSC	341	1159	3659	0.085	0.099	0.773	0.227	0.743	0.694	0.124
NDEG = 2000										
SwarnSeq	1554	446	2446	0.389	0.038	0.223	0.777	0.822	0.816	0.518

DEGSeq	687	1313	3313	0.172	0.112	0.657	0.344	0.759	0.706	0.229
DESeq2	41	1959	3959	0.010	0.167	0.980	0.021	0.712	0.624	0.014
DESingle	497	1503	3503	0.124	0.128	0.752	0.249	0.745	0.682	0.166
EdgeR	522	1478	3478	0.131	0.126	0.739	0.261	0.747	0.685	0.174
Limma	243	1757	3757	0.061	0.150	0.879	0.122	0.727	0.650	0.081
DECENT	465	1535	3535	0.116	0.131	0.768	0.233	0.743	0.678	0.155
MAST	622	1378	3378	0.156	0.117	0.689	0.311	0.754	0.698	0.207
Monocle	737	1263	3263	0.184	0.108	0.632	0.369	0.763	0.713	0.246
NODES	757	1243	3243	0.189	0.106	0.622	0.379	0.764	0.715	0.252
scDD	385	1615	3615	0.096	0.138	0.808	0.193	0.737	0.668	0.128
BPSC	452	1548	3548	0.113	0.132	0.774	0.226	0.742	0.676	0.151
NDEG = 2500										
SwarnSeq	1792	708	2208	0.448	0.060	0.283	0.717	0.833	0.815	0.551
DEGSeq	826	1674	3174	0.207	0.143	0.670	0.330	0.760	0.692	0.254
DESeq2	54	2446	3946	0.014	0.208	0.978	0.022	0.702	0.594	0.017
DESingle	618	1882	3382	0.155	0.160	0.753	0.247	0.745	0.666	0.190
EdgeR	655	1845	3345	0.164	0.157	0.738	0.262	0.748	0.670	0.202
Limma	308	2192	3692	0.077	0.187	0.877	0.123	0.721	0.626	0.095
DECENT	573	1927	3427	0.143	0.164	0.771	0.229	0.741	0.660	0.176
MAST	761	1739	3239	0.190	0.148	0.696	0.304	0.756	0.684	0.234
Monocle	899	1601	3101	0.225	0.136	0.640	0.360	0.766	0.701	0.277
NODES	948	1552	3052	0.237	0.132	0.621	0.379	0.770	0.708	0.292
scDD	476	2024	3524	0.119	0.172	0.810	0.190	0.734	0.648	0.146
BPSC	589	1911	3411	0.147	0.163	0.764	0.236	0.742	0.662	0.181
NDEG = 3000										
SwarnSeq	2003	997	1997	0.501	0.085	0.332	0.668	0.843	0.810	0.572
DEGSeq	988	2012	3012	0.247	0.171	0.671	0.329	0.764	0.681	0.282
DESeq2	79	2921	3921	0.020	0.249	0.974	0.026	0.692	0.565	0.023
DESingle	736	2264	3264	0.184	0.193	0.755	0.245	0.744	0.649	0.210
EdgeR	813	2187	3187	0.203	0.186	0.729	0.271	0.750	0.659	0.232
Limma	372	2628	3628	0.093	0.224	0.876	0.124	0.715	0.603	0.106
DECENT	701	2299	3299	0.175	0.196	0.766	0.234	0.741	0.645	0.200
MAST	890	2110	3110	0.223	0.179	0.703	0.297	0.756	0.669	0.254
Monocle	1039	1961	2961	0.260	0.167	0.654	0.346	0.768	0.687	0.297
NODES	1164	1836	2836	0.291	0.156	0.612	0.388	0.777	0.703	0.333
scDD	608	2392	3392	0.152	0.204	0.797	0.203	0.734	0.633	0.174
BPSC	731	2269	3269	0.183	0.193	0.756	0.244	0.743	0.648	0.209

Table 6.7. Performance evaluation metrics for GSE53638 (Data 2) scRNA-seq data.

NDEG = 500										
Methods	**TP**	**FP**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	258	242	2742	0.086	0.019	0.484	0.516	0.819	0.809	0.147

DEGSeq	145	355	2855	0.048	0.028	0.710	0.290	0.811	0.795	0.083
DESeq2	20	480	2980	0.007	0.038	0.960	0.040	0.803	0.779	0.011
DESingle	122	378	2878	0.041	0.030	0.756	0.244	0.810	0.792	0.070
EdgeR	72	428	2928	0.024	0.034	0.856	0.144	0.807	0.785	0.041
Limma	54	446	2946	0.018	0.035	0.892	0.108	0.805	0.783	0.031
DECENT	14	486	2986	0.005	0.038	0.972	0.028	0.803	0.778	0.008
MAST	130	370	2870	0.043	0.029	0.740	0.260	0.810	0.793	0.074
Monocle	89	411	2911	0.030	0.033	0.822	0.178	0.808	0.788	0.051
NODES	51	449	2949	0.017	0.036	0.898	0.102	0.805	0.783	0.029
ScDD	130	370	2870	0.043	0.029	0.740	0.260	0.810	0.793	0.074
BPSC	55	445	2945	0.018	0.035	0.890	0.110	0.805	0.783	0.031
NDEG = 1000										
SwarnSeq	585	415	2415	0.195	0.033	0.415	0.585	0.835	0.819	0.293
DEGSeq	272	728	2728	0.091	0.058	0.728	0.272	0.814	0.779	0.136
DESeq2	26	974	2974	0.009	0.077	0.974	0.026	0.797	0.747	0.013
DESingle	214	786	2786	0.071	0.062	0.786	0.214	0.810	0.772	0.107
EdgeR	129	871	2871	0.043	0.069	0.871	0.129	0.804	0.761	0.065
Limma	81	919	2919	0.027	0.073	0.919	0.081	0.801	0.755	0.041
DECENT	17	982	2983	0.006	0.078	0.983	0.017	0.796	0.746	0.009
MAST	242	758	2758	0.081	0.060	0.758	0.242	0.812	0.775	0.121
Monocle	155	845	2845	0.052	0.067	0.845	0.155	0.806	0.764	0.078
NODES	84	916	2916	0.028	0.072	0.916	0.084	0.801	0.755	0.042
ScDD	248	752	2752	0.083	0.060	0.752	0.248	0.812	0.776	0.124
BPSC	89	911	2911	0.030	0.072	0.911	0.089	0.801	0.756	0.045
NDEG = 1500										
SwarnSeq	767	733	2233	0.256	0.058	0.489	0.511	0.842	0.810	0.341
DEGSeq	417	1083	2583	0.139	0.086	0.722	0.278	0.817	0.766	0.185
DESeq2	40	1460	2960	0.013	0.116	0.973	0.027	0.791	0.717	0.018
DESingle	303	1197	2697	0.101	0.095	0.798	0.202	0.809	0.751	0.135
EdgeR	208	1292	2792	0.069	0.102	0.861	0.139	0.802	0.739	0.092
Limma	100	1400	2900	0.033	0.111	0.933	0.067	0.795	0.725	0.044
DECENT	35	1464	2965	0.012	0.116	0.977	0.023	0.790	0.717	0.016
MAST	363	1137	2637	0.121	0.090	0.758	0.242	0.813	0.759	0.161
Monocle	224	1276	2776	0.075	0.101	0.851	0.149	0.804	0.741	0.100
NODES	94	1406	2906	0.031	0.111	0.937	0.063	0.794	0.724	0.042
ScDD	342	1158	2658	0.114	0.092	0.772	0.228	0.812	0.756	0.152
BPSC	128	1372	2872	0.043	0.109	0.915	0.085	0.797	0.729	0.057
NDEG = 2000										
SwarnSeq	904	1096	2096	0.301	0.087	0.548	0.452	0.846	0.796	0.362
DEGSeq	533	1467	2467	0.178	0.116	0.734	0.267	0.819	0.748	0.213
DESeq2	71	1929	2929	0.024	0.153	0.965	0.036	0.785	0.689	0.028
DESingle	403	1597	2597	0.134	0.126	0.799	0.202	0.810	0.732	0.161
EdgeR	287	1713	2713	0.096	0.136	0.857	0.144	0.801	0.717	0.115

Limma	128	1872	2872	0.043	0.148	0.936	0.064	0.789	0.697	0.051
DECENT	86	1913	2914	0.029	0.151	0.957	0.043	0.786	0.691	0.034
MAST	482	1518	2518	0.161	0.120	0.759	0.241	0.815	0.742	0.193
Monocle	316	1684	2684	0.105	0.133	0.842	0.158	0.803	0.721	0.126
NODES	108	1892	2892	0.036	0.150	0.946	0.054	0.788	0.694	0.043
ScDD	449	1551	2551	0.150	0.123	0.776	0.225	0.813	0.738	0.180
BPSC	190	1810	2810	0.063	0.143	0.905	0.095	0.794	0.705	0.076
NDEG = 2500										
SwarnSeq	1031	1469	1969	0.344	0.116	0.588	0.412	0.850	0.780	0.375
DEGSeq	628	1872	2372	0.209	0.148	0.749	0.251	0.819	0.729	0.228
DESeq2	115	2385	2885	0.038	0.189	0.954	0.046	0.780	0.663	0.042
DESingle	520	1980	2480	0.173	0.157	0.792	0.208	0.811	0.715	0.189
EdgeR	411	2089	2589	0.137	0.165	0.836	0.164	0.803	0.701	0.149
Limma	159	2341	2841	0.053	0.185	0.936	0.064	0.784	0.669	0.058
DECENT	185	2314	2815	0.062	0.183	0.926	0.074	0.786	0.672	0.067
MAST	609	1891	2391	0.203	0.150	0.756	0.244	0.818	0.726	0.221
Monocle	426	2074	2574	0.142	0.164	0.830	0.170	0.804	0.703	0.155
NODES	117	2383	2883	0.039	0.189	0.953	0.047	0.781	0.663	0.043
ScDD	568	1932	2432	0.189	0.153	0.773	0.227	0.815	0.721	0.207
BPSC	271	2229	2729	0.090	0.176	0.892	0.108	0.792	0.683	0.099
NDEG = 3000										
SwarnSeq	1145	1855	1855	0.382	0.147	0.618	0.382	0.853	0.763	0.382
DEGSeq	717	2283	2283	0.239	0.181	0.761	0.239	0.819	0.708	0.239
DESeq2	192	2808	2808	0.064	0.222	0.936	0.064	0.778	0.641	0.064
DESingle	648	2352	2352	0.216	0.186	0.784	0.216	0.814	0.699	0.216
EdgeR	552	2448	2448	0.184	0.194	0.816	0.184	0.806	0.687	0.184
Limma	213	2787	2787	0.071	0.221	0.929	0.071	0.779	0.643	0.071
DECENT	368	2628	2632	0.123	0.208	0.877	0.123	0.792	0.664	0.123
MAST	724	2276	2276	0.241	0.180	0.759	0.241	0.820	0.709	0.241
Monocle	564	2436	2436	0.188	0.193	0.812	0.188	0.807	0.688	0.188
NODES	136	2864	2864	0.045	0.227	0.955	0.045	0.773	0.634	0.045
ScDD	659	2341	2341	0.220	0.185	0.780	0.220	0.815	0.701	0.220
BPSC	398	2602	2602	0.133	0.206	0.867	0.133	0.794	0.667	0.133

Table 6.8. Performance evaluation metrics for GSE53638 (Data 3) for scRNA-seq data.

					NDEG = 500					
Methods	**TP**	**FP**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	136	364	2864	0.045	0.030	0.728	0.272	0.803	0.785	0.078
DEGSeq	114	386	2886	0.038	0.032	0.772	0.228	0.801	0.782	0.065
DESeq2	21	479	2979	0.007	0.040	0.958	0.042	0.795	0.770	0.012
DESingle	213	287	2787	0.071	0.024	0.574	0.426	0.808	0.795	0.122
EdgeR	151	349	2849	0.050	0.029	0.698	0.302	0.804	0.787	0.086

Limma	161	339	2839	0.054	0.028	0.678	0.322	0.804	0.788	0.092
DECENT	14	486	2986	0.005	0.040	0.972	0.028	0.794	0.769	0.008
MAST	145	355	2855	0.048	0.030	0.710	0.290	0.803	0.786	0.083
Monocle	94	406	2906	0.031	0.034	0.812	0.188	0.800	0.779	0.054
NODES	88	412	2912	0.029	0.034	0.824	0.176	0.799	0.779	0.050
scDD	42	458	2958	0.014	0.038	0.916	0.084	0.796	0.772	0.024
BPSC	81	419	2919	0.027	0.035	0.838	0.162	0.799	0.778	0.046
NDEG = 1000										
SwarnSeq	374	626	2626	0.125	0.052	0.626	0.374	0.813	0.783	0.187
DEGSeq	249	751	2751	0.083	0.063	0.751	0.249	0.804	0.767	0.125
DESeq2	49	951	2951	0.016	0.079	0.951	0.049	0.789	0.740	0.025
DESingle	342	658	2658	0.114	0.055	0.658	0.342	0.810	0.779	0.171
EdgeR	271	729	2729	0.090	0.061	0.729	0.271	0.805	0.770	0.136
Limma	253	747	2747	0.084	0.062	0.747	0.253	0.804	0.767	0.127
DECENT	52	948	2948	0.017	0.079	0.948	0.052	0.790	0.740	0.026
MAST	257	743	2743	0.086	0.062	0.743	0.257	0.804	0.768	0.129
Monocle	184	816	2816	0.061	0.068	0.816	0.184	0.799	0.758	0.092
NODES	163	837	2837	0.054	0.070	0.837	0.163	0.798	0.755	0.082
scDD	57	943	2943	0.019	0.078	0.943	0.057	0.790	0.741	0.029
BPSC	140	860	2860	0.047	0.072	0.860	0.140	0.796	0.752	0.070
NDEG = 1500										
SwarnSeq	656	844	2344	0.219	0.070	0.563	0.437	0.827	0.788	0.292
DEGSeq	358	1142	2642	0.119	0.095	0.761	0.239	0.804	0.748	0.159
DESeq2	100	1400	2900	0.033	0.117	0.933	0.067	0.785	0.714	0.044
DESingle	469	1031	2531	0.156	0.086	0.687	0.313	0.813	0.763	0.208
EdgeR	397	1103	2603	0.132	0.092	0.735	0.265	0.807	0.753	0.176
Limma	316	1184	2684	0.105	0.099	0.789	0.211	0.801	0.742	0.140
DECENT	136	1364	2864	0.045	0.114	0.909	0.091	0.788	0.718	0.060
MAST	373	1127	2627	0.124	0.094	0.751	0.249	0.806	0.750	0.166
Monocle	286	1214	2714	0.095	0.101	0.809	0.191	0.799	0.738	0.127
NODES	250	1250	2750	0.083	0.104	0.833	0.167	0.796	0.734	0.111
scDD	62	1438	2938	0.021	0.120	0.959	0.041	0.783	0.709	0.028
BPSC	229	1271	2771	0.076	0.106	0.847	0.153	0.795	0.731	0.102
NDEG =2000										
SwarnSeq	847	1153	2153	0.282	0.096	0.577	0.424	0.835	0.780	0.339
DEGSeq	433	1567	2567	0.144	0.130	0.784	0.217	0.803	0.725	0.173
DESeq2	178	1822	2822	0.059	0.152	0.911	0.089	0.783	0.691	0.071
DESingle	593	1407	2407	0.198	0.117	0.704	0.297	0.815	0.746	0.237
EdgeR	541	1459	2459	0.180	0.121	0.730	0.271	0.811	0.739	0.216
Limma	390	1610	2610	0.130	0.134	0.805	0.195	0.799	0.719	0.156
DECENT	260	1740	2740	0.087	0.145	0.870	0.130	0.789	0.702	0.104
MAST	505	1495	2495	0.168	0.124	0.748	0.253	0.808	0.734	0.202
Monocle	426	1574	2574	0.142	0.131	0.787	0.213	0.802	0.724	0.170
NODES	357	1643	2643	0.119	0.137	0.822	0.179	0.797	0.714	0.143

scDD	69	1931	2931	0.023	0.161	0.966	0.035	0.775	0.676	0.028
BPSC	351	1649	2649	0.117	0.137	0.825	0.176	0.796	0.714	0.140
					NDEG = 2500					
SwarnSeq	1005	1495	1995	0.335	0.124	0.598	0.402	0.841	0.768	0.365
DEGSeq	562	1938	2438	0.187	0.161	0.775	0.225	0.805	0.709	0.204
DESeq2	260	2240	2740	0.087	0.186	0.896	0.104	0.781	0.668	0.095
DESingle	717	1783	2283	0.239	0.148	0.713	0.287	0.818	0.729	0.261
EdgeR	689	1811	2311	0.230	0.151	0.724	0.276	0.815	0.725	0.251
Limma	482	2018	2518	0.161	0.168	0.807	0.193	0.799	0.698	0.175
DECENT	416	2084	2584	0.139	0.173	0.834	0.166	0.793	0.689	0.151
MAST	635	1865	2365	0.212	0.155	0.746	0.254	0.811	0.718	0.231
Monocle	558	1942	2442	0.186	0.162	0.777	0.223	0.805	0.708	0.203
NODES	469	2031	2531	0.156	0.169	0.812	0.188	0.798	0.696	0.171
scDD	75	2425	2925	0.025	0.202	0.970	0.030	0.766	0.644	0.027
BPSC	478	2022	2522	0.159	0.168	0.809	0.191	0.798	0.697	0.174
					NDEG = 3000					
SwarnSeq	1146	1854	1854	0.382	0.154	0.618	0.382	0.846	0.753	0.382
DEGSeq	685	2315	2315	0.228	0.193	0.772	0.228	0.807	0.692	0.228
DESeq2	317	2683	2683	0.106	0.223	0.894	0.106	0.777	0.643	0.106
DESingle	844	2156	2156	0.281	0.179	0.719	0.281	0.821	0.713	0.281
EdgeR	862	2138	2138	0.287	0.178	0.713	0.287	0.822	0.715	0.287
Limma	577	2423	2423	0.192	0.202	0.808	0.192	0.798	0.677	0.192
DECENT	618	2382	2382	0.206	0.198	0.794	0.206	0.802	0.683	0.206
MAST	785	2215	2215	0.262	0.184	0.738	0.262	0.816	0.705	0.262
Monocle	710	2290	2290	0.237	0.191	0.763	0.237	0.809	0.695	0.237
NODES	578	2422	2422	0.193	0.202	0.807	0.193	0.798	0.677	0.193
scDD	89	2911	2911	0.030	0.242	0.970	0.030	0.758	0.612	0.030
BPSC	645	2355	2355	0.215	0.196	0.785	0.215	0.804	0.686	0.215

Table 6.9. Performance evaluation metrics for GSE65525 scRNA-seq data.

					NDEG = 500					
Methods	**TP**	**FP**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	500	0	2500	0.167	0.000	0.000	1.000	0.893	0.896	0.286
DEGSeq	191	309	2809	0.064	0.015	0.618	0.382	0.880	0.870	0.109
DESeq2	155	345	2845	0.052	0.016	0.690	0.310	0.879	0.867	0.089
DESingle	221	279	2779	0.074	0.013	0.558	0.442	0.882	0.872	0.126
EdgeR	206	294	2794	0.069	0.014	0.588	0.412	0.881	0.871	0.118
Limma	38	462	2962	0.013	0.022	0.924	0.076	0.874	0.857	0.022
DECENT	178	322	2822	0.059	0.015	0.644	0.356	0.880	0.869	0.102
MAST	77	423	2923	0.026	0.020	0.846	0.154	0.875	0.860	0.044
Monocle	134	366	2866	0.045	0.017	0.732	0.268	0.878	0.865	0.077
NODES	59	441	2941	0.020	0.021	0.882	0.118	0.875	0.859	0.034
scDD	72	428	2928	0.024	0.020	0.856	0.144	0.875	0.860	0.041
BPSC	135	365	2865	0.045	0.017	0.730	0.270	0.878	0.865	0.077

NDEG=1000										
SwarnSeq	1000	0	2000	0.333	0.000	0.000	1.000	0.913	0.917	0.500
DEGSeq	294	706	2706	0.098	0.034	0.706	0.294	0.882	0.858	0.147
DESeq2	226	774	2774	0.075	0.037	0.774	0.226	0.879	0.852	0.113
DESingle	306	694	2694	0.102	0.033	0.694	0.306	0.883	0.859	0.153
EdgeR	288	712	2712	0.096	0.034	0.712	0.288	0.882	0.857	0.144
Limma	85	915	2915	0.028	0.044	0.915	0.085	0.873	0.840	0.043
DECENT	326	674	2674	0.109	0.032	0.674	0.326	0.884	0.860	0.163
MAST	255	745	2745	0.085	0.036	0.745	0.255	0.880	0.854	0.128
Monocle	283	717	2717	0.094	0.034	0.717	0.283	0.882	0.857	0.142
NODES	105	895	2895	0.035	0.043	0.895	0.105	0.874	0.842	0.053
scDD	72	928	2928	0.024	0.044	0.928	0.072	0.872	0.839	0.036
BPSC	204	796	2796	0.068	0.038	0.796	0.204	0.878	0.850	0.102
NDEG = 1500										
SwarnSeq	1442	58	1558	0.481	0.003	0.039	0.961	0.931	0.933	0.641
DEGSeq	402	1098	2598	0.134	0.052	0.732	0.268	0.884	0.846	0.179
DESeq2	263	1237	2737	0.088	0.059	0.825	0.175	0.878	0.834	0.117
DESingle	371	1129	2629	0.124	0.054	0.753	0.247	0.883	0.843	0.165
EdgeR	348	1152	2652	0.116	0.055	0.768	0.232	0.882	0.841	0.155
Limma	123	1377	2877	0.041	0.066	0.918	0.082	0.872	0.822	0.055
DECENT	438	1062	2562	0.146	0.051	0.708	0.292	0.886	0.849	0.195
MAST	431	1069	2569	0.144	0.051	0.713	0.287	0.886	0.848	0.192
Monocle	381	1119	2619	0.127	0.053	0.746	0.254	0.883	0.844	0.169
NODES	146	1354	2854	0.049	0.065	0.903	0.097	0.873	0.824	0.065
scDD	72	1428	2928	0.024	0.068	0.952	0.048	0.870	0.818	0.032
BPSC	259	1241	2741	0.086	0.059	0.827	0.173	0.878	0.834	0.115
NDEG = 2000										
SwarnSeq	1442	558	1558	0.481	0.027	0.279	0.721	0.929	0.912	0.577
DEGSeq	503	1497	2497	0.168	0.071	0.749	0.252	0.886	0.833	0.201
DESeq2	294	1706	2706	0.098	0.081	0.853	0.147	0.877	0.816	0.118
DESingle	440	1560	2560	0.147	0.074	0.780	0.220	0.883	0.828	0.176
EdgeR	423	1577	2577	0.141	0.075	0.789	0.212	0.883	0.827	0.169
Limma	200	1800	2800	0.067	0.086	0.900	0.100	0.872	0.808	0.080
DECENT	532	1468	2468	0.177	0.070	0.734	0.266	0.888	0.836	0.213
MAST	591	1409	2409	0.197	0.067	0.705	0.296	0.890	0.841	0.236
Monocle	473	1527	2527	0.158	0.073	0.764	0.237	0.885	0.831	0.189
NODES	173	1827	2827	0.058	0.087	0.914	0.087	0.871	0.806	0.069
scDD	72	1928	2928	0.024	0.092	0.964	0.036	0.867	0.797	0.029
BPSC	298	1702	2702	0.099	0.081	0.851	0.149	0.877	0.816	0.119
NDEG = 2500										
SwarnSeq	1442	1058	1558	0.481	0.050	0.423	0.577	0.927	0.891	0.524
DEGSeq	589	1911	2411	0.196	0.091	0.764	0.236	0.888	0.820	0.214
DESeq2	323	2177	2677	0.108	0.104	0.871	0.129	0.875	0.797	0.117
DESingle	496	2004	2504	0.165	0.096	0.802	0.198	0.883	0.812	0.180

EdgeR	488	2012	2512	0.163	0.096	0.805	0.195	0.883	0.811	0.177
Limma	351	2149	2649	0.117	0.103	0.860	0.140	0.877	0.800	0.128
DECENT	617	1883	2383	0.206	0.090	0.753	0.247	0.889	0.822	0.224
MAST	750	1750	2250	0.250	0.084	0.700	0.300	0.895	0.833	0.273
Monocle	553	1947	2447	0.184	0.093	0.779	0.221	0.886	0.817	0.201
NODES	199	2301	2801	0.066	0.110	0.920	0.080	0.869	0.787	0.072
scDD	72	2428	2928	0.024	0.116	0.971	0.029	0.864	0.776	0.026
BPSC	326	2174	2674	0.109	0.104	0.870	0.130	0.875	0.798	0.119
					NDEG = 3000					
SwarnSeq	1442	1558	1558	0.481	0.074	0.519	0.481	0.926	0.870	0.481
DEGSeq	719	2281	2281	0.240	0.109	0.760	0.240	0.891	0.810	0.240
DESeq2	350	2650	2650	0.117	0.126	0.883	0.117	0.874	0.779	0.117
DESingle	552	2448	2448	0.184	0.117	0.816	0.184	0.883	0.796	0.184
EdgeR	545	2455	2455	0.182	0.117	0.818	0.182	0.883	0.795	0.182
Limma	496	2504	2504	0.165	0.119	0.835	0.165	0.881	0.791	0.165
DECENT	704	2296	2296	0.235	0.110	0.765	0.235	0.890	0.808	0.235
MAST	906	2094	2094	0.302	0.100	0.698	0.302	0.900	0.825	0.302
Monocle	632	2368	2368	0.211	0.113	0.789	0.211	0.887	0.802	0.211
NODES	217	2783	2783	0.072	0.133	0.928	0.072	0.867	0.768	0.072
scDD	72	2928	2928	0.024	0.140	0.976	0.024	0.860	0.756	0.024
BPSC	364	2636	2636	0.121	0.126	0.879	0.121	0.874	0.780	0.121

Table 6.10. Performance evaluation metrics for Tung's (GSE77288) data.

					NDEG = 500					
Methods	**TP**	**FP**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	482	18	2518	0.161	0.001	0.036	0.964	0.837	0.841	0.275
DEGSeq	107	393	2893	0.036	0.030	0.786	0.214	0.813	0.794	0.061
DESeq2	134	366	2866	0.045	0.028	0.732	0.268	0.815	0.797	0.077
DESingle	227	273	2773	0.076	0.021	0.546	0.454	0.821	0.809	0.130
EdgeR	199	301	2801	0.066	0.023	0.602	0.398	0.819	0.806	0.114
Limma	292	208	2708	0.097	0.016	0.416	0.584	0.825	0.817	0.167
DECENT	172	328	2828	0.057	0.025	0.656	0.344	0.817	0.802	0.098
MAST	159	341	2841	0.053	0.026	0.682	0.318	0.816	0.801	0.091
Monocle	96	404	2904	0.032	0.031	0.808	0.192	0.812	0.793	0.055
NODES	187	313	2813	0.062	0.024	0.626	0.374	0.818	0.804	0.107
scDD	141	359	2859	0.047	0.028	0.718	0.282	0.815	0.798	0.081
BPSC	114	386	2886	0.038	0.030	0.772	0.228	0.813	0.795	0.065
					NDEG = 1000					
SwarnSeq	946	54	2054	0.315	0.004	0.054	0.946	0.863	0.868	0.473
DEGSeq	159	841	2841	0.053	0.065	0.841	0.159	0.810	0.770	0.080
DESeq2	234	766	2766	0.078	0.059	0.766	0.234	0.815	0.779	0.117
DESingle	426	574	2574	0.142	0.044	0.574	0.426	0.828	0.803	0.213
EdgeR	347	653	2653	0.116	0.050	0.653	0.347	0.824	0.794	0.174

Limma	607	393	2393	0.202	0.030	0.393	0.607	0.840	0.825	0.304
DECENT	351	649	2649	0.117	0.050	0.649	0.351	0.823	0.793	0.176
MAST	330	670	2670	0.110	0.052	0.670	0.330	0.821	0.791	0.165
Monocle	263	737	2737	0.088	0.057	0.737	0.263	0.817	0.782	0.132
NODES	355	645	2645	0.118	0.050	0.645	0.355	0.823	0.794	0.178
scDD	218	782	2782	0.073	0.060	0.782	0.218	0.814	0.777	0.109
BPSC	213	787	2787	0.071	0.061	0.787	0.213	0.814	0.776	0.107
NDEG = 1500										
SwarnSeq	1311	189	1689	0.437	0.015	0.126	0.874	0.883	0.882	0.583
DEGSeq	293	1207	2707	0.098	0.093	0.805	0.195	0.813	0.755	0.130
DESeq2	338	1162	2662	0.113	0.090	0.775	0.225	0.816	0.760	0.150
DESingle	612	888	2388	0.204	0.069	0.592	0.408	0.835	0.795	0.272
EdgeR	486	1014	2514	0.162	0.077	0.676	0.324	0.828	0.782	0.216
Limma	887	613	2113	0.296	0.047	0.409	0.591	0.854	0.829	0.394
DECENT	518	982	2482	0.173	0.076	0.655	0.345	0.828	0.783	0.230
MAST	497	1003	2503	0.166	0.077	0.669	0.331	0.827	0.780	0.221
Monocle	398	1102	2602	0.133	0.085	0.735	0.265	0.820	0.768	0.177
NODES	523	977	2477	0.174	0.075	0.651	0.349	0.829	0.784	0.232
scDD	279	1221	2721	0.093	0.094	0.814	0.186	0.812	0.753	0.124
BPSC	321	1179	2679	0.107	0.091	0.786	0.214	0.815	0.758	0.143
NDEG = 2000										
SwarnSeq	1532	468	1468	0.511	0.036	0.234	0.766	0.895	0.879	0.613
DEGSeq	420	1580	2580	0.140	0.121	0.790	0.210	0.816	0.740	0.168
DESeq2	426	1574	2574	0.142	0.121	0.787	0.213	0.816	0.740	0.170
DESingle	758	1242	2242	0.253	0.096	0.621	0.379	0.839	0.782	0.303
EdgeR	625	1375	2375	0.208	0.104	0.688	0.313	0.833	0.769	0.250
Limma	1104	896	1896	0.368	0.069	0.448	0.552	0.864	0.825	0.442
DECENT	664	1336	2336	0.221	0.103	0.668	0.332	0.833	0.770	0.266
MAST	642	1358	2358	0.214	0.105	0.679	0.321	0.831	0.767	0.257
Monocle	491	1509	2509	0.164	0.116	0.755	0.246	0.820	0.748	0.196
NODES	636	1364	2364	0.212	0.105	0.682	0.318	0.831	0.766	0.254
scDD	327	1673	2673	0.109	0.129	0.837	0.164	0.808	0.728	0.131
BPSC	416	1584	2584	0.139	0.122	0.792	0.208	0.815	0.739	0.166
NDEG = 2500										
SwarnSeq	1713	787	1287	0.571	0.061	0.315	0.685	0.904	0.870	0.623
DEGSeq	514	1986	2486	0.171	0.152	0.794	0.206	0.816	0.721	0.187
DESeq2	525	1975	2475	0.175	0.152	0.790	0.210	0.816	0.721	0.191
DESingle	901	1599	2099	0.300	0.123	0.640	0.360	0.844	0.768	0.328
EdgeR	769	1731	2231	0.256	0.130	0.692	0.308	0.838	0.757	0.280
Limma	1293	1207	1707	0.431	0.093	0.483	0.517	0.873	0.817	0.470
DECENT	791	1709	2209	0.264	0.131	0.684	0.316	0.837	0.755	0.288
MAST	772	1728	2228	0.257	0.133	0.691	0.309	0.834	0.752	0.281
Monocle	598	1902	2402	0.199	0.147	0.761	0.239	0.821	0.730	0.217

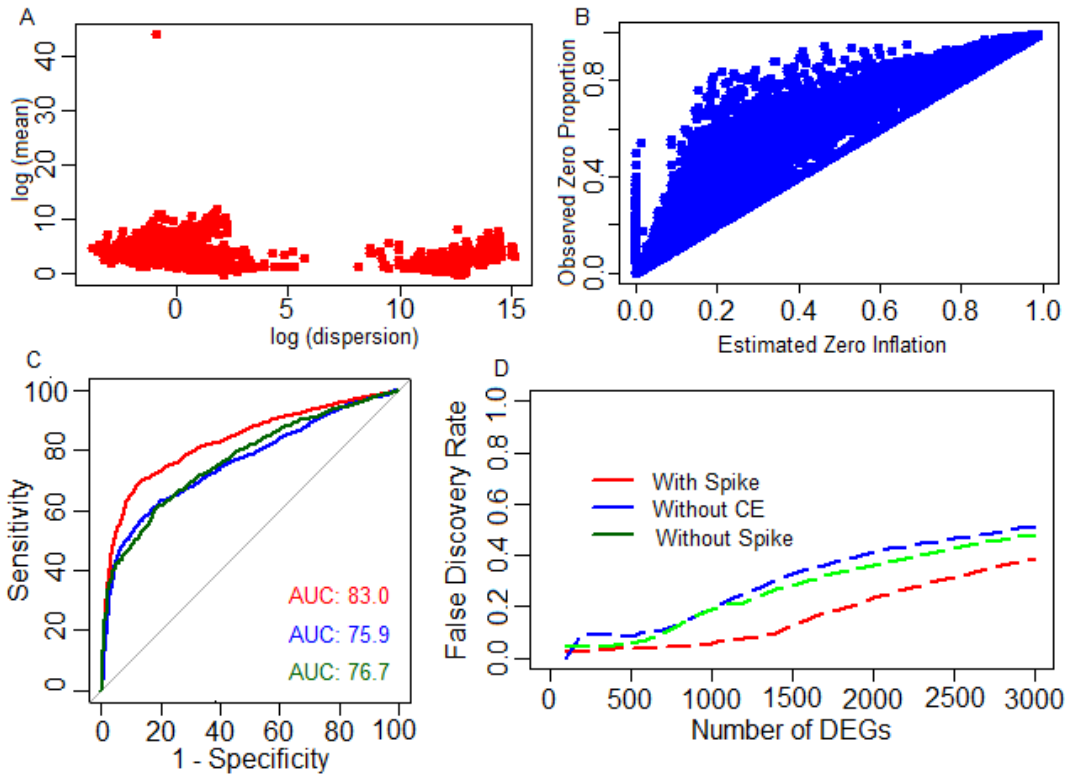
NODES	752	1748	2248	0.251	0.135	0.699	0.301	0.833	0.750	0.273
scDD	394	2106	2606	0.131	0.163	0.842	0.158	0.806	0.705	0.143
BPSC	514	1986	2486	0.171	0.153	0.794	0.206	0.815	0.720	0.187
				NDEG = 3000						
SwarnSeq	1846	1154	1154	0.615	0.089	0.385	0.615	0.911	0.855	0.615
DEGSeq	611	2389	2389	0.204	0.183	0.796	0.204	0.817	0.702	0.204
DESeq2	629	2371	2371	0.210	0.183	0.790	0.210	0.817	0.703	0.210
DESingle	1024	1976	1976	0.341	0.152	0.659	0.341	0.848	0.752	0.341
EdgeR	909	2091	2091	0.303	0.156	0.697	0.303	0.844	0.745	0.303
Limma	1456	1544	1544	0.485	0.119	0.515	0.485	0.881	0.806	0.485
DECENT	901	2099	2099	0.300	0.160	0.700	0.300	0.840	0.739	0.300
MAST	886	2114	2114	0.295	0.163	0.705	0.295	0.837	0.735	0.295
Monocle	698	2302	2302	0.233	0.178	0.767	0.233	0.822	0.711	0.233
NODES	836	2164	2164	0.279	0.167	0.721	0.279	0.833	0.729	0.279
scDD	461	2539	2539	0.154	0.196	0.846	0.154	0.804	0.682	0.154
BPSC	614	2386	2386	0.205	0.184	0.795	0.205	0.816	0.701	0.205

NDEG: Number of differentially expressed genes; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Effect of spike-in on performance

We evaluated the performance of the SwarnSeq method on data from GSE77288 for which spike-in and molecular concentration data is publicly available. For this purpose, we considered the following comparison settings: (a) spike-in data available; (b) spike-in not available (capture rates estimated from the data); (c) data unadjusted with cell capture rates. In other words, this comparison setting allowed us to examine the impact of external spike-ins and further capture rates on the DE performance of SwarnSeq method. The results are shown in Figure 6.11 and Table 6.11. It was observed that the SwarnSeq performed better when capture rates were estimated from external spike-ins as assessed in terms of AUC (Figure 6.11). However, there was a decrease in AUC value when the capture rates of cells were estimated from the count data (Figure 6.11). Further, the SwarnSeq had the least AUC when the observed counts were not adjusted with cell capture rates.

Figure 6.11. Performance analysis of SwarnSeq method in presence of spike-ins. (A) Scatter plot showing the relationship of mean and dispersion parameters of the genes. (B) Scatter plot comparing the observed value of zero proportions and estimated zero inflation parameters of genes (C) ROC curves are shown for SwarnSeq method (i) when spike-in information is considered (red); (ii) when spike-in data are not considered and capture efficiencies are estimated from the data (green); and (iii) Unadjusted for capture efficiency. (D) FDR curves of SwarnSeq method are shown for: (i) when spike-in information is considered (red); (ii) when spike-in data are not considered and capture efficiencies are estimated from the data (green); (iii) Unadjusted for capture efficiency. (E) Various performance measures are listed for SwarnSeq method under different conditions.



E

	TP	FP	TN	FN	TPR	FPR	FDR	PPR	NPV	ACC	F1
With Spike	1532	468	12487	1468	0.511	0.036	0.234	0.766	0.895	0.879	0.613
Without CE	1175	825	12130	1825	0.392	0.064	0.413	0.588	0.869	0.834	0.470
W/O Spike	1257	713	12242	1743	0.419	0.055	0.362	0.638	0.875	0.846	0.506

Table 6.11. Performance of SwarnSeq method under three different scenarios.

	TP	FP	TN	FN	TPR	FPR	PPR	NPV	ACC	F1
					NDEG = 500					
With Spike	482	18	12937	2518	0.161	0.001	0.964	0.837	0.841	0.275
Unadjusted	457	43	12912	2543	0.152	0.003	0.914	0.835	0.838	0.261
Without spike	468	27	12928	2532	0.156	0.002	0.945	0.836	0.840	0.268
					NDEG = 1000					
With Spike	946	54	12901	2054	0.315	0.004	0.946	0.863	0.868	0.473

Unadjusted	809	191	12764	2191	0.270	0.015	0.809	0.853	0.851	0.405
Without spike	795	187	12768	2205	0.265	0.014	0.810	0.853	0.850	0.399
					NDEG = 1500					
With Spike	1311	189	12766	1689	0.437	0.015	0.874	0.883	0.882	0.583
Unadjusted	1008	492	12463	1992	0.336	0.038	0.672	0.862	0.844	0.448
Without spike	1059	416	12539	1941	0.353	0.032	0.718	0.866	0.852	0.473
					NDEG = 2500					
With Spike	1713	787	12168	1287	0.571	0.061	0.685	0.904	0.870	0.623
Unadjusted	1342	1158	11797	1658	0.447	0.089	0.537	0.877	0.824	0.488
Without spike	1414	1053	11902	1586	0.471	0.081	0.573	0.882	0.835	0.517
					NDEG = 3000					
With Spike	1846	1154	11801	1154	0.615	0.089	0.615	0.911	0.855	0.615
Unadjusted	1466	1534	11421	1534	0.489	0.118	0.489	0.882	0.808	0.489
Without spike	1542	1424	11531	1458	0.514	0.110	0.520	0.888	0.819	0.517

Under the FDR based comparison setting, the SwarnSeq had the smallest FDR values, when the capture rates of cells were estimated from the spike-in data (Figure 6.11). Further, SwarnSeq performed poorly when the observed counts were not adjusted with capture rates of the cells, as compared to the adjusted scRNA-seq counts. The results from the third comparison setting, *i.e.* comparative analysis based on performance metrics, are shown in Figure 6.11 and Table 6.11. It was found that when the capture rates were estimated from the spike-ins and incorporated in SwarnSeq, its performance was better as compared to other two situations (Figure 6.11, Table 6.11). Thus, we have convincingly demonstrated the viability of using the external spike-in capture rates for endogenous RNA in SwarnSeq modeling, and subsequently found its DE performance is both robust and better.

SwarnSeq as Differential Zero Inflation tool

The SwarnSeq method provides an excellent platform for performing DZI analysis of genes. DZI genes are detected using a ZINB model under a GLM framework. For identification of DZI genes on real dataset, we set 1E-10 as threshold for

adjusted *p-values* computed through the SwarnSeq. For instance, at this threshold we identified 2936 DZI genes for GSE29087 data. This means, 2936 genes have a significant difference in the number of cells whose expressions are zeros across two cellular groups. Similar interpretations can be made for other datasets.

Our SwarnSeq model provides an opportunity to classify the influential genes into gene types with respect to their differential zero inflation and expression. Through this, the identified genes can be grouped into various gene types, and the results are shown in Table 6.12.

Table 6.12 Classification of DE and DZI genes.

Datasets	DEG	DEZIG	DZIG	Non-DEG
GSE29087	4930	2789	149	3567
GSE53638 (data1)	2406	278	408	11771
GSE53638 (data 2)	1831	2789	5013	6004
GSE53638 (data 3)	1733	3101	4673	5507
GSE65525	2033	15194	5799	929
GSE75790	3993	9874	2865	3852
GSE92495	757	324	5	14438
GSE109999	5694	6386	71	903
GSE111108	27	7187	87	10021
GSE115469	24	7745	6296	3231
GSE77288	1426	119	619	13791

DEG: Differentially Expressed; DZIG: Differentially Zero Inflated; DEZIG: Both DEG and DZIG

For instance, the GSE29087 data, the SwarnSeq identified 4930 genes as DEG, 2789 genes as DEZIG, and 149 as only DZIG (Table 6.12). This means that out of 15234 genes, the mean expression of non-zero counts of 4930 genes are expressed differentially across the two cellular groups. While, for 2789 genes, there is a significant proportion of cells whose expressions are zero across two cellular populations (however the mean of non-zeros counts of these genes in the remaining cells are significantly different) and only 149 genes had a significant

number of cells as zero expressions across the cellular populations (Table 6.12). Similar type of interpretations can be made for other datasets from Table 6.12.

Discussion

In this Chapter, we presented SwarnSeq, an improved statistical method, for performing analysis on counts data derived from scRNA-seq study. Our method is capable of performing reliable statistical tests on gene mean abundance, zero inflation, and classification of influential genes derived from scRNA-seq counts expression data. It uses the ZINB model to model the observed UMI counts. Further, the UMI provides an excellent opportunity to model the transcriptional capturing process. In other words, the observed counts data are adjusted with cell capture rates through a simple binomial model. Moreover, RNA spike-ins data including the external RNA spike-ins [253], can give valuable insights into the technical variation in scRNA-seq study. This raises a key question of whether and how to use spike-ins in data analyses. For instance, when they are available, they can be used to estimate the capture rates for cells. This property is well integrated in our SwarnSeq approach. Thus, SwarnSeq is capable of modeling capture rates using spike-ins data, if they are available and can estimate the capture rates from the observed data, if spike-ins are not available. We established statistical a theory for adjusting the UMI counts data with the molecular capturing process derived from real scRNA-seq experiments. Moreover, the SwarnSeq operates through various analytical steps including, pre-processing, normalization, estimation of gene parameters, detection of DE genes, and DZI genes, selection of top genes, and classification of genes into sub-types. The SwarnSeq method employs

different normalization methods such as modified median normalization [206] and trimmed mean of *M* values [204] to remove the amplification bias from the data. Thus, SwarnSeq is compatible with different normalization strategies.

Here, we established the statistical basis for the distributional nature of the observed scRNA-seq count in presence of cell capture rates. Further, we have empirically shown the suitability of the ZINB model for fitting zero inflated, and overdispersed count data over other count models, such as NB, PD, HD, and ZIPD. Moreover, the study of ZINB over NB model for estimation of parameters indicated that the latter overestimated the dispersion to accommodate excess overdispersion and underestimated the mean to accommodate the extra zeros present in scRNA-seq data. In UMI data, factors such as technical noise, dropout events, and low molecule capturing have substantial overdispersion and zero-inflation, and a NB model is not appropriate. Hence, we implemented a ZINB model in our SwarnSeq method to fit the observed scRNA-seq count data and to obtain better estimates of the gene-wise means and dispersions.

The SwarnSeq method models the unwanted variation in mean transcript abundance of genes attributed to different sources, such as cellular groups, cell clusters, and other cell co-variates. This means, it provides reliable MLEs of the effects of the cellular groups, cell clusters, and cell co-variates using the EM algorithm. Further, it detects the influential genes which are DE under a GLM framework. Here, these genes are identified based on the statistical significance values adjusted over multiple hypothesis testing. This provides statistically meaningful and biologically interpretable values in [0, 1] for genes in scRNA-seq

data. The benchmarking of methods indicated the better performance of our SwarnSeq method over other methods. This comparative analysis was carried out on three different comparison settings, *i.e.,* AUROC, FDR, and other performance metrics on multiple real scRNA-seq datasets.

The SwarnSeq method can also be extended to carry out other types of tests, including the differential testing of zero proportions of genes across the cell populations. Here, we considered the zero-inflation parameter of genes as a function of the effects of cellular populations, cell clusters, and other cell co-variates. Then, a linear logit model was used to test for biological differences in zero inflation. To statistically measure this, a statistical significance value adjusted over multiple hypothesis testing, was assigned to each gene. This measure provided biologically interpretable values to genes, which showed there was significant difference in the proportion of zero expressions across the cellular populations due to technical variation, dropout events, least transcript abundance, *etc*. The available scRNA-seq tools mostly focused on performing DE analysis of genes and ignores the zero-inflation analysis which is an integral part of the scRNA-seq experiments. Therefore, our SwarnSeq method can perform DZI analysis including DE analysis of genes using the observed scRNA-seq counts data adjusted over molecular capturing process. Additionally, it also provides option for classifying the detected influential genes into various gene types according to their differential expression and zero inflation.

Multilevel statistical models fitted with an EM algorithm are computationally intensive and time consuming. ZINB models are implemented in several tools like

DEsingle [211], DECENT [209], which are time consuming. For instance, a data with 500 cells and 3000 genes DECENT took 120 min, while the largest dataset (GSE115469: 5466 cells and 17316 genes) took time up to 24 hours to finish on a 10-core DELL PC with 32 GB RAM with Intel(R) Xeon(R) v4 CPUS @ 2.60 GHz. The SwarnSeq method required less computational time than DECENT and DEsingle with much superior performance along with additional features. Besides, it can even be used on a PC or workstation computer for analyzing large scRNA-seq datasets. The benchmarking of the SwarnSeq method on multiple real datasets over a wide range of statistical criteria indicated its better and robust performance over the existing methods. Further, the SwarnSeq method will surely help the experimental biologist and genome researchers to identify true DE genes for their experiments.

“A single cell is regarded as the true biological atom…”

G H Lewes

CHAPTER 7

STATISTICAL APPROACH FOR GENE SET ANALYSIS WITH QUANTITATIVE TRAIT LOCI FOR RNA-SEQUENCING DATA

Background

RNA-seq is a powerful technique for studying GE dynamics and regulation in human and nonhuman genomes. Recently, RNA-Seq has surpassed the Microarrays by providing better quantification of GE for very high and low expressed genes, and higher levels of accuracy and reproducibility [284]. Here, the expression of genes are measured in terms of discrete read counts obtained through mapping the sequence reads to reference genome followed by quantification of transcripts abundance [75]. Further, DE analysis is one of powerful downstream analysis performed on the RNA-seq count data to detect DE genes with higher resolution than Microarrays across the two different experimental conditions [205]. Besides, it also allows to study alternative splicing [285], new coding and noncoding RNA transcripts and long noncoding RNAs [286,287]. In other words, the RNA-seq is much more popular and efficient, as it answers a much wider range of questions than Microarrays. Moreover, to interpret the long list DE genes in the context of the underlying phenotypic differences and to gain insights into biological mechanisms [8], secondary genomic analytics, such as GSA is usually popular practice. Expressly, the GSA allows to interpret the high-throughput RNA-seq count data in the context of broader biological context.

GSA methods were initially developed for Microarrays, but later extended to RNA-seq [15]. Here, preparation of ranked gene list (*i.e.,* DE analysis) is a major process, which depends on the nature and distributional properties of the data. For instance, GSA approaches for Microarrays deal with continuous data and expected to follow Gaussian distribution. Contrarily, GE in RNA-seq are non-negative counts (discrete in nature) and assumed to follow a NBD model. Therefore, it may be improper to use GSA techniques meant for Microarray data directly to RNA-seq data. Initially, GSA for RNA-eq data analysis was adopted from Microarrays with the help of data transformation, subsequently new approaches exclusively for RNA-seq were also developed [74]. For instance, VOOM-normalization was used for normalizing the read counts for sequence-depths, then Microarrays GSA approaches are applied on the normalized data [78]. Then, specialized GSA methods for RNA-seq were developed, which includes GOseq [76]. It performs over-representation of GO categories enriched with a long list of DE genes in RNA-Seq data. Further, an easy-to-use web application, iDEP was developed for in-depth analysis of RNA-seq data [77]. Both the methods belong to the ORA category of the GSA, which uses the GO and pathways information to analyze the RNA-seq data [76,77]. These GSA methods only consider the number of DE genes alone and ignore any values associated with them such as read counts, DE score, *etc.* By discarding this data, ORA treats each gene equally by assuming that each gene is independent of the others, which is quite unrealistic in biology [21]. Further, ORA typically focus on the genes in gene set and discards the others. Apart from this, GSA methods based on gene enrichment statistic(s), such as AbsFilterGSEA

[81,82], seqGSEA [83], ssGSEA [288], EGSEA [84], GSVA [79], GSEPD [85], and RNA-Enrich [86] were developed exclusively for RNA-seq data analysis. Further, the reviews of these methods and their comparison can be found in recent studies available in literature [15,289]. However, these techniques are also suffered from limitations, such as they only use DE score to prepare ranked transcript list but ignore this information for gene set testing. Also these approaches use data transformation technique, through which overdispersion, zero inflation, count nature and other inherent nature of RNA-seq data are lost [289].

The contemporary GSA approaches in RNA-seq mostly use GO and pathways information for analyzing gene set [76,77,79,81–86,288], and very useful in establishing links of gene sets with underlying biological processes. However, in plant and complex disease biology, such approaches may not able to establish any formal relation between the underlying genotypes and the trait/phenotype, as most of the traits are quantitative in nature and controlled by polygenes [13,17,132,133]. Apart from the GO and pathways, other biological annotations information, such as QTL, expression QTL, *etc.* are available in public domain databases that may be effectively used for GSA to gain biological insights into the etiology of complex diseases in humans as well as other organisms. For this purpose, a statistical approach and tool was developed to perform GSA with genetically enriched QTL data [17] for Microarrays. This approach has immense use for performing trait/QTL enrichment analysis of gene sets and further, QTL enriched gene sets can be used for molecular breeding programs for biotic/abiotic stress engineering in plants. However, it has some serious limitations, such as only

consider the genes which overlapped with the QTL regions, but failed to consider their corresponding DE scores, treats each gene equally by assuming each gene as independently and identically distributed which is contrary to the real biology. GSAQ uses only the most significant genes, while discards other genes. For instance, a gene input list from Microarray is obtained by setting the arbitrary threshold(s) for FC and *p-values* as 1.5 and 0.01, respectively. With this method, marginally less significant genes (*e.g.*, FC ~1.499 and *p-value* ~ 0.011) are missed, resulting in information loss for some key genes. Under these circumstances, the statistical methodologies for GSA with QTL requires further improvements and advances, which will be very helpful in unraveling genotype-phenotype relationships in plants or in complex diseases.

In this chapter, we propose an improved statistical approach, *i.e.* GSQSeq, for analyzing genes with trait enriched QTL data for RNA-seq studies. This approach considers the genes present in the gene set along with their corresponding scores to analyze in presence of the trait specific QTL data. Here, the enrichment significance of the gene sets was assessed through the p-values computed using the developed test statistic(s). Further, we assessed the performance of the proposed method with the existing ones using performance metrics such as FDR, and $-\log_{10}$(p-*value*) on multiple real crop datasets. For this purpose, we used 7 expression datasets derived from Microarrays and RNA-seq studies in rice. Our analytical findings indicate that the developed approach outperformed the existing method for detecting trait enriched gene sets. For the

benefit of users, we developed GSQSeq R package based on the developed methodology.

Material and Methods

Real Microarray Datasets

Rice GE experimental datasets were collected from GEO database of NCBI for platforms GPL2025 [140]. Here, we used the rice data, as it is a model crop plant, huge amount GE and other related biological datasets are available publicly, and its genome is well annotated. These GE datasets were generated under biotic (fungal (Blast), and insect (Brown plant hopper)), and abiotic (cold and drought) stresses in rice. The QTL datasets for the stresses in rice, *viz.* drought, cold, insect, and fungal, were collected from the Gramene QTL database (http://www.gramene.org/qtl/) [169]. The detail description for the datasets is given in Chapter 4.

Real RNA-seq dataset

The raw sequence datasets of rice (Japonica Group) under salinity stress were collected from the Sequence Read Archive (SRA) database of NCBI (https://www.ncbi.nlm.nih.gov/sra/). The datasets were generated from Illumina HiSeq 2000 with platform GPL13834 (in GEO). This platform consists 323 samples and 29 series of *Oryza sativa*. Among these datasets, we used sequence data pertains to GSE109341, submitted by Formentin *et al*. in January 2018 and last updated June 2018 to test the proposed method [290]. Unlike other datasets, GSE109341 has quite sufficient large number (24; case 12: control: 12) of samples belonging two different contrasting rice cultivars. Further, the sequence datasets

were generated from root and leaf tissue samples under untreated and treated plants. Each sample was made of 6 pooled plants with three biological replicates.

RNA-seq preprocessing and read alignment

The single-end illumina raw sequence reads were collected from SRA database using SRA toolkit (v. 2.9.1-1). Then the raw reads were preprocessed with Trimmomatic toolkit (v. 0.38), which involves removal of adapter sequences, quality filtering, *etc*. Further, the overall quality of preprocessed results was manually inspected using the quality reports generated by FastQC (v. 0.11.7). Moreover, the preprocessed reads were mapped with HISAT (v. 2.1.0) on the *Oryza sativa* v. Nipponbare reference genome, downloaded from the MSU Rice Genome Annotation Project version 7.0 (http://rice.plantbiology.msu.edu/) [291]. The mapping of sequence reads to the reference genome allows to identify their genomic positions. Gene coordinates file (.GFF3) was collected from MSU rice genome browser [291], which also help to map the reads spanning splice junctions.

Transcript assembly and quantification

The success of analysis of RNA-seq data requires the accurate reconstructions and proper quantification of all the isoforms expressed from each gene. Here, we executed the StringTie tool (v. 1.3.4d) to assemble transcripts from the RNA-seq reads that have been aligned to the genome, which primarily involves two steps. First, grouping the reads into distinct gene loci and then assembling each locus into as many isoforms. After assembling the transcripts with StringTie, we used *gffcompare* tool to assess the success of matching the assembled transcript with pre-annotated genes, either fully or partially.

However, the given experiment involved multiple RNA-seq samples generated for two varieties (with two tissue samples) under two different contrasting conditions (salinity treated vs. untreated). Hence, genes and transcripts present in one sample rarely identical with other samples due to varied sequencing depth. So, they need to be assembled in a consistent manner for which the mapping results for individual samples can be compared. For this purpose, we executed the merge function implemented in *StringTie* tool, which prepares a final list of genes by merging all the genes found in any of the samples.

Notations

Let, Y_{ij}: read counts of i^{th} (i = 1, 2, …, N) of gene in j^{th} (j = 1, 2, …, M) sample/library; **Ω**: collection of all genes present in the RNA-seq data (*i.e.* whole gene list); ***G****:* gene set selected from **Ω**; *N:* size of **Ω**; *M*: number of samples/libraries; *n*: size of ***G***; ***Q***: set of associated QTLs; D_i: differential gene expression score for i^{th} gene; T_i be the threshold placed at the i^{th} position in gene ranked list, which divides the gene list into *G* and $G^c = (\boldsymbol{\Omega} - G)$.

Proposed GSQSeq Approach

Earlier developed GSAQ approach was based on the over representation analysis of the QTL hit genes (*i.e*. genes overlapped with QTL regions) in the selected gene set through hypergeometric test [17]. This approach only considered the genes in the selected gene set but ignored their corresponding DE scores. Hence, we developed the GSQSeq approach which is capable of integrating the available DE scores of the selected genes with QTL analysis of the gene set. For this purpose, we developed a scoring function for the gene set, GSQ, that combines features

from over-representation and shifted expression-based approaches [292]. Here, GSQ is computed using hypergeometric distribution based on enrichment score weighted with the DE scores computed through tests such as t-test, FC, *etc*. In other words, GSQ feeds on gene list (preferably ordered based on DE score) along with the corresponding vector of DE scores and classification of genes in input list as *G* and G^c based on a threshold. Then, it calculates the GSQ score, given in Eq. 7.1, for every gene set of the ordered gene list taken at each of the threshold values [128] by using the following procedure.

Then, GSQ uses the following function to calculate the difference between the sum of differential gene expression test scores for *G* and G^c using Eq. 7.1.

$$SD_{GQ} = \sum_{i \in Q}^{i \in G} D_i - \sum_{i \in Q}^{i \in G'} D_i \tag{7.1}$$

This calculation is repeated for each threshold value, T_i.

Therefore, to perform the gene set analysis with the underlying trait specific QTLs for RNA-seq data under a sound computing framework, we developed the GSQSeq approach. In other words, it can be used to evaluate the statistical significance of selected gene sets related to specific trait based on available QTL information. Under GSQSeq approach, the following hypothesises can be constructed for testing purpose.

H_0: Genes in *G* are at most as often overlapped with the QTL regions as the genes in G' (*i.e.* $SD_{GQ} = 0$)

H_1: Genes in *G* are more often overlapped with the QTL regions as compared to genes in G' (*i.e.* $SD_{GQ} > 0$)

The above constructed null hypothesis is a competitive one as it considers the genes from both *G* and G' [5]. In other words, the H_0 tells that the QTL hit gene set members and non-members are distributed randomly across the gene list. Further, the QTL hits of the genes present in *G* can be determined through the indicator function given in Eq. 7.2.

$$I_q(g) = \begin{cases} 1 & if\ g^c[a,b] \in q^c[d,e] \\ 0 & if\ g^c[a,b] \notin q^c[d,e] \end{cases} \tag{7.2}$$

where, $g \in G$; *a* and *b* represent start and stop positions (in terms of base pairs) in chromosome *c* of the gene *g*; $q \in Q$; *d* and *e* represent the start and stop positions (in base pairs) in the chromosome *c* of the QTL *q.*

Further, SD_{GQ} cannot be used for GSA, as it is unstable due to different sizes of gene sets, G and G'. Hence, GSQ uses a Z-score transformation given in Eq. 7.3.

$$GSQ = \frac{SD_{GQ} - E(SD_{GQ})}{\sqrt{V(SD_{GQ})}} \tag{7.3}$$

where, $E(SD_{GQ})$ and $V(SD_{GQ})$ are the expected value and variance of the SD_{GQ} respectively. Further, we obtained the distribution the test statistic, SD_{GQ}, under the H_0, and the expressions for the mean and variance of the test statistic can be obtained and is given in Eq. 7.4 and 7.5 respectively.

$$E(SD_{GQ}) = 2E(X)E(N_{GQ}) - NE(X) \tag{7.4}$$

$$V(SD_{GQ}) = 4\left(\frac{V(X)}{N-1}(E(N_{GQ})(N - E(N_{GQ})) - V(N_{GQ}) + E(X)^2 V(N_{GQ})\right) \tag{7.5}$$

where, *X*: differential gene expression test scores of the genes in the gene set, N_{GQ}: Number of gene set members in *G* got QTL hits, *E(.)*: expected value and $V(.)$: variance.

The N_{GQ} in Eq. 7.4 and 7.5 can be expressed using Eq. 7.2 and given in Eq. 7.6.

$$N_{GQ} = \sum_{q \in Q} \sum_{g \in G} I_q(g) \tag{7.6}$$

The N_{GQ} in Eq. 7.6 follows hypergeometric model and its PMF can be given as:

$$\mathrm{P}\left[N_{GQ} = v\right] = \frac{\binom{V}{v}\binom{N-V}{n-v}}{\binom{N}{n}} \tag{7.7}$$

where, *V*: total number of genes covered by the QTLs in the whole **Ω** and *v*: number of genes in ***G*** that are covered by QTLs. The expected value and variance of N_{GQ}, given in Eq.7.6, can be expressed in Eq. 7.8, and 7.9, respectively.

$$E\left(N_{GQ}\right) = \frac{nV}{N} \tag{7.9}$$

$$V\left(N_{GQ}\right) = \frac{nV(N-V)(N-n)}{(N-1)N^2} \tag{7.10}$$

Under H_0, the GSQ statistic, given in Eq. 7.3, follows a standard normal distribution asymptomatically at least, *i.e.* $GSQ \sim N(0,1)$. Through this property, the statistical significance value for the selected gene set, *G*, was computed. Similarly, this procedure was repeated for all the *K* gene sets obtained through placing the threshold, T_k, (k = 1, 2, …, K) at *K* different places in the ranked gene list. Then, we adjusted the statistical significance values for the gene sets through the multiple hypothesis testing correction, and the procedure is given as follows.

Let, $p_1, p_2, \ldots, p_K$ be the corresponding *p-values* for all the *K* gene sets, and *α* be the level of significance. Here, we assume that all gene sets are equally important for the trait development, hence, we employed Hochberg procedure [172] for correcting the multiple testing, and to compute the adjusted (*adj.*) *p-values* for gene sets. The detail procedure for computation of *adj. p-values* and FDR through Hochberg procedure is given in Chapter 3 (Eq. 3.24). Further, based on the computed *adj. p-values,* the underlying QTLs enrichment significance of the selected gene sets was assessed. In other words, lesser value of *adj. p-value* indicates more QTL enrichment of the selected gene set for the target trait development, and *vice-versa*. The outlines and key analytical steps of the proposed GSQSeq approach are shown in Figure 7.1.

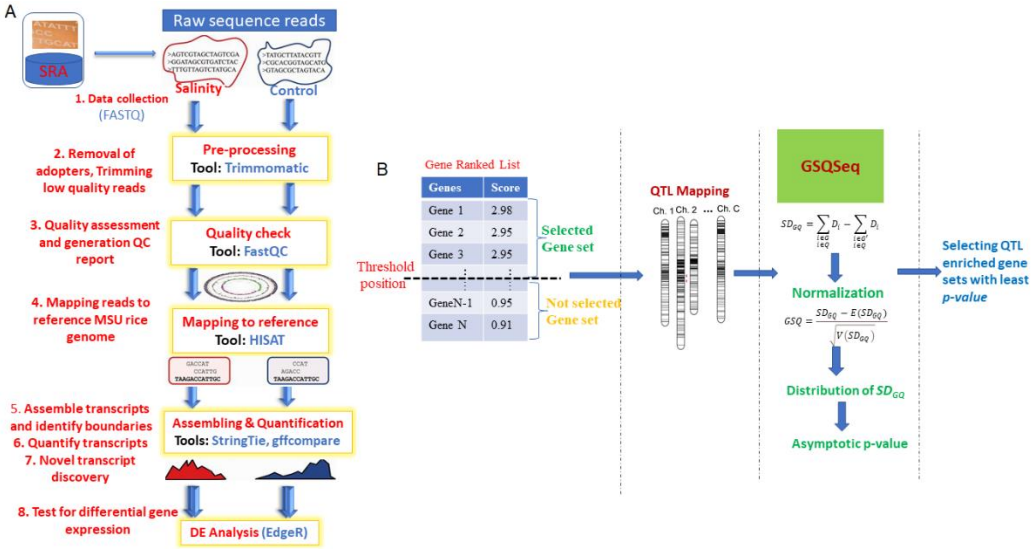


Figure 7.1. Outlines and analytical steps of the GSQSeq approach. (A) Various steps undertaken in RNA-seq data analysis. (B) Analytical steps in GSQSeq approach.

Results and Discussion

From the SRA database of NCBI, a total of 542,309,740 single end reads (with 50 base pair length) were obtained for 24 libraries. Further, the average number of

reads per library was 22,596,239 with a CV 0.169 (=16.9 %). After pre-processing with Trimomatic, the above summary statistic reduced to 22,353,215 as mean library size with CV 0.171 (=17.1%). However, through pre-processing, the average library size was reduced as compared to that of raw sequence datasets. But the variability among the library remained unchanged. Then, most of the pre-processed reads (94.3%) were successfully mapped to the rice reference genome. Of these, 2.87% were mapped more than one positions and were discarded from further analysis.

Genes ranked list preparation

Rice RNA-seq data

The processed sequence count data was used for DE analysis for each samples/library belonging to two contrasting classes, *i.e.* treated *vs.* untreated. The DE analysis was performed through edgeR R package (*v* 3.30.3) implemented in R software (*v*. 4.0.1). The DE test statistic(s) for the genes were computed through LRT statistic(s). Based on the absolute value of the LRT statistic, the genes are arranged in descending order for the preparation of gene ranked list. Then, different values of the thresholds (T_i) are placed on the gene ranked list to select different gene sets. Through this process, gene sets of sizes such as 200, 300, 400, …, 2000 are selected from the ranked gene list.

Rice Microarray data

The raw CEL files of these collected samples for the cold, drought, fungal and insect stresses were processed using RMA algorithm available in *affy* Bioconductor package of R. This includes background correction, quantile

normalization and summarization by the median polish approach. The log2 scale transformed expression data from the RMA for these selected experimental samples were used for the preparation of the gene ranked list through DE analysis. Here, the DE analysis was performed through t-test and the test statistic(s) for the genes computed through the t-test. The genes are arranged in descending order for the preparation of gene ranked list. Then, different values of the T_i are placed in various positions on the gene ranked list to select different gene sets. Through this process, gene sets of sizes such as 200, 300, 400, …, 2000 are selected from the ranked gene lists for each dataset.

Distribution of GSQ statistic

The distribution of the NQhits statistic computed through existing GSAQ approach (given in Chapter 4) over different selected gene sets for the different stresses are shown in Figure 7.2A. Further, the distribution of the GSQ statistic(s) computed from the GSQSeq approach is also shown on Figure 7.2B.

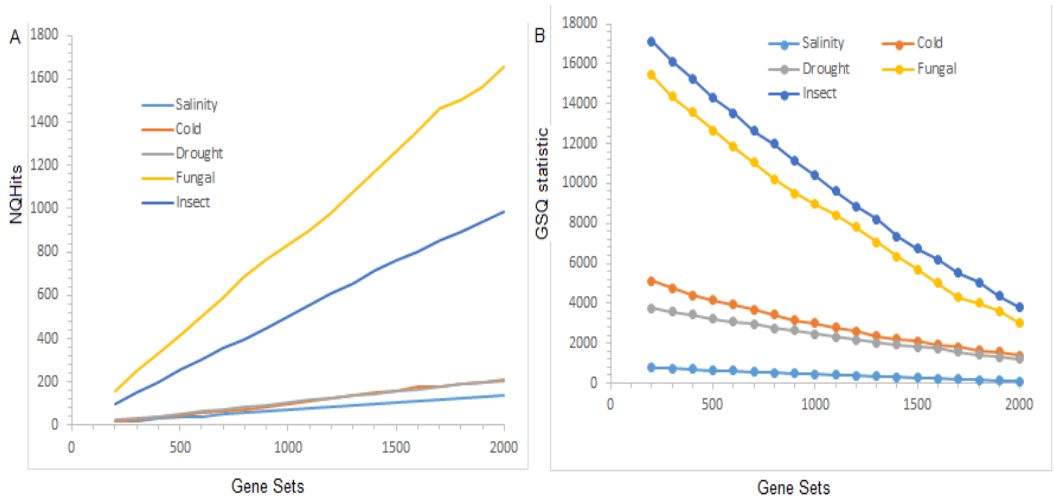


Figure 7.2. Distribution of NQhits and GSQ test statistics(s). The distribution of NQhits (A) and GSQ (B) test statistic(s) over the selected gene sets

The distribution of the NQhits computed from the GSAQ approach indicated that the values of the NQhits statistic(s) was found to be higher for fungal stress followed by insect stress as compared to other datasets (Figure 7.2A). This is due to the fact that the higher number (76) of QTLs are reported for this stress followed by 57 in bacterial stress. In other words, the NQhits is a linear function of the number of genes present in gene sets, number of QTLs reported for that stress and length of the QTL regions (Figure 7.2A). Similar interpretations can be for the distribution of the GSQ statistic(s) (Figure 7.2B). However, the NQhits statistic did not consider the DE scores of the genes present in the gene set. Here, it is worthy to note that the GSQ is a function of the number of genes along with their respective DE scores in the gene set, number of QTLs reported for that stress and length of the QTL regions

Proposed approach for Gene set analysis with QTLs

The NQhits and GSQ statistics failed to tell the trait specific enrichment of gene sets or association of genotype-phenotype relation. Therefore, we proposed GSQSeq approach to test the trait specific enrichments of the gene sets with underlying QTLs. Further, we explored the ability of the proposed GSQSeq approach along with existing GSVQ and GSAQ approaches to provide biologically meaningful insights (*e.g.*, establishing genotype-trait specific phenotype associations) in the real high-throughput GE datasets derived from RNA-seq and Microarrays studies. For all the three tested GSA approaches, we searched significantly associated gene sets enriched with underlying QTLs, which were selected by a particular gene selection method (*e.g.*, t-test in Microarrays, edgeR

in RNA-seq) in each of the datasets. The results from such analysis are shown in Tables 7.1, 7.2.

Table 7.1. Performance analysis of GSQSeq approach on real Microarray datasets.

	Drought			Fungal			Insect		
GS	GSVQ	GSAQ	GSQSeq	GSVQ	GSAQ	GSQSeq	GSVQ	GSAQ	GSQSeq
200	0.83	11.80	276.94	0.00	0.00	254.48	0.65	6.36	229.17
300	0.64	9.29	252.76	0.31	0.87	220.28	0.94	13.44	234.13
400	0.45	1.86	254.48	0.06	0.00	202.94	0.69	8.45	236.10
500	0.81	11.66	252.28	0.22	0.02	190.28	1.08	18.20	228.86
600	0.79	12.97	252.16	0.23	0.00	181.58	0.92	13.93	228.69
700	0.49	2.75	252.06	0.34	0.62	219.58	1.23	21.99	228.56
800	0.78	11.56	251.98	1.04	10.44	219.50	0.81	12.97	228.47
900	0.70	10.90	251.92	0.98	5.69	219.43	1.09	17.49	228.39
1000	0.93	14.56	251.86	0.14	0.16	219.38	1.09	15.47	228.32
1100	1.01	15.11	251.81	0.00	0.00	219.33	1.43	23.98	228.26
1200	1.08	19.30	251.76	0.00	0.00	219.28	1.62	25.14	228.21
1300	1.00	18.04	293.50	0.04	0.00	219.24	1.43	24.35	228.17
1400	0.93	15.60	276.15	0.19	0.37	219.20	2.28	28.35	228.12
1500	1.01	13.95	252.76	0.34	1.22	219.17	1.99	27.67	228.09
1600	0.69	9.23	254.48	0.94	6.27	219.13	1.50	21.05	228.05
1700	1.01	14.48	220.28	2.29	19.08	219.10	1.62	21.52	228.02
1800	1.09	16.36	202.94	0.14	0.03	219.08	1.04	16.17	227.99
1900	1.10	17.91	190.28	0.00	0.00	219.05	1.19	20.39	227.96
2000	1.04	19.27	181.58	0.01	0.00	219.02	1.08	13.90	227.94

GSAQ: Gene Set Analysis with QTL; GSVQ: Gene Set Validation test with QTL; GSQSeq: Gene Set Analysis with QTL for RNA-seq

Table 7.2. Performance analysis of GSQSeq approach on RNA-seq and Microarray datasets based on *-log$_{10}$(p-values*).

	Salinity			Cold		
Gene sets	GSVQ	GSAQ	GSQSeq	GSVQ	GSAQ	GSQSeq
200	1.05	18.85	222.71	0.05	1.00	222.46
300	0.67	8.29	212.71	0.01	2.00	223.02
400	0.69	8.02	211.71	0.08	2.00	224.30
500	0.82	12.99	302.23	0.02	2.00	225.83
600	0.62	5.85	270.66	0.02	2.00	226.60
700	0.75	10.74	226.11	0.01	1.30	227.13
800	0.77	13.23	197.34	0.00	1.30	228.46
900	0.89	14.98	168.77	0.00	1.30	229.41
1000	0.73	9.76	159.48	0.01	1.30	230.45

1100	0.85	14.63	138.47	0.02	1.30	231.46
1200	0.71	11.52	132.49	0.03	1.30	302.59
1300	0.91	17.85	112.50	0.02	1.30	249.00
1400	0.85	14.15	105.22	0.02	0.00	236.69
1500	0.72	8.29	101.95	0.03	1.19	232.29
1600	0.83	11.51	89.95	0.04	1.30	219.10
1700	0.71	8.09	87.65	0.02	1.46	213.22
1800	0.82	10.50	77.54	0.01	1.70	196.50
1900	0.93	14.24	68.47	0.01	2.30	203.26
2000	0.81	10.89	67.15	0.00	2.00	180.41

GSAQ: Gene Set Analysis with QTL; GSVQ: Gene Set Validation test with QTL; GSQSeq: Gene Set Analysis with QTL for RNA-seq

For salinity stress RNA-seq data, the magnitude of −log10 (*p-values*) from GSQSeq was found to be much higher than that of existing GSVQ and GSAQ approaches (Table 7.2). This indicated that, GSQSeq approach more often rejected H_0 (*i.e*. equal salinity QTL enrichment of both selected and not selected gene sets) as compared to GSVQ and GSAQ approaches. Therefore, it was found that salinity trait specific analysis of gene sets derived from RNA-seq study was successful through GSQSeq as compared to GSVQ and GSAQ. In other words, GSQSeq approach performed better in terms of detecting the QTL enriched gene sets compared to the existing approaches. In order to cross validate these findings on the same RNA-seq data related to salinity stress, we computed FDR for the GSQSeq, GSAQ and GSVQ approaches for all the gene sets. The results are given in Tables 7.3 and 7.4. It was observed that the value of FDR from proposed GSQSeq approach for all these gene sets are far below than that of existing GSAQ and GSVQ approaches (Table 7.3). Therefore, it can be inferred that the proposed GSQSeq is more robust than the GSAQ and GSVQ for performing gene set enrichment testing with salinity trait specific QTLs.

For cold stress data derived from Microarrays, the values of −log10 (*p-values*) from GSQSeq was observed to be higher than that of existing GSVQ and GSAQ approaches over all the selected gene sets (Table 7.1). This indicated that, GSQSeq approach more often rejected H_0 (*i.e*. equal salinity QTL enrichment of both selected and not selected gene sets) as compared to GSVQ and GSAQ approaches. Further, the FDR values computed through the proposed GSQSeq approach for all these selected gene sets were found to be least followed by GSAQ than GSVQ approach (Table 7.4). Similar findings were observed for drought, fungal and insect stress datasets in rice (Tables 7.1-7.4). Therefore, it can be concluded that the proposed GSQSeq is much better and robust than the GSAQ and GSVQ for performing gene set enrichment testing with the underlying QTLs for Microarrays based GE study. Furthermore, across all the considered datasets we found much greater consistency in QTL specific gene set enrichment analysis across five different stress scenarios, *viz*. salinity, cold, drought, fungal and insect, by using GSQSeq than GSVQ and GSAQ (Tables 7.1-7.4).

Table 7.3. FDR based analysis of GSA approaches on RNA-seq and Microarray datasets.

	Salinity			Cold			
Gene sets	GSVQ	GSAQ	GSQSeq	GSVQ	GSAQ	GSQSeq	GSVQ
200	0.224	2.69E-18	6.99E-160	0.995	0.10	1.58E-249	0.226
300	0.224	6.14E-09	6.50E-139	0.995	0.50	2.96E-237	0.254
400	0.224	9.97E-09	5.63E-133	0.995	0.68	6.90E-233	0.352
500	0.224	2.45E-13	2.82E-302	0.995	0.46	9.96E-220	0.226
600	0.238	1.40E-06	8.31E-271	0.995	0.23	7.09E-214	0.226
700	0.224	2.86E-11	2.46E-226	0.995	0.01	3.32E-197	0.344
800	0.224	1.61E-13	1.23E-197	0.995	0.18	1.78E-236	0.226
900	0.224	6.67E-15	4.08E-169	0.995	0.14	2.07E-236	0.241
1000	0.224	2.37E-10	6.99E-160	0.995	0.10	2.37E-236	0.223
1100	0.224	1.12E-14	6.50E-139	0.995	0.06	2.67E-236	0.223
1200	0.224	5.87E-12	5.63E-133	0.995	0.02	4.47E-303	0.223

1300	0.224	1.35E-17	4.98E-113	0.995	0.02	1.58E-249	0.223
1400	0.224	2.25E-14	8.88E-106	0.995	1	2.96E-237	0.223
1500	0.224	6.14E-09	1.53E-102	0.995	1	6.90E-233	0.223
1600	0.224	5.87E-12	1.43E-90	0.995	1	9.96E-220	0.241
1700	0.224	9.00E-09	2.63E-88	0.995	1	7.09E-214	0.223
1800	0.224	4.59E-11	3.22E-78	0.995	1	3.32E-197	0.223
1900	0.224	2.19E-14	3.60E-69	0.995	1	6.10E-204	0.223
2000	0.224	2.24E-11	7.10E-68	0.995	1	3.92E-181	0.223

GSAQ: Gene Set Analysis with QTL; GSVQ: Gene Set Validation test with QTL; GSQSeq: Gene Set Analysis with QTL for RNA-seq

Table 7.4. FDR based analysis of the GSA approaches on real Microarray datasets.

	Drought			**Fungal**			**Insect**	
GS	GSAQ	GSQSeq	GSVQ	GSAQ	GSQSeq	GSVQ	GSAQ	GSQSeq
200	2.52E-12	1.04E-276	0.996	1	6.24E-221	0.224	4.33E-07	5.1E-230
300	6.14E-10	2.20E-253	0.996	0.42	8.79E-253	0.141	4.36E-14	3E-225
400	0.013834	4.47E-255	0.996	1.00	1.10E-252	0.215	3.77E-09	1.8E-220
500	3.21E-12	6.24E-221	0.996	1.00	6.59E-249	0.121	1.19E-18	1.1E-215
600	1.83E-13	8.79E-253	0.996	1.00	3.95E-245	0.141	1.58E-14	6.2E-211
700	0.00188	1.10E-252	0.996	0.65	2.37E-241	0.121	3.24E-22	3.7E-206
800	3.73E-12	1.32E-252	0.549	3.42E-10	1.42E-237	0.172	1.19E-13	2.2E-201
900	1.61E-11	1.54E-252	0.549	9.63E-06	8.54E-234	0.121	5.55E-18	1.3E-196
1000	6.61E-15	1.76E-252	0.996	1	5.12E-230	0.121	4.98E-16	7.5E-192
1100	2.12E-15	1.98E-252	0.996	1	3.07E-226	0.101	4.01E-24	4.4E-187
1200	5.11E-19	2.20E-252	0.996	1	1.84E-222	0.101	4.56E-25	2.6E-182
1300	5.74E-18	4.96E-294	0.996	1	1.11E-218	0.101	2.13E-24	1.5E-177
1400	7.92E-16	1.04E-276	0.996	1	6.64E-215	0.098	8.54E-28	9E-173
1500	2.14E-14	2.20E-253	0.996	0.226411	3.98E-211	0.098	2.04E-27	5.3E-168
1600	6.54E-10	4.47E-255	0.549	3.44E-06	2.39E-207	0.101	2.13E-21	3.1E-163
1700	6.94E-15	6.24E-221	0.097	1.57E-18	1.43E-203	0.101	8.14E-22	1.9E-158
1800	1.65E-16	1.30E-203	0.996	0.150939	8.61E-200	0.123	1.06E-16	1.1E-153
1900	5.89E-18	5.57E-191	0.996	2.64E-01	5.16E-196	0.121	8.51E-21	6.4E-149
2000	5.11E-19	2.65E-182	0.996	3.77E-01	3.10E-192	0.121	1.58E-14	3.8E-144

The proposed GSQSeq approach is an improved way to perform the trait specific analysis of gene sets to establish genotype (polygenes)-phenotype (quantitative trait) association testing with the help of genetically rich QTL data. Further, it is more biologically appealing to establish association of genes

(genotype) in the selected gene set with underlying QTLs (traits/phenotypes). However, in the existing GSVQ and GSAQ approaches, the genes in gene sets are taken as input to the hypergeometric distribution for performing trait enrichment analysis. These approaches violate the basic assumptions (*i.e*. sampling units must be drawn without replacement) and did not consider the DE scores of the genes present in gene set. Thus, they expected to have poor performance in terms of gene set enrichment. Hence, the proposed GSQSeq approach was found to be more successful and effective to detect trait specific QTLs enriched gene sets than the existing approaches.

The proposed GSQSeq approach allowed one to statistically test the gene set for enrichment with the underlying QTLs (*i.e*. rejection of null hypothesis of random association of selected genes with QTLs). Further, a *p-value* was assigned to each selected gene set, which is more scientific and statistically meaningful to genome researchers and experimental biologists (as value lies between 0 and 1). The gene sets with lower *p-values* are considered as more enriched with the underlying trait specific QTLs and *vice-versa*. It may be noted that the proposed GSQSeq technique is a two-stage approach. First, it deals with the selection of gene sets through DE analysis of large GE data. Second, it assesses the QTL enrichment significance of gene sets by using a developed parametric testing procedure. This analysis eases the interpretation of a large-scale experiment by identifying trait specific enriched gene sets. Here, rather than focusing on individual QTL hit genes, researchers can focus on gene sets

(polygenes), which tend to be more reproducible and more interpretable (for quantitative traits).

The proposed GSQSeq can be considered as a valuable tool for performing gene(s) enrichment analysis in molecular plant breeding context. Further, it provides a valuable tool for integrating the GE data from scRNA-seq or bulk RNA-seq or Microarrays with genetically rich QTL data to identify potential QTL enriched gene sets or set of QTL candidate genes, which may act as valuable input or hypothesis for the plant breeders for designing breeding experiments. Due to the unavailability of scRNA-seq datasets for crops, we are unable to test the performance of GSQSeq approach on rice scRNA-seq datasets, which will be done in future.

"Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis …"

R A Fisher

CHAPTER 8

DEVELOPED SOFTWARE PACKAGES

Gene Selection using 'BSM' R Software Package

Selection of biologically relevant genes from high dimensional expression data is a key research problem in gene expression genomics. Most of the available gene selection tools are either based on relevancy or redundancy measure, which are usually adjudged through post selection classification accuracy. Through these tools, the ranking of genes was done on a single high-dimensional expression data, which leads to the selection of spuriously associated and redundant genes. Therefore, in Chapter 3, we developed a BSM statistical approach through combining SVM wrapper with MRMR under a sound statistical setup for the selection of biologically relevant genes. Here, the genes are selected through statistical significance values computed using a NP test statistic under a bootstrap based subject sampling model. Based on this, we developed an R software package which includes BSM R package and accompanying documentation with examples. This package is available at https://github.com/sam-uofl/BSM. This software is capable of computing weights for gene selection through MRMR and SVM, SVM-MRMR, and also provide functions for computing *p-values,* and adjusted *p-values* through BSM approach for different parameter options. Further,

it also allowed different functions for selecting relevant gene sets through existing MRMR, SVM, SVM-MRMR, and proposed BSM gene selection approaches.

Gene Set Analysis with 'GSAQ' R software package

The analysis of gene sets is usually carried out based on gene ontology terms and known biological pathways. These approaches may not establish any formal relation between genotype and trait specific phenotype. Therefore, in Chapter 4, we proposed an innovative GSAQ statistical approach for interpreting expression data in context of gene sets with traits. To facilitate the use of the proposed approach, we have developed resources that are freely available from the CRAN site of R. This resource includes the GSAQ R package, accompanying documentation and model real data examples. This package can be freely downloaded from https://cran.r-project.org/web/packages/GSAQ. This software is capable of computing *NQhits* statistic and performing QTL specific gene set enrichment analysis through the proposed GSAQ and existing GSVQ approaches. Besides, it can also be used for selection of relevant gene sets from high-dimensional GE data through different gene selection methods, obtaining QTL candidate genes and getting chromosome and QTL wise distributions of genes in selected gene set.

Analysis of scRNA-seq data using 'SwarnSeq' R Package

scRNA-seq is gradually replacing bulk RNA-seq and Microarrays for high-throughput studies of gene expression. The DE analysis is the major downstream analysis of scRNA-seq data, used to detect DE genes to gain insights into the underlying complex biological processes. The DE analysis in presence of noises

from biological (*e.g.,* stochasticity of gene expression, heterogeneous cell types, cell cycle) and technical sources (*e.g.,* dropout events, zero-inflation, low input mRNA molecules, low cell capture rates, amplification bias) remain a key challenge in scRNA-seq. So, in Chapter 6 we present a novel statistical approach for DE, and other downstream analysis that considers the molecular capture process in scRNA-seq data modeling. Our novel approach is implemented in an R software package, namely, SwarnSeq. The developed R package can be availed at https://github.com/sam-uofl/SwarnSeq. Further, our SwarnSeq R package provides *OptimCluster* function for getting the optimum number of cell clusters from scRNA-seq count data. Additionally, it also provides option for estimation of capture rates of cells using different methods, *e.g.* MLE, regression, *etc.*, whether RNA spike-in data is available or not. The function SwarnSeq implemented in SwarnSeq R package can be executed for estimating the parameters for each gene, *i.e.* mean, dispersion, zero inflation, effects of groups, cell clusters and cell level auxiliary information on zero-inflation as well as means of non-zero counts. *SwarnSeqLRT* function provides option for results from DE analysis and DZI analysis, when the observed UMI counts are adjusted for molecular capture rates. Moreover, functions like *SwarnUnadjSeq* and *SwarnUnadjLRT* are implemented for parameter estimation and DE, and DZI analyses respectively, when the users do not need to adjust the count data for capture efficiency. The top influential genes detected through SwarnSeq approach can be selected and classified through the implemented *SwarnClass* and *SwarnTopTags* functions, respectively. Different

options are provided in the SwarnSeq R package for adjusting the capturing process and correcting amplification bias through different normalization methods.

GSA for RNA-seq/scRNA-seq study using 'GSQSeq' R Package

In chapter 7, we presented a statistical approach for performing gene set analysis with QTLs for RNA-seq/scRNA-seq data. This approach considers the genes present in the gene set along with their corresponding DE scores to analyze in presence of the trait specific QTL data. Here, the enrichment significance of the gene sets is assessed through the *p-values* computed using the developed test statistic(s). Based on this developed approach, we developed 'GSQSeq' R software package. This is available at https://github.com/sam-uofl/GSQSeq . This is capable of computing GSQ scores for gene sets, and statistical significance values for QTL enrichment of the sets. Further, it also provides function to perform the enrichment analysis of the gene sets with QTL through the existing GSVQ.

The pre-processed scRNA-seq datasets, R codes and reference genes used in the comparative study (Chapter 5) are available in the RoopSeq GitHub project directory at https://github.com/sam-uofl/RoopSeq.

"In the next 10 years, data science and software will do more for medicine than all of the biological sciences together…"

Vinod Khosla

CHAPTER 9

GENERAL DISCUSSION AND CONCLUSION

In the last 15 years since its inception, GSA has become an extremely popular approach for secondary analysis of high-throughput genomic data obtained Microarrays, RNA-seq and scRNA-seq studies. It has been successfully used to gain biological insights into the etiology of various complex diseases in humans as well as other model organisms, including mammals and other cellular organisms. GSA has immense benefits in terms of biological interpretation of results as well as numerous computational advantages over single gene studies [52]. It also enhances biologically meaningful interpretation of results and reproducibility of important gene lists yielded by independent studies [12,15,17,26,67]. In other words, the cumulative effects of the genes distributed in a gene set is considered in a single analysis, and has more statistical power as compared to the univariate counterparts [26]. Despite the wider usefulness of GSA, there are limited number of studies found in the literature, which consider the wider gamut of high throughput genomic studies. Hence, we have presented a detailed overview of GSA in high-throughput genomic studies in Chapter 2. This also summarized the commonalities of GSA approaches used in key genomic studies in terms of their execution, underlying null hypotheses, nature of test statistic, sampling models, common execution, and analytical steps, *etc*.

Over the years, a diverse set of methods for performing GSA has been proposed for Microarrays, RNA-seq and GWAS data analysis and the increased application of these methods has exposed several factors that affect the interpretations of GSA results. These factors include the null hypothesis being tested, the underlying sampling/permutation procedure, and the nature and distribution of test statistic(s). All of these factors play a significant role for choosing proper GSA for the data analysis. Researchers have also identified a variety of circumstances that can lead to faulty findings; hence, proper care is suggested to avoid misleading results. Several individual studies have been conducted over time to summarize GSA approaches for each type of genomic study. In Chapter 2, we summarize a comprehensive review of GSA approaches in terms of statistical structure, execution, and classification for three different high-throughput genomic studies. Several approaches and tools have evolved over time, individually for each type of genomic study. Thus, instead of individually reviewing them, we present the classification of GSA approaches for Microarrays, RNA-seq and GWAS into different generations along with underlying statistical methodologies/tests, and special features. Many earlier reviews of GSA are data independent studies [5,9,15], but our study is data dependent and comprehensive.

This study presented in Chapter 2 will serve as a catalogue and provide guidelines to genome researchers and experimental biologists for choosing the proper GSA based on several factors. Here, we reported several challenges, which need to be addressed by statisticians and biologists collectively to develop the next generation of GSA approaches. These new approaches will be able to analyze

high-throughput data more efficiently in order to better understand the biological systems, and to increase the specificity, sensitivity, utility, and relevance of GSA.

In GE genomics, the key step is to select relevant genes or gene sets for performing GSA. Further, the selected genes which can be used as predictors for the development of statistical/classification models to handle high dimensionality in GE data. Therefore, in Chapter 3, we present an improved statistical approach for gene selection from high-dimensional GE data, which is both effective in reducing redundancy among the genes and improves biological relevancy of genes with the target trait. Here, the genes are selected based on the assessment of the statistical significance of the self-contained null hypothesis under a sound computational framework. Usually, thousand(s) of null hypotheses are usually tested simultaneously in GE data analysis which increased the chance of selection of false positive genes. Hence, through the proposed BSM approach an adjusted *p-value* was assigned to each gene after multiple test adjustments, and relevant genes were selected based on the adjusted *p-values*. The BSM approach is based on the NP test statistic(s) which does not depend on the distribution of the GE data unlike t-test. Further, the bootstrap procedure in the BSM can minimize the redundancy among genes as well as reduce the spurious association of genes with traits during gene selection. The proposed approach is also less computationally expensive compared to SVM-RFE, and SVM-MRMR and can be implemented on a personal or workstation computer for analyzing large GE datasets. Furthermore, we used a comprehensive framework of performance analysis of the gene selection methods under statistical necessary, and biological

relevant criteria. More specifically the tested gene selection methods include SVM-RFE from Wrapper, SVM-MRMR and proposed BSM from Hybrids (Embedded) and the remaining 7 from the filter categories. The comparative analysis revealed the proposed approach has the features of an ideal technique of gene selection, as it performed better under both statistically necessary, and biologically relevant criteria. Moreover, this study provides a systematic and rigorous evaluation of the gene selection methods under a multi-criteria decision setup on multiple real datasets. It will also provide a framework to the researchers to comparatively study the available methods, which will guide genome researchers and experimental biologists to select the best method(s) objectively. The proposed approach may provide a statistical template for combing other filter and wrapper gene selection methods under a sound, and effective computational environment.

After obtaining the gene sets from high-throughput expression data, it is necessary to perform GSA with genetically rich QTL data to establish the genotype-phenotypes link. So, in Chapter 5, an innovative statistical approach for analyzing gene sets with QTL is presented. The proposed GSAQ approach is a new way to perform the enrichment analysis of gene sets to establish genotype (polygenes)-phenotype (quantitative trait) association testing with the help of genetically rich trait specific loci data. Further, it is more biologically appealing to establish association of genes (genotype) in the selected gene set with underlying QTLs (traits/phenotypes). The GSAQ approach has number of unique features, (i) eases the interpretation of a large scale experiment by identifying trait specific enriched gene set; (ii) provides a statistically sound a framework for performing

GSA with QTLS, as it is based on a competitive null hypothesis and gene sampling model; (iii) helps in prioritizing QTL candidate genes or QTL enriched gene sets under a sound computational setup, which would be very helpful in unraveling genotype-to-phenotype relationships; (iv) provides biologically relevant criteria for performance analysis of gene selection methods.

The proposed GSAQ approach can be considered as a valuable tool for performing gene(s) enrichment analysis in plant breeding context. Further, the GSAQ approach provides a valuable platform for integrating the GE data with genetically rich QTL data to identify potential QTL enriched gene sets or set of QTL candidate genes, which may act as valuable input or hypothesis for the plant breeders for designing breeding experiments. In Chapter 4, we have statistically established the credibility of the proposed method (GSAQ) by comparing its performance with the only existing approach (GSVQ) through a statistically strong criterion, *i.e.* FDR, in five different stress scenarios in rice. But, in case of crop biotechnology and breeding, very little amount of work has been done to confirm these results. However, these results can provide guidelines to the biotechnologists and breeders to validate the *in-silico* results in a wet lab condition.

RNA-seq and scRNA-seq have completely overtaken the Microarrays to study the expression dynamics of genes at the tissue and individual cell resolution level, respectively. Therefore, the GSA approaches need to be extended to these studies. More specifically, scRNA-seq is a rapidly growing field in gene expression genomics, and DE is a popular downstream analysis performed on such data. So, newer and better methods were or being introduced over the years in the literature,

which greatly vary based on their utility, basic statistical concepts, models fitted to the data, the test statistic(s) used, *etc.* It is pertinent for the users to be updated on the recent development, the current status of the available methods, and further to evaluate and choose the best method for their real data applications. Under these considerations, we presented a comprehensive study of the available DE methods for scRNA-seq data analysis in Chapter 5. Instead of individually reviewing them, we introduced the classification of the available methods, along with their unique features and limitations. Further, in our comparative study, we have performed a systematic comparison of popular methods/tools extensively used for DE analysis of scRNA-seq data, which broadly covers all the classes of the DE methods. These methods include seven from dedicated single-cell methods, four from the general category, and the remaining methods from bulk RNA-seq. In Chapter 5, we focus on the most straightforward experimental design (*i.e.,* comparing two cell groups), but many real studies require more complex structures, which not all of the tested methods can accommodate. The main strengths of our comparative study include (*i*) use of multiple real scRNA-seq datasets with different cell sizes to capture true distributional nature and diversity of single-cell data; (*ii*) assessment of the methods based on the individual-centric performance metrics; (*iii*) performance analysis of techniques based on multi-criteria setup; (*iv*) combined data analysis through TOPSIS approach; (*v*) similarity analysis of the tested methods. Under the individual performance metric centric evaluations, it is not possible to find the globally best performing option for DE analysis of scRNA-seq data, as particular metric provides different results for different data. However, their performances are

data-dependent and mostly positively related to the total number of cells and cells per group in the data. Further, the tested methods' performance mostly depends on the type of test statistic(s) employed to compute the DE statistic for genes. To search for the best option for DE analysis in scRNA-seq, we first used the TOPSIS method under the MCDM setup and found that different methods performed well for different datasets. Moreover, our integrated data analysis through the TOPSIS technique ably revealed the consistently best practices for DE analysis of scRNA-seq data irrespective of the evaluation criteria. The crucial conclusions from our work that was overlooked in all former studies can be summarized as (*i*) bulk RNA-seq DE methods are competitive and even better than most of the single-cell methods; (*ii*) possible to find the globally best method through combined data analysis; (*iii*) there exist similarities among the performance of the DE methods. In the future, the researchers may consider carrying out an extensive comparison of methods for DE analysis of scRNA-seq data under more complex experimental designs. This study will serve as a catalog and provide guidelines to genome researchers and experimental biologists to choose the best option objectively. In this chapter, we reported the existing limitations of the available methods which need to be addressed by statisticians and biologists collectively to develop innovative and efficient approaches. These new approaches will be able to analyze UMI data more efficiently to better understand the biological systems and increase the specificity, sensitivity, utility, and relevance of single-cell studies.

As DE analysis is a key process in GSA, so, we present an improved and novel statistical approach for analysis of scRNA-seq counts data in Chapter 6. This

approach can perform analysis including DE, DZI, classification of genes, estimation of cell capture rates, and determination of optimum number of cell clusters with strong statistical basis. Here, we provided all the background statistical theory, data example, preliminary data and real experimental data analysis results for our SwarnSeq model. The benchmarking of the SwarnSeq method on multiple real datasets over a wide range of statistical criteria indicated its better performance over the existing methods. Further, the SwarnSeq method will surely help the experimental biologist and genome researchers to identify true DE genes for their experiments. Moreover, our comparison framework may be adopted for further comparative study of scRNA-seq DE tools. In future, parameter estimation procedure, like Empirical Bayes shrinkage method can be implemented in the SwarnSeq to estimate the gene specific dispersion, and that will enhance its performance. The SwarnSeq assumes the factors, such as cellular populations, cell clusters and other co-variates, have fixed effects on means and zero inflations. This assumption may be unrealistic from a biological standpoint (some may have random effects). Therefore, researchers may think of random or mixed effect models in SwarnSeq in the future to improve its performance.

After selection of the DE genes, the gene sets derived from RNA-seq studies are analyzed with QTL data. So, we present a statistical method for GSA of RNA-seq/scRNA-seq data in Chapter 7. This approach considers the genes present in the gene set along with their corresponding scores to analyze in presence of the trait specific QTL data. Here, the enrichment significance of the gene sets is assessed through the adjusted p-values computed using the

developed test statistic(s). GSAQ approach presented in Chapter 4 has some serious limitations, such as only consider the genes which overlapped with the QTL regions, but failed to consider their corresponding DE scores, treats each gene equally by assuming each gene as independently and identically distributed which is contrary to the real biology. Further, it uses only the most significant genes, while discards other genes. Unlike GSAQ, GSQSeq considers the significant genes along with their DE scores for performing GSA in RNA-seq data. This technique performs better than its predecessor to perform GSA in high-throughput genomic studies. Such concepts are very useful in establishing links of gene sets with the underlying trait/phenotypes in plant and complex disease biology, as most of the traits are quantitative in nature and controlled by polygenes. In future attempts may be made by the computational biologists and bioinformaticians to develop next generations of (Topology based) GSA with QTLs approaches using Graph/network theory, multivariate and regression analytical techniques. Besides, the limitations and shortcomings of the available GSA and DE methods, reported in various Chapters 2-6, need to be addressed by statisticians and biologists collectively to develop efficient approaches. These new approaches will be able to analyze high-throughput genomic data more efficiently to better understand the biological systems and increase the specificity, sensitivity, utility, and relevance of high-throughput genomic studies, such as RNA-seq, scRNA-seq.

"If you cannot do bioinformatics, then you may not do or understand the biology…"

Anonymous

REFERENCES

1. Marx V. The big challenges of big data. Nature. 2013;498: 255–260. doi:10.1038/498255a
2. Wang J, Chen L, Wang Y, Zhang J, Liang Y, Xu D. A Computational Systems Biology Study for Understanding Salt Tolerance Mechanism in Rice. Xu Y, editor. PLoS One. 2013;8: e64929. doi:10.1371/journal.pone.0064929
3. Das S, Meher PK, Rai A, Bhar LM, Mandal BN. Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (*Glycine max* L.). PLoS One. 2017;12: e0169605. doi:10.1371/journal.pone.0169605
4. Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D. Prediction of Drought-Resistant Genes in Arabidopsis thaliana Using SVM-RFE. Zhu D, editor. PLoS One. 2011;6: e21750. doi:10.1371/journal.pone.0021750
5. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23: 980–987. doi:10.1093/bioinformatics/btm051
6. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biology. 2003. doi:10.1186/gb-2003-4-4-210
7. Kuo L, Yu F, Zhao Y. *Statistical Methods for Identifying Differentially Expressed Genes in Replicated Microarray Experiments: A Review. Statistical Advances in the Biomedical Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. pp. 341–363. doi:10.1002/9780470181218.
8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102: 15545–15550. doi:10.1073/pnas.0506580102
9. Fridley BL, Patch C. Gene set analysis of SNP data: benefits, challenges, and future directions. Eur J Hum Genet. 2011;19: 837–843. doi:10.1038/ejhg.2011.57
10. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics. 2007;23: 3251–3253. doi:10.1093/bioinformatics/btm369
11. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34: 267–273. doi:10.1038/ng1180
12. de Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties

of gene-set analysis. Nat Rev Genet. 2016;17: 353–364. doi:10.1038/nrg.2016.29
13. Mooney MA, Wilmot B. Gene set analysis: A step-by-step guide. Am J Med Genet Part B Neuropsychiatr Genet. 2015. doi:10.1002/ajmg.b.32328
14. Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. Genomics. 2011;98: 1–8. doi:10.1016/j.ygeno.2011.04.006
15. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. Brief Bioinform. 2016;17: 393–407. doi:10.1093/bib/bbv069
16. Goeman JJ, Van de Geer S, De Kort F, van Houwellingen HC. A global test for groups fo genes: Testing association with a clinical outcome. Bioinformatics. 2004. doi:10.1093/bioinformatics/btg382
17. Das S, Rai A, Mishra DC, Rai SN. Statistical Approach for Gene Set Analysis with Trait Specific Quantitative Trait Loci. Sci Rep. 2018;8: 2391. doi:10.1038/s41598-018-19736-w
18. Das S, Rai A, Mishra DC, Rai SN. Statistical approach for selection of biologically informative genes. Gene. 2018;655. doi:10.1016/j.gene.2018.02.044
19. Wang X, Cairns MJ. Gene set enrichment analysis of RNA-Seq data: integrating differential expression and splicing. BMC Bioinformatics. 2013;14: S16. doi:10.1186/1471-2105-14-S5-S16
20. Rahmatallah Y, Zybailov B, Emmert-Streib F, Glazko G. GSAR: Bioconductor package for Gene Set analysis in R. BMC Bioinformatics. 2017. doi:10.1186/s12859-017-1482-6
21. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. Ouzounis CA, editor. PLoS Comput Biol. 2012;8: e1002375. doi:10.1371/journal.pcbi.1002375
22. Khatri P, Draghici S, Ostermeier GC, Krawetz SA. Profiling Gene Expression Using Onto-Express. Genomics. 2002;79: 266–270. doi:10.1006/geno.2002.6698
23. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003.
24. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. AgriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017. doi:10.1093/nar/gkx382
25. Beibarth T, Speed TP. GOstat: Find statistically overrepresented Gene Ontologies with a group of genes. Bioinformatics. 2004. doi:10.1093/bioinformatics/bth088
26. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007;1: 107–129. doi:10.1214/07-AOAS101
27. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex. Neurochem Res. 2004;29: 1213–1222. doi:10.1023/B:NERE.0000023608.29741.45

28. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics. 2005;21: 2988–2993. doi:10.1093/bioinformatics/bti457
29. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci. 2005;102: 13544–13549. doi:10.1073/pnas.0506577102
30. Kim SY, Volsky DJ. PAGE: Parametric analysis of gene set enrichment. BMC Bioinformatics. 2005. doi:10.1186/1471-2105-6-144
31. Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics. 2007. doi:10.1093/bioinformatics/btl599
32. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005;21: 1943–1949. doi:10.1093/bioinformatics/bti260
33. Glazko G V., Emmert-Streib F. Unite and conquer: Univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics. 2009. doi:10.1093/bioinformatics/btp406
34. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35: W169–W175. doi:10.1093/nar/gkm415
35. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet. 2002;31: 19–20. doi:10.1038/ng0502-19
36. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 2003.
37. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004. doi:10.1093/bioinformatics/btg455
38. Berriz GF, King OD, Bryant B, Sander C, Roth FP. Characterizing gene sets with FuncAssociate. Bioinformatics. 2003. doi:10.1093/bioinformatics/btg363
39. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol. 2004. doi:10.1186/gb-2004-5-12-r101
40. Castillo-Davis CI, Hartl DL. GeneMerge - Post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 2003. doi:10.1093/bioinformatics/btg114
41. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. 2008. doi:10.1093/nar/gkn276
42. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009. doi:10.1093/bioinformatics/btp101
43. Robinson MD, Grigull J, Mohammad N, Hughes TR. FunSpec: A web-based

cluster interpreter for yeast. BMC Bioinformatics. 2002;3. doi:10.1186/1471-2105-3-35
44. Martínez-Cruz LA, Rubio A, Martínez-Chantar ML, Labarga A, Barrio I, Podhorski A, et al. GARBAN: Genomic analysis and rapid biological annotation of cDNA microarray and proteomic data. Bioinformatics. 2003. doi:10.1093/bioinformatics/btg291
45. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004;20: 3710–3715. doi:10.1093/bioinformatics/bth456
46. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 2013. doi:10.1093/nar/gkt439
47. Sun H, Fang H, Chen T, Perkins R, Tong W. GOFFA: Gene Ontology for Functional Analysis - A FDA Gene Ontology tool for analysis of genomic and proteomic data. BMC Bioinformatics. 2006. doi:10.1186/1471-2105-7-S2-S23
48. Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 2006;34: W293–W297. doi:10.1093/nar/gkl031
49. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): A web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics. 2004. doi:10.1186/1471-2105-5-16
50. Luo W, Brouwer C. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013. doi:10.1093/bioinformatics/btt285
51. Yi M, Horton JD, Cohen JC, Hobbs HH, Stephens RM. WholePathwayScope: A comprehensive pathway-based analysis tool for high-throughput data. BMC Bioinformatics. 2006. doi:10.1186/1471-2105-7-30
52. Newton MA, Quintana FA, den Boon JA, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat. 2007. doi:10.1214/07-AOAS104
53. Cao W, Li Y, Liu D, Chen C, Xu Y. Statistical and Biological Evaluation of Different Gene Set Analysis Methods. Procedia Environ Sci. 2011;8: 693–699. doi:10.1016/j.proenv.2011.10.106
54. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics. 2007. doi:10.1186/1471-2105-8-242
55. Smyth GK. limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. doi:10.1007/0-387-29362-0_23
56. Breslin T, Edén P, Krogh M. Comparing functional annotation analyses with Catmap. BMC Bioinformatics. 2004. doi:10.1186/1471-2105-5-193
57. Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ. T-profiler: Scoring the activity of predefined groups of genes using gene expression data. Nucleic Acids Res. 2005. doi:10.1093/nar/gki484

58. Henegar C, Cancello R, Rome S, Vidal H, Clément K, Zucker J-D. Clustering Biological Annotations and Gene Expression Data To Identify Putatively Co-Regulated Biological Processes. J Bioinform Comput Biol. 2006;04: 833–852. doi:10.1142/S0219720006002181
59. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail-advanced gene set enrichment analysis. Nucleic Acids Res. 2007. doi:10.1093/nar/gkm323
60. Kim S-B, Yang S, Kim S-K, Kim SC, Woo HG, Volsky DJ, et al. GAzer: gene set analyzer. Bioinformatics. 2007;23: 1697–1699. doi:10.1093/bioinformatics/btm144
61. Wu D, Smyth GK. Camera: A competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012. doi:10.1093/nar/gks461
62. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: Generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics. 2009. doi:10.1186/1471-2105-10-161
63. Frost HR, Li Z, Moore JH. Spectral gene set enrichment (SGSE). BMC Bioinformatics. 2015;16: 70. doi:10.1186/s12859-015-0490-7
64. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene Sets Net Correlations Analysis (GSNCA): A multivariate differential coexpression test for gene sets. Bioinformatics. 2014. doi:10.1093/bioinformatics/btt687
65. Hsueh HM, Tsai CA. Gene set analysis using sufficient dimension reduction. BMC Bioinformatics. 2016. doi:10.1186/s12859-016-0928-6
66. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006;38: 500–501. doi:10.1038/ng0506-500
67. Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res. 2013;41: W98–W103. doi:10.1093/nar/gkt281
68. Wu X, Hasan M Al, Chen JY. Pathway and network analysis in proteomics. J Theor Biol. 2014. doi:10.1016/j.jtbi.2014.05.031
69. Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. Stat Appl Genet Mol Biol. 2005. doi:10.2202/1544-6115.1055
70. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009. doi:10.1093/bioinformatics/btn577
71. Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformatics. 2012. doi:10.1186/1471-2105-13-226
72. Glaab E, Baudot A, Krasnogor N, Valencia A. TopoGSA: Network topological gene set analysis. Bioinformatics. 2010. doi:10.1093/bioinformatics/btq131
73. Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 2013;41: e19–e19. doi:10.1093/nar/gks866
74. Rahmatallah Y, Emmert-Streib F, Glazko G. Comparative evaluation of gene set analysis approaches for RNA-Seq data. BMC Bioinformatics. 2014;15:

397. doi:10.1186/s12859-014-0397-8
75. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biology. 2016. doi:10.1186/s13059-016-0881-8
76. Young MD, Davidson N, Wakefield MJ, Smyth GK, Oshlack A. goseq : Gene Ontology testing for RNA-seq datasets Reading data. Genome Biol. 2010.
77. Ge SX, Son EW, Yao R. iDEP: An integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics. 2018. doi:10.1186/s12859-018-2486-6
78. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: Rotation gene set tests for complex microarray experiments. Bioinformatics. 2010. doi:10.1093/bioinformatics/btq401
79. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics. 2013. doi:10.1186/1471-2105-14-7
80. Fridley BL, Jenkins GD, Grill DE, Kennedy RB, Poland GA, Oberg AL. Soft truncation thresholding for gene set analysis of RNA-seq data: Application to a vaccine study. Sci Rep. 2013. doi:10.1038/srep02898
81. Yoon S, Kim SY, Nam D. Improving gene-set enrichment analysis of RNA-Seq data with small replicates. PLoS One. 2016. doi:10.1371/journal.pone.0165919
82. Xiong Q, Mukherjee S, Furey TS. GSAASeqSP: A toolset for gene set association analysis of RNA-Seq data. Sci Rep. 2014. doi:10.1038/srep06347
83. Wang X, Cairns MJ. SeqGSEA: A Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu090
84. Alhamdoosh M, Ng M, Wilson NJ, Sheridan JM, Huynh H, Wilson MJ, et al. Combining multiple tools outperforms individual methods in gene set enrichment analyses. Bioinformatics. 2017. doi:10.1093/bioinformatics/btw623
85. Stamm K, Tomita-Mitchell A, Bozdag S. GSEPD: A Bioconductor package for RNA-seq gene set enrichment and projection display. BMC Bioinformatics. 2019. doi:10.1186/s12859-019-2697-5
86. Lee C, Patil S, Sartor MA. RNA-Enrich: A cut-off free functional enrichment testing method for RNA-seq with improved detection power. Bioinformatics. 2016. doi:10.1093/bioinformatics/btv694
87. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. Am J Hum Genet. 2010. doi:10.1016/j.ajhg.2010.05.002
88. Nam D, Kim J, Kim S-Y, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. Nucleic Acids Res. 2010;38: W749–W754. doi:10.1093/nar/gkq428
89. Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet. 2007. doi:10.1086/522374

90. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. Schork NJ, editor. PLoS Genet. 2009;5: e1000384. doi:10.1371/journal.pgen.1000384
91. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010. doi:10.1002/gepi.2045
92. Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am J Hum Genet. 2008. doi:10.1016/j.ajhg.2008.06.024
93. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011. doi:10.1016/j.ajhg.2011.05.029
94. Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, et al. Gene set-based analysis of polymorphisms: Finding pathways or biological processes associated to traits in genome-wide association studies. Nucleic Acids Res. 2009. doi:10.1093/nar/gkp481
95. O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, et al. The SNP ratio test: Pathway analysis of genome-wide association datasets. Bioinformatics. 2009. doi:10.1093/bioinformatics/btp448
96. Chen X, Wang L, Hu B, Guo M, Barnard J, Zhu X. Pathway-based analysis for genome-wide association studies using supervised principal components. Genet Epidemiol. 2010;34: 716–724. doi:10.1002/gepi.20532
97. Luo L, Zhu Y, Xiong M. Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. Eur J Hum Genet. 2013;21: 217–224. doi:10.1038/ejhg.2012.141
98. Kim JH, Karnovsky A, Mahavisno V, Weymouth T, Pande M, Dolinoy DC, et al. LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. BMC Genomics. 2012. doi:10.1186/1471-2164-13-526
99. Sun R, Hui S, Bader GD, Lin X, Kraft P. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. Barsh GS, editor. PLOS Genet. 2019;15: e1007530. doi:10.1371/journal.pgen.1007530
100. Schwarz DF, Hädicke O, Erdmann J, Ziegler A, Bayer D, Möller S. SNPtoGO: Characterizing SNPs by enriched GO terms. Bioinformatics. 2008. doi:10.1093/bioinformatics/btm551
101. Holmans P, Green EK, Pahwa JS, Ferreira MAR, Purcell SM, Sklar P, et al. Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. Am J Hum Genet. 2009. doi:10.1016/j.ajhg.2009.05.011
102. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, et al. Pathway analysis by adaptive combination of P-values. Genet Epidemiol. 2009. doi:10.1002/gepi.20422
103. Bessarabova M, Ishkin A, et al. Knowledge-based analysis of proteomics data. BMC Bioinforma 2012 1316. 2012. doi:10.1186/1471-2105-13-S16-S13
104. Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, Ritchie MD, et al. Genetic analysis of biological pathway data through genomic

randomization. Hum Genet. 2011. doi:10.1007/s00439-011-0956-2
105. Moskvina V, O'Dushlaine C, Purcell S, Craddock N, Holmans P, O'Donovan MC. Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. Genet Epidemiol. 2011. doi:10.1002/gepi.20636
106. Lee PH, O'dushlaine C, Thomas B, Purcell SM. INRICH: Interval-based enrichment analysis for genome-wide association studies. Bioinformatics. 2012. doi:10.1093/bioinformatics/bts191
107. Araki H, Knapp C, Tsai P, Print C. GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. FEBS Open Bio. 2012;2: 76–82. doi:10.1016/j.fob.2012.04.003
108. Ayellet VS, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 2010. doi:10.1371/journal.pgen.1001058
109. Li MX, Kwan JSH, Sham PC. HYST: A hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. Am J Hum Genet. 2012. doi:10.1016/j.ajhg.2012.08.004
110. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–75. doi:10.1086/519795
111. Lips ES, Cornelisse LN, Toonen RF, Min JL, Hultman CM, Holmans PA, et al. Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. Mol Psychiatry. 2012;17: 996–1006. doi:10.1038/mp.2011.117
112. Pedroso I, Lourdusamy A, Rietschel M, Nöthen MM, Cichon S, McGuffin P, et al. Common genetic variants and gene-expression changes associated with bipolar disorder are over-represented in brain signaling pathway genes. Biol Psychiatry. 2012. doi:10.1016/j.biopsych.2011.12.031
113. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008;24: 2784–2785. doi:10.1093/bioinformatics/btn516
114. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, et al. Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease. Am J Hum Genet. 2009. doi:10.1016/j.ajhg.2009.01.026
115. Zhang K, Chang S, Cui S, Guo L, Zhang L, Wang J. ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework. Nucleic Acids Res. 2011;39: W437–W443. doi:10.1093/nar/gkr391
116. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. Nucleic Acids Res. 2010;38: W90–W95. doi:10.1093/nar/gkq324
117. Zhang K, Chang S, Guo L, Wang J. I-GSEA4GWAS v2: a web server for functional analysis of SNPs in trait-associated pathways identified from genome-wide association study. Protein Cell. 2015;6: 221–224.

doi:10.1007/s13238-014-0114-4
118. Jia P, Zheng S, Long J, Zheng W, Zhao Z. dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. Bioinformatics. 2011;27: 95–102. doi:10.1093/bioinformatics/btq615
119. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014;30: 523–530. doi:10.1093/bioinformatics/btt703
120. Wang L, Matsushita T, Madireddy L, Mousavi P, Baranzini SE. PINBPA: Cytoscape app for network analysis of GWAS data. Bioinformatics. 2015;31: 262–264. doi:10.1093/bioinformatics/btu644
121. Kutmon M, van Iersel MP, Bohler A, Kelder T, Nunes N, Pico AR, et al. PathVisio 3: An Extendable Pathway Analysis Toolbox. Murphy RF, editor. PLOS Comput Biol. 2015;11: e1004085. doi:10.1371/journal.pcbi.1004085
122. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27: 431–432. doi:10.1093/bioinformatics/btq675
123. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 2015. doi:10.1371/journal.pcbi.1004219
124. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2014;15: 504–518. doi:10.1093/bib/bbt002
125. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25: 25–29. doi:10.1038/75556
126. Kanehisa M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32: 277D – 280. doi:10.1093/nar/gkh063
127. Carbon S, Dietze H, Lewis SE, Mungall CJ, Munoz-Torres MC, Basu S, et al. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017;45: D331–D338. doi:10.1093/nar/gkw1108
128. Mishra P, Törönen P, Leino Y, Holm L. Gene set analysis: Limitations in popular existing methods and proposed improvements. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu374
129. Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza TM, Mukherjee S, et al. Comparative study of gene set enrichment methods. BMC Bioinformatics. 2009. doi:10.1186/1471-2105-10-275
130. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One. 2013. doi:10.1371/journal.pone.0079217
131. Pers TH. Gene set analysis for interpreting genetic studies. Human Molecular Genetics. 2016. doi:10.1093/hmg/ddw249
132. Sullivan PF, Posthuma D. Biological pathways and networks implicated in psychiatric disorders. Curr Opin Behav Sci. 2015;2: 58–68. doi:10.1016/j.cobeha.2014.09.003
133. Nurnberger JI, Koller DL, et al. Identification of Pathways for Bipolar Disorder.

JAMA Psychiatry. 2014;71: 657. doi:10.1001/jamapsychiatry.2014.176
134. Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJM. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. Hum Mol Genet. 2011. doi:10.1093/hmg/ddr248
135. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res. 2016. doi:10.1177/0962280212460441
136. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Gene-set analysis and reduction. Brief Bioinform. 2008;10: 24–34. doi:10.1093/bib/bbn042
137. Reuter JA, Spacek D V., Snyder MP. High-Throughput Sequencing Technologies. Mol Cell. 2015;58: 586–597. doi:10.1016/j.molcel.2015.05.004
138. Trevino V, Falciani F, Barrera-Saldaña HA. DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. Mol Med. 2007;13: 527–541. doi:10.2119/2006-00107.
139. Charpe AM. DNA Microarray. Advances in Biotechnology. New Delhi: Springer India; 2014. pp. 71–104. doi:10.1007/978-81-322-1554-7_6
140. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41: D991–D995. doi:10.1093/nar/gks1193
141. Golub TR, Slonim DK, Tamayo P, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science (80). 1999;286: 531–537. doi:10.1126/science.286.5439.531
142. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002. doi:10.1023/A:1012487302797
143. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23: 2507–2517. doi:10.1093/bioinformatics/btm344
144. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7: 3. doi:10.1186/1471-2105-7-3
145. Das S, Pandey P, Rai A, Mohapatra C. A computational system biology approach to construct gene regulatory networks for salinity response in rice (Oryza sativa). Indian J Agric Sci. 2015;85.
146. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med. 2004. doi:10.1016/j.artmed.2004.01.007
147. Lazar C, Taminau J, Meganck S, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinforma. 2012. doi:10.1109/TCBB.2012.33
148. Das S, Meher PK, et al. Inferring gene regulatory networks using Kendall's tau correlation coefficient and identification of salinity stress responsive genes in rice. Curr Sci. 2017;112. doi:10.18520/cs/v112/i06/1257-1262
149. Ding C, Peng H. Minimum redundancy feature selection from microarray

gene expression data. Computational Systems Bioinformatics CSB2003 Proceedings of the 2003 IEEE Bioinformatics Conference CSB2003. IEEE Comput. Soc; 2003. pp. 523–528. doi:10.1109/CSB.2003.1227396
150. Chen YW, Lin CJ. Combining SVMs with various feature selection strategies. Stud Fuzziness Soft Comput. 2006. doi:10.1007/978-3-540-35488-8_13
151. Hossain A, Willan AR, Beyene J. An improved method on wilcoxon rank sum test for gene selection from microarray experiments. Commun Stat Simul Comput. 2013. doi:10.1080/03610918.2012.667479
152. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics. 2002. doi:10.1093/bioinformatics/18.11.1454
153. Cheng T, Wang Y, Bryant SH. FSelector: a Ruby gem for feature selection. Bioinformatics. 2012;28: 2851–2852. doi:10.1093/bioinformatics/bts528
154. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics. 2017;18: 9. doi:10.1186/s12859-016-1423-9
155. Ding C, Peng H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. J Bioinform Comput Biol. 2005;03: 185–205. doi:10.1142/S0219720005001004
156. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997. doi:10.1016/s0004-3702(97)00043-x
157. Guyon I. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn. 1998. doi:10.1109/5254.708428
158. Duan KB, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience. 2005. doi:10.1109/TNB.2005.853657
159. Mundra PA, Rajapakse JC. SVM-RFE With MRMR Filter for Gene Selection. IEEE Trans Nanobioscience. 2010;9: 31–37. doi:10.1109/TNB.2009.2035284
160. Sohn I, Owzar K, George SL, Kim S, Jung SH. A permutation-based multiple testing method for time-course microarray experiments. BMC Bioinformatics. 2009. doi:10.1186/1471-2105-10-336
161. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47–e47. doi:10.1093/nar/gkv007
162. Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I. Fewer permutations, more accurate P-values. Bioinformatics. 2009. doi:10.1093/bioinformatics/btp211
163. Lai C, Reinders MJT, van't Veer LJ, Wessels LFA. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics. 2006. doi:10.1186/1471-2105-7-235
164. Kursa MB. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics. 2014. doi:10.1186/1471-2105-15-8
165. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE

Trans Pattern Anal Mach Intell. 2005. doi:10.1109/TPAMI.2005.159
166. Tiwari S, SL K, Kumar V, et al. Mapping QTLs for Salt Tolerance in Rice (Oryza sativa L.) by Bulked Segregant Analysis of Recombinant Inbred Lines Using 50K SNP Chip. Yadav RS, editor. PLoS One. 2016;11: e0153610. doi:10.1371/journal.pone.0153610
167. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004. doi:10.1093/nar/gkh036
168. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy - Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004. doi:10.1093/bioinformatics/btg405
169. Ware D. Gramene: a resource for comparative grass genomics. Nucleic Acids Res. 2002. doi:10.1093/nar/30.1.103
170. Sahani M, Linden J. Advances in neural information processing systems Processing Systems: Proceedings from the 2002 …. 2003. doi:http://dx.doi.org/10.1016/j.oraloncology.2016.02.011
171. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. An Introduction to the Bootstrap. Boston, MA: Springer US; 1993. doi:10.1007/978-1-4899-4541-9
172. Benjamini Y, Hochberg Y. Multiple Hypotheses Testing with Weights. Scand J Stat. 1997;24: 407–418. doi:10.1111/1467-9469.00072
173. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann Appl Stat. 2011;5: 1752–1779. doi:10.1214/11-AOAS466
174. Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. J Thorac Dis. 2017;9: 1725–1729. doi:10.21037/jtd.2017.05.34
175. Mazandu GK, Mulder NJ. Information content-based gene ontology functional similarity measures: Which one to use for a given biological data type? PLoS One. 2014. doi:10.1371/journal.pone.0113859
176. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation. Bioinformatics. 2003. doi:10.1093/bioinformatics/btg153
177. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007. doi:10.1093/bioinformatics/btm087
178. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: Improvements and new features. Nucleic Acids Res. 2007. doi:10.1093/nar/gkl976
179. Naeem H, Zimmer R, et al. Rigorous assessment of gene set enrichment tests. Bioinformatics. 2012. doi:10.1093/bioinformatics/bts164
180. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: Performance evaluation and usage guidelines. Brief Bioinform. 2012. doi:10.1093/bib/bbr049
181. Bargsten JW, Nap J-P, Sanchez-Perez GF, van Dijk ADJ. Prioritization of candidate genes in QTL regions based on associations between traits and biological processes. BMC Plant Biol. 2014;14: 330. doi:10.1186/s12870-014-0330-3

182. Gentleman RC, Carey VJ, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004. doi:10.1186/gb-2004-5-10-r80
183. Irizarry RA, Hobbs B, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003. doi:10.1093/biostatistics/4.2.249
184. Bland M. Do Baseline P-Values Follow a Uniform Distribution in Randomised Trials? PLoS One. 2013. doi:10.1371/journal.pone.0076010
185. Riley JW, Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM. The American Soldier: Adjustment During Army Life. Am Sociol Rev. 1949;14: 557. doi:10.2307/2087216
186. Won S, Morris N, Lu Q, Elston RC. Choosing an optimal method to combine P -values. Stat Med. 2009;28: 1537–1553. doi:10.1002/sim.3569
187. Lury DA, Fisher RA. Statistical Methods for Research Workers. Stat. 1972. doi:10.2307/2986695
188. Satopää VA, Baron J, et al. Combining multiple probability predictions using a simple logit model. Int J Forecast. 2014;30: 344–356. doi:10.1016/j.ijforecast.2013.09.009
189. Strimmer K. fdrtool: A versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics. 2008. doi:10.1093/bioinformatics/btn209
190. Strimmer K. A unified approach to false discovery rate estimation. BMC Bioinformatics. 2008;9: 303. doi:10.1186/1471-2105-9-303
191. Wang Y, Tetko I V., et al. Gene selection from microarray data for cancer classification - A machine learning approach. Comput Biol Chem. 2005. doi:10.1016/j.compbiolchem.2004.11.001
192. Miao Z, Zhang X. Differential expression analyses for single-cell RNA-Seq: old questions on new data. Quant Biol. 2016. d:10.1007/s40484-016-0089-7
193. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. Comparison of methods to detect differentially expressed genes between single-cell populations. Brief Bioinform. 2016; bbw057. doi:10.1093/bib/bbw057
194. Sandberg R. Entering the era of single-cell transcriptomics in biology and medicine. Nature Methods. 2014. doi:10.1038/nmeth.2764
195. Trapnell C. Defining cell types and states with single-cell genomics. Genome Research. 2015. doi:10.1101/gr.190595.115
196. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011. doi:10.1101/gr.110882.110
197. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7: 39921. doi:10.1038/srep39921
198. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biology. 2016. doi:10.1186/s13059-016-0927-y
199. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The Technology and Biology of Single-Cell RNA Sequencing. Molecular Cell.

2015. doi:10.1016/j.molcel.2015.04.005
200. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics. 2015. doi:10.1038/nrg3833
201. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics. 2019. doi:10.1186/s12859-019-2599-6
202. Finak G, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16: 278. 10.1186/s13059-015-0844-5
203. Van den Berge K, Perraudeau F, Soneson C, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19: 24. doi:10.1186/s13059-018-1406-4
204. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140. doi:10.1093/bioinformatics/btp616
205. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106
206. Love MI, Anders S, Huber W. Differential analysis of count data - the DESeq2 package. Genome Biology. 2014. doi:110.1186/s13059-014-0550-8
207. Fujita K, Iwaki M, Yanagida T. Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA. Nat Commun. 2016. doi:10.1038/ncomms13788
208. Wang J, Huang M, Torre E, Dueck H, Shaffer S, Murray J, et al. Gene expression distribution deconvolution in single-cell RNA sequencing. Proc Natl Acad Sci U S A. 2018. doi:10.1073/pnas.1721085115
209. Ye C, Speed TP, Salim A. DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data. Berger B, editor. Bioinformatics. 2019;35: 5155–5162. doi:10.1093/bioinformatics/btz453
210. Van den Berge K, Soneson C, Love MI, Robinson MD, Clement L. zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications. doi.org. 2017. doi:10.1101/157982
211. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Berger B, editor. Bioinformatics. 2018;34: 3223–3224. doi:10.1093/bioinformatics/bty332
212. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014. doi:10.1038/nbt.2859
213. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14: 309–315. doi:10.1038/nmeth.4150
214. Kharchenko P V., Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11: 740–742. doi:10.1038/nmeth.2967
215. Mou T, Deng W, Gu F, Pawitan Y, Vu TN. Reproducibility of Methods to

Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. Front Genet. 2020. doi:10.3389/fgene.2019.01331
216. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018. doi:10.1038/nmeth.4612
217. Dal Molin A, Baruzzo G, Di Camillo B. Single-cell RNA-sequencing: Assessment of differential expression analysis methods. Front Genet. 2017. doi:10.3389/fgene.2017.00062
218. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2009. doi:10.1093/bioinformatics/btp612
219. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. Stat Appl Genet Mol Biol. 2011. doi:10.2202/1544-6115.1637
220. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29: 1035–1043. doi:10.1093/bioinformatics/btt087
221. Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, et al. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics. 2016. doi:10.1093/bioinformatics/btw202
222. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016. doi:10.1186/s13059-016-1077-y
223. Sengupta D, Rayan NA, Lim M, Lim B, Prabhakar S. Fast, scalable and accurate differential expression analysis for single cells. bioRxiv. 2016. doi:10.1101/049734
224. Welch BL. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. Biometrika. 1947. doi:10.2307/2332510
225. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bull. 1945;1: 80. doi:10.2307/3001968
226. Seyednasrollah F, Rantanen K, Jaakkola P, Elo LL. ROTS: Reproducible RNA-seq biomarker detector - Prognostic markers for clear cell renal cell cancer. Nucleic Acids Res. 2016. doi:10.1093/nar/gkv806
227. Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: A robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. Bioinformatics. 2016. doi:10.1093/bioinformatics/btv634
228. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15: R29. doi:10.1186/gb-2014-15-2-r29
229. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010. doi:10.1186/gb-2010-11-3-r25
230. Hardcastle TJ, Kelly KA. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010. doi:10.1186/1471-2105-11-422

231. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013. doi:10.1038/nbt.2450
232. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013. doi:10.1177/0962280211428386
233. Frazee A, Pertea G, Jaffe A, Langmead B, Salzberg S, Leek J. Flexible analysis of transcriptome assemblies with Ballgown. bioRxiv. 2014. doi:10.1101/003665
234. Auer PL, Doerge RW. A two-stage poisson model for testing RNA-Seq data. Stat Appl Genet Mol Biol. 2011. doi:10.2202/1544-6115.1627
235. Elo LL, Filén S, Lahesmaa R, Aittokallio T. Reproducibility-optimized test statistic for ranking genes in microarray studies. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2008. doi:10.1109/tcbb.2007.1078
236. Paulson JN, Colin Stine O, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013. doi:10.1038/nmeth.2658
237. Delmans M, Hemberg M. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics. 2016. doi:10.1186/s12859-016-0944-6
238. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. PLoS Comput Biol. 2015. doi:10.1371/journal.pcbi.1004575
239. Zhang W, Wei Y, Zhang D, Xu EY. ZIAQ: a quantile regression method for differential expression analysis of single-cell RNA-seq data. Cowen L, editor. Bioinformatics. 2020;36: 3124–3130. doi:10.1093/bioinformatics/btaa098
240. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods. 2018;145: 25–32. doi:10.1016/j.ymeth.2018.04.017
241. Jia C, Hu Y, Kelly D, Kim J, Li M, Zhang NR. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res. 2017. doi:10.1093/nar/gkx754
242. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9: 284. doi:10.1038/s41467-017-02554-5
243. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. Morris Q, editor. PLOS Comput Biol. 2015;11: e1004333. doi:10.1371/journal.pcbi.1004333
244. Chen W, Li Y, Easton J, Finkelstein D, Wu G, Chen X. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. Genome Biol. 2018;19: 70. doi:10.1186/s13059-018-1438-9
245. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Commun. 2020. doi:10.1038/s41467-020-14766-3
246. Wu Z, Zhang Y, Stitzel ML, Wu H. Two-phase differential expression analysis

for single cell RNA-seq. Bioinformatics. 2018. doi:10.1093/bioinformatics/bty329
247. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Res. 2011. doi:10.1101/gr.124321.111
248. Van De Wiel MA, Leday GGR, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2013. doi:10.1093/biostatistics/kxs031
249. Srivastava S, Chen L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res. 2010. doi:10.1093/nar/gkq670
250. Chu C, Fang Z, Hua X, et al. deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies. BMC Genomics. 2015. doi:10.1186/s12864-015-1676-0
251. Qiu X, Mao Q, Tang Y, Wang L, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017. doi:10.1038/nmeth.4402
252. Sekula M, Gaskins J, Datta S. Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects. Biometrics. 2019. doi:10.1111/biom.13074
253. Jiang L, Schlesinger F, Davis CA, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011. doi:10.1101/gr.121095.111
254. Moliner A, Enfors P, Ibáñez CF, Andäng M. Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials. Stem Cells Dev. 2008. doi:10.1089/scd.2007.0211
255. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv. 2014. doi:10.1101/003236
256. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015. doi:10.1016/j.cell.2015.04.044
257. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. Nat Methods. 2017. doi:10.1038/nmeth.4179
258. Savas P, Virassamy B, Ye C, Salim A, Mintoff CP, Caramia F, et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. Nat Med. 2018. doi:10.1038/s41591-018-0078-7
259. Grün D, Kester L, Van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014. doi:10.1038/nmeth.2930
260. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017. doi:10.1016/j.molcel.2017.01.023
261. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Softw. 2008. doi:10.18637/jss.v027.i08
262. Kemp CD, Kemp AW. Some properties of the “Hermite” distribution.

Biometrika. 1965;52: 381–394. doi:10.1093/biomet/52.3-4.381
263. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007. doi:10.1093/bioinformatics/btm453
264. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 2008. doi:10.1093/biostatistics/kxm030
265. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;25: 1026–1032. doi:10.1093/bioinformatics/btp113
266. Yoon K, Hwang C-L. Multiple Attribute Decision Making. Multiple Attribute Decision Making. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc.; 1995. doi:10.4135/9781412985161
267. Khezrian M, Jahan A, Kadir WMNW, Ibrahim S. An approach for web service selection based on confidence level of decision maker. PLoS One. 2014. doi:10.1371/journal.pone.0097831
268. Ahn BS. Compatible weighting method with rank order centroid: Maximum entropy ordered weighted averaging approach. Eur J Oper Res. 2011. doi:10.1016/j.ejor.2011.02.017
269. Assari A. Role of public participation in sustainability of historical city: usage of TOPSIS method. Indian J Sci Technol. 2012. doi:10.17485/ijst/2012/v5i3.2
270. Moriña D, Higueras M, Puig P, Oliveira M. Generalized hermite distribution modelling with the R package hermite. R J. 2015. doi:10.32614/rj-2015-035
271. Long JS, Freese J. Regression Models for Categorical Dependent Variables Using STATA. Sociology The Journal Of The British Sociological Association. 2001. doi:10.1186/2051-3933-2-4
272. Tasic B, Menon V, Nguyen TN, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016;19: 335–346. doi:10.1038/nn.4216
273. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet. 2019;10. doi:10.3389/fgene.2019.00317
274. Zeisel A, Moz-Manchado AB, Codeluppi S, Lönnerberg P, Manno G La, Juréus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (80-). 2015. doi:10.1126/science.aaa1934
275. Tian L, Su S, Dong X, et al. scPipe: A flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLoS Comput Biol. 2018. doi:10.1371/journal.pcbi.1006361
276. Ramsköld D, Luo S, Wang YC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012. doi:10.1038/nbt.2282
277. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Rep. 2012. doi:10.1016/j.celrep.2012.08.003
278. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research. 2018;7:

1141. doi:10.12688/f1000research.15666.1
279. Petropoulos S, Edsgärd D, Reinius B, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016. doi:10.1016/j.cell.2016.03.023
280. MacParland SA, Liu JC, Ma XZ, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018. doi:10.1038/s41467-018-06318-7
281. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. J R Stat Soc Ser B. 1977;39: 1–22. doi:10.1111/j.2517-6161.1977.tb01600.x
282. McKinnon KIM. Convergence of the Nelder-Mead simplex method to a nonstationary point. SIAM J Optim. 1998. doi:10.1137/S1052623496303482
283. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011. doi:10.1186/1471-2105-12-77
284. Ledford H. The death of microarrays? Nature. 2008;455: 847–847. doi:10.1038/455847a
285. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008. doi:10.1038/nature07509
286. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (80-). 2008. doi:10.1126/science.1162228
287. Wilhelm BT, Marguerat S, Watt S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008. doi:10.1038/nature07002
288. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009. doi:10.1038/nature08460
289. Das S, McClain CJ, Rai SN. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. Entropy. 2020;22: 427. doi:10.3390/e22040427
290. Formentin E, Sudiro C, Perin G, et al. Transcriptome and Cell Physiological Analyses in Different Rice Cultivars Provide New Insights Into Adaptive and Salinity Stress Responses. Front Plant Sci. 2018;9. doi:10.3389/fpls.2018.00204
291. Kawahara Y, de la Bastide M, Hamilton JP, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6: 4. doi:10.1186/1939-8433-6-4
292. Törönen P, Ojala PJ, Marttinen P, Holm L. Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. BMC Bioinformatics. 2009;10: 307. doi:10.1186/1471-2105-10-30.

APPENDIX I

ACRONYMS

Acronyms	Full form
ACC	Accuracy
AIC	Akaike Information Criterion
AUC	Area Under Curve
AUROC	Area Under Receiver Operating Characteristics curve
BIC	Bayesian Information Criterion
BP	Biological Process
BSM	Bootstrap-SVM-MRMR
BSS	Between cluster Sum of Squares
CA	Classification Accuracy
CC	Cellular Component
DE	Differential Expression
DEG	Differentially Expressed Genes
DEZIG	Differentially Expressed and Differentially Zero Inflated Genes
df	degree of freedom
DZI	Differential Zero-Inflated
DZIG	Differentially Zero Inflated Genes
ECM	Expected Conditional Maximization
EM	Expected Maximization
EMD	Earth Mover's Distance
ERCC	External RNA Controls Consortium
ES	Enrichment Score
F1	F1 score
FC	Fold Change
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
FP	True Negative
FPR	False Positive Rate
GAM	Generalized Additive Model
GE	Gene Expression
GEO	Gene Expression Omnibus
GL	Generalized Linear
GO	Gene Ontology
GR	Gain Ratio
GSA	Gene Set Analysis
GSAQ	Gene Sets Analysis with trait specific QTLs
GSEA	Gene Set Enrichment Analysis
GSQSeq	Gene Set Analysis with QTL for RNA-seq
GSVA	Gene Set Variation Analysis
GSVQ	Gene Set Validation with QTLs

GWAS	Genome Wide Association Study
HD	Hermite Distribution
iDEP	integrated Differential Expression and Pathway
IG	Information Gain
iid	independently and identically distributed
KS	Kolmogorov-Smirnov test
LRT	Likelihood Ratio Test
MCDM	Multiple Criteria Decision Making
MF	Molecular Function
MLE	Maximum Likelihood
MRMR	Maximum Relevance and Minimum Redundancy
NB	Negative Binomial
NIS	Negative Ideal Solution
NP	Non-Parametric
NPV	Negative Prediction value
ORA	Over Representation Analysis
PCR	Pearson's Correlation
PD	Poisson Distribution
PIS	Positive Ideal Solution
PMF	Probability Mass Function
PPV	Positive Prediction Rate
QLF	Quasi-Likelihood F test
QTL	Quantitative Trait Loci
RF	Random Forest
RMA	Robust Multichip Average
RNA-seq	RNA-sequencing
rv	Random Variable
scRNA-seq	single cell RNA-sequencing
SE	Standard Error
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
SRC	Spearman's Rank Correlation
SU	Symmetrical Uncertainty
SVM	Support Vector Machine
SVM-MRMR	SVM-RFE with MRMR
SVM-RFE	Support Vector Machine-Recursive Feature Elimination
TOPSIS	Technique for Order Performance by Similarity to Ideal Solution
TP	True Positives
TPR	True Positive Rate
TSS	Total Sum of Squares
UMI	Unique Molecular Identifier
Wilcox	Wilcoxon Signed Rank Test
WSS	Within cluster Sum of Squares
ZIM	Zero Inflated Model
ZINB	Zero Inflated Negative Binomial
ZIPD	Zero Inflated Poisson Distribution

APPENDIX II

Objective function of Support Vector Machine

To maximize the distance between the hyper planes in Eq. 3.5 (Chapter 3), the objective function $\frac{\|k\|^2}{\mathbf{2}}$ is minimized. Hence, the objective function, *J*, for this case vs. control classification problem becomes.

$$J = \|k\|^2/2$$

Through Taylor series expansion, the objective function, J, can be approximated as (at *k*=c):

$$J = \frac{1}{2}\left\{\|c\|^2 + \frac{\partial}{\partial k}\|k\|^2{}_{k=c}(k-c) + \frac{1}{2!}\frac{\partial^2}{\partial k^2}\|k\|^2{}_{k=c}(k-c)^2 + \frac{1}{3!}\|k\|^2{}_{k=c}(k-c)^3 + \cdots\right\}$$

$$= \frac{1}{2}\left\{\|c\|^2 + \frac{\partial J_2}{\partial k}{}_{k=c}(k-c) + \frac{1}{2!}\frac{\partial^2 J_2}{\partial k^2}{}_{k=c}(k-c)^2 + \frac{1}{3!}\frac{\partial^3 J_2}{\partial k^3}{}_{k=c}(k-c)^3 + \cdots\right\}$$

Differentiating both sides with respect to *k,* and ignoring the second and higher order derivatives terms from the expression, we have

$$\frac{\partial J}{\partial k} = \frac{1}{2}\left\{0 + \frac{\partial J}{\partial k}\left(\frac{\partial J}{\partial k}{}_{k=c}(k-c)\right)\right\}$$

$$\frac{\partial J}{\partial k} = \frac{1}{2}\left\{(k-c)\frac{\partial^2 J}{\partial k^2} + \frac{\partial J}{\partial k}.1)\right\}$$

$$\frac{\partial J}{\partial k} = \frac{1}{2}\frac{\partial^2 J}{\partial k^2}(k-c)$$

Replace *k* with Δk in above expression and ignoring the constant *c*. Now, the above expression becomes, as below.

$$\lim_{\Delta k}\frac{\Delta J}{\Delta k} = \frac{1}{2}\frac{\partial^2 J}{\partial k^2}(\Delta k)$$

$$\Delta J = \frac{1}{2}\frac{\partial^2 J}{\partial k^2}(\Delta k)^2$$

Further, ΔJ attributed to i^{th} gene can be expressed as:

$$\Delta J(i) = \frac{1}{2}\frac{\partial^2 J_2}{\partial {k_i}^2}(\Delta k_i)^2$$

APPENDIX III

Distribution of observed scRNA-seq UMI counts

In Chapter 6, $Z_{ijk} \sim ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$ and $\rho_{ijk} = (Y_{ijk}|Z_{ijk} = z) \sim B(z, p_{ijk})$

The Probability Mass Function (PMF) of Z_{ijk} is expressed as:

$$P[Z_{ijk} = z] = \begin{cases} \pi_{ijk} + (1 - \pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}} & when\ z = 0 \\ (1 - \pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z};\ z > 0 \end{cases}$$

$$P[Y_{ijk} = y | Z_{ijk} = z] = \binom{z}{y} {p_{ijk}}^{y}(1 - p_{ijk})^{z-y}$$

The joint PMF of Y_{ijk} and Z_{ijk} can be written as:

$$P[Y_{ijk} = y, Z_{ijk} = z | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] = P[Y_{ijk} = y | Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = z | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}]$$

Now, the marginal probability distribution of Y_{ijk} can be given as:

$$P[Y_{ijk} = y | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] = \sum_z P[Y_{ijk} = y | Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = z | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}]$$

<u>Case-1</u>: For zero count $(Y_{ijk} = 0)$ **case**

$$[Y_{ijk} = 0 | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] = \sum_z P[Y_{ijk} = 0 | Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = z | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}]$$

$$= P[Y_{ijk} = 0 | Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = 0 | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}] + \sum_{z=1}^{\infty} P[Y_{ijk} = 0 | Z_{ijk} = z, p_{ijk}]P[\, Z_{ijk} = z | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}]$$

$$= \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}} + \sum_{z=1}^{\infty}\left\{(1-p_{ijk})^{z}(1 - \pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}\right\}$$

$$= \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left[\sum_{z=0}^{\infty}\left\{(1-p_{ijk})^{z}\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}\right\}\right]$$

$$= \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left[\sum_{z=0}^{\infty}\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{z}\left(1 - \frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right]\left(1-\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{-\theta_{ijk}}$$

$$= \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(1-\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{-\theta_{ijk}}$$

$$= \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{\theta_{ijk}}$$

<u>Case-2</u>: For non-zero counts, *i.e.* $Y_{ijk}(>0) = t = 1, 2, 3, \ldots$

$$P\left[Y_{ijk}=t\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk},p_{ijk}\right] = \sum_{z\geq t}P\left[Y_{ijk}=t\middle|Z_{ijk}=z,p_{ijk}\right]P\left[\,Z_{ijk}=z\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk}\right]$$

$$= \sum_{z\geq t}\binom{z}{t}{p_{ijk}}^{t}(1-p_{ijk})^{z-t}(1-\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}$$

$$= (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\sum_{z\geq t}\binom{z}{t}{p_{ijk}}^{t}(1-p_{ijk})^{z-t}\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}$$

$$= \frac{(1-\pi_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{p_{ijk}}{1-p_{ijk}}\right)^{t}\sum_{z\geq t}\frac{G(z+\theta_{ijk})}{G(z-t+1)}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}(1-p_{ijk})^{z}$$

Let, $z' = z - t$

$$= \frac{(1-\pi_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{p_{ijk}}{1-p_{ijk}}\right)^{t}\sum_{z'=0}\frac{G(z'+t+\theta_{ijk})}{G(z'+1)}\left(\frac{\mu_{ijk(1-p_{ijk})}}{\theta_{ijk}+\mu_{ijk}}\right)^{z'+t}$$

$$= (1 - \pi_{ijk})\frac{G(t+\theta_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{t}\sum_{z'=0}\frac{G(z'+t+\theta_{ijk})}{G(z'+1)G(t+\theta_{ijk})}\left(\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{z'}$$

$$= (1-\pi_{ijk})\frac{G(t+\theta_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{t}\left(1-\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{-(t+\theta_{ijk})}$$

$$= (1-\pi_{ijk})\frac{G(t+\theta_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{t}$$

So, the distribution of Y_{ijk} is also $ZINB(\pi_{ijk}, \mu_{ijk}p_{ijk}, \theta_{ijk})$. Now, the PMF of Y_{ijk} can be expressed as:

Let, $\mu_{ijk}p_{ijk} = \mu'_{ijk}$

$$P[Y_{ijk} = y|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] = \begin{cases} \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{\theta_{ijk}} \\ (1-\pi_{ijk})\frac{G(t+\theta_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu'_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{y} \end{cases}$$

The expected value and variance of observed read counts can be obtained as:

$$E(Y) = \sum_{y=0}^{\infty} y\, f_{zinb}(y)$$

$$= 0 f_{zinb}(0) + \sum_{y=1}^{\infty} y\, f_{zinb}(y)$$

$$= (1-\pi_{ijk})\sum_{y=0}^{\infty} y\, f_{NB}(y) = (1-\pi_{ijk})\mu$$

$$Var(Y) = E(Y^2) - \{E(Y)\}^2$$

$$E(Y^2) = \sum_{y=0}^{\infty} y^2 f_{zinb}(y) = (1-\pi_{ijk})(\mu'_{ijk} + \frac{\mu'^{2}_{ijk}}{\theta_{ijk}} + \mu'^{2}_{ijk})$$

Now, $Var(Y) = (1-\pi_{ijk})\left(\mu'_{ijk} + \frac{\mu'^{2}_{ijk}}{\theta_{ijk}} + \mu'^{2}_{ijk}\right) - (1-\pi_{ijk})^2 \mu'^{2}_{ijk}$

$$= (1-\pi_{ijk})\mu\left(1 + \pi_{ijk}\mu'_{ijk} + \frac{\mu'_{ijk}}{\theta_{ijk}}\right)$$

$E(Y_{ijk}) = (1-\pi_{ijk})\mu_{ijk}p_{ijk}$

$Var(Y_{ijk}) = (1-\pi_{ijk})\mu_{ijk}p_{ijk}(1 + \pi_{ijk}\mu_{ijk}p_{ijk} + \frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}})$

APPENDIX IV

Distribution of sample mean and variance of observed UMI counts

The expected values of gene-wise sample mean, and sample variance of the observed scRNA-seq UMI count data can be obtained as follows.

Here, we assume that the cell capture rates of the genes remain same, *i.e.* $p_{ij1} = p_{ij2} = \cdots = p_{ijK} = p_{ij}$, and the model parameters, μ_{ijk} and θ_{ijk} for the genes remain same across the cells. Let, $\bar{Y}_{..k}$: sample mean expression values of k^{th} gene and its expected values can be given as:

$$E(\bar{Y}_{..k}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}E(Y_{ijk})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}E\{E(Y_{ijk}|Z_{ijk})\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}E(Z_{ijk}p_{ijk})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}(\mu_{ijk}p_{ijk})$$

Under the assumption, *i.e.* $p_{ij1} = p_{ij2} = \cdots = p_{ijK} = p_{ij}$, the above expression can be also written as:

$$E(\bar{Y}_{..k}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}(\mu_k p_{ij})$$

$$= \frac{1}{N}\mu_k\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}p_{ij} = \mu_k\bar{p}_{..}$$

Further, the variance of the observed scRNA-seq data can be obtained as:

$$V(Y_{ijk}) = E\{V(Y_{ijk}|Z_{ijk})\} + V\{E(Y_{ijk}|Z_{ijk})\}$$

$$= E(Z_{ijk}p_{ijk})(1-p_{ijk}) + V(Z_{ijk}p_{ijk})$$

$$= p_{ijk}(1-p_{ijk})\mu_{ijk} + p_{ijk}^2(\mu_{ijk} + {\mu_{ijk}^2}/{\theta_{ijk}})$$

$$= \mu_{ijk} p_{ijk} (1 + {}^{\mu_{ijk} p_{ijk}}/_{\theta_{ijk}})$$

Under the above assumptions, the $V(Y_{ijk})$ becomes:

$$V(Y_{ijk}) = \mu_k p_{ij} (1 + {}^{\mu_k p_{ij}}/_{\theta_k})$$

Let, S_k^2 be the sample variance of k^{th} gene. Then its expected value can be derived as:

$$E(S_k^2) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(M_i - 1)} \sum_{j=1}^{M_i} \left\{ V(Y_{ijk}) + E(Y_{ijk})^2 \right\} - \frac{1}{N(N-1)} \sum_{i \neq i'=1}^{N} \frac{1}{M_i (M_i - 1)} \sum_{j \neq j'=1}^{M_i} E(Y_{ijk}) E(Y_{i'j'k})$$

$$= \mu_k \bar{p}_{..} + \frac{{\mu_k}^2}{\theta_k} \overline{p_{ij}^2} + {\mu_k}^2 var(p_{ij})$$

where, $\bar{p}_{..}$, $\overline{p_{ij}^2}$ are the means of p_{ij} and p_{ij}^2 respectively and $var(p_{ij})$ is variance of p_{ij}

CURRICULUM VITAE

Name: Samarendra Das

UofL id: s0das009
Student id: 5235902
Ph. D. Candidate
University of Louisville, Ky, USA
502-554-0118
S0das009@louisville.edu

EDUCATION

INSTITUTION AND LOCATION	DEGREE	COMPLETION	FIELD OF STUDY
Orissa University of Agriculture and Technology, Bhubaneswar 755001, Odisha, India	B.S.	08/2009	Agriculture
Indian Agricultural Research Institute, New Delhi 110012, India	M.S.	08/2011	Agricultural Statistics
University of Louisville, Louisville, Ky 40292, USA	Ph.D.	12/2020	Bioinformatics

RESEARCH EXPERIENCE

Ph. D. Candidate, University of Louisville, USA
(August 2017 – December 2020)
Proposed Dissertation Title: Statistical Approaches of Gene Set Analysis with Quantitative Trait Loci for High-Throughput Genomic Studies

- Responsible for experiments designing, statistical methodology and tools development for high-throughput genomic studies, big data analytics, data mining.
- Responsible for software development, data analysis, manuscript drafting and revision, research grant writing.

Mentor: Dr. Shesh Nath Rai
Professor, Dept. of Bioinformatics and Biostatistics
University of Louisville, Ky, USA
Email id: shesh.rai@louisville.edu
Ph. 502/472-9120

Student Assistant, James Graham Brown Cancer Center
(November 2019 – December 2020)

- Responsible for assisting the supervisor on preparing the teaching materials, data analysis, grant writing, reviewing manuscripts, drafting manuscripts and other activities.

- Responsible for R codes and package development
- Supervised by: Shesh N. Rai, Ph.D.
  Director, Biostatistics and Bioinformatics facility, James Brown Cancer Center
  University of Louisville, Ky, USA

Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India
(January 2013 – August 2017)
Project 1: Modelling and construction of transcriptional regulatory networks using time-series gene expression data (AGENIASRISIL201401100030)
Role: Principal Investigator (05/2014 –01/2017)
Funded by Indian Council of Agricultural Research, New Delhi, India

Project 2: Development of gene selection approaches for classification of crop gene expression data. (AGENIASRISIL201503000067)
Role: Principal Investigator (10/2015 – 10/2018)
Funded by Indian Council of Agricultural Research, New Delhi, India

Project 3: Development of Rank based Stability Measures for Selecting Genotypes. (AGENIASRISIL201502500062)
Role: Co-Principal Investigator (09/2015 – 08/2017)
Funded by Indian Council of Agricultural Research, New Delhi, India

- Responsible for research project writing, leading research projects as PI and Co-PI, project writing, manuscript drafting, reviewing manuscripts and project reports.
- Responsible for development of statistical approaches, models, and software for high dimensional gene expression data.
- Responsible for development of data mining techniques genomic data analysis.

Research Student, ICAR-Indian Agricultural Research Institute, New Delhi, India
(August 2009 – August 2012)
- Responsible for analyzing agricultural crop data, writing codes in SAS, R.
- Assisting supervisor for preparing teaching materials

Mentor(s): A. K. Paul, Ph.D., Principal Scientist
Anil Rai, Ph.D., Principal Scientist
ICAR-Indian Agricultural Research Institute, New Delhi, India

TEACHING AND TRAINING EXPERIENCE

Faculty, ICAR-Indian Agricultural Research Institute, New Delhi, India
(January 2015 – August 2017)
- Responsible for teaching the following academic courses to MS and Ph.D. students
  - ✓ Mathematical Methods in Statistics (AS 551)
  - ✓ Advanced Statistical Genetics (AS 602)
  - ✓ Bioinformatics-I (AS-571)
  - ✓ Mathematical Foundations in Computer Application (CA-551/BI-503)
  - ✓ Senior Certificate Courses (Module I and II)
  - ✓ Biological Network Modelling and Analysis (BI-614)

Training Instructor
- Responsible for organizing various CAFT and Winter School training programs for faculties of State Agricultural universities, ICAR-institutes

- Organized one Centre for Advanced Faculty Training program on “Recent Analytical Techniques in Statistical Genetics and Genomics” during January 17 – February 06, 2017 as program coordinator.
- Organized one Winter School on “Advanced Statistical Techniques in Genetics and Genomics” as course co-director during 2-22 March 2017.
- Administrative management of training programs

EMPLOYMENT HISTORY

01/2013 - 05/2013	Foundation Trainee, ICAR-National Academy of Agricultural Research Management, Hyderabad, India
06/2013 -09/2013	Attachment Trainee, National Institute of Biomedical Genomics, Kalyani, West Bengal, India
10/2013 - 07/2017	Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India
08/2017 - 12/2020	Ph.D. graduate student, University of Louisville, USA
11/2019 - 12/2020	Student Assistant, James Graham Brown Cancer Center, Louisville, USA

AWARDS

a)	Nehru Memorial Gold Medal Award, Indian Agricultural Statistics Research Institute, New Delhi, India (2009 – 2011)
b)	MN Das Young Scientist Award (awarded as Runner UP), Society of Statistics, Computers and Applications in its 18th Annual National Conference (February 18-20, 2016) organized at University of Jammu, Jammu, India.
c)	MN Das Memorial Young Scientist appreciation certificate, Society of Statistics, Computers and Applications for the year 2016-17.
d)	Best paper award in the field of Statistical Theory and Methodology, Indian Society of Agricultural Statistics (2018), New Delhi, India
e)	Student Assistantship Award, James Graham Brown Cancer Center, Louisville, USA (2019)
f)	Graduate Dean's Citations award, University of Louisville, Ky, USA (2020)

FELLOWSHIPS

a)	Netaji Subash-ICAR International fellowship (2016-17), Indian Council of Agricultural Research, New Delhi, India
b)	IARI Scholarship (2011-12), Indian Agricultural Research Institute, New Delhi, India
c)	ICAR Junior Research Fellowship (JRF) by Indian Council of Agricultural Research, New Delhi, India (2009 – 11)
d)	Got the financial aid/ graduate assistantship from Dept. Hepatobiology and Toxicology, University of Louisville, Ky, USA (2017 - 2020)

PROFESSIONAL SOCIETY

a)	Life member, Society of Biotechnology and Bioinformatics, OUAT, Odisha, India
b)	Student member, American Statistical Association, USA

PUBLICATIONS

1.	*Dissertation project*: Statistical Approaches of Gene Set Analysis with Quantitative Trait Loci for High-Throughput Genomic Studies Publications: 1. Das, S. and Rai, S.N. (2020) Statistical Approach for Biologically Relevant Gene Selection from High-Throughput Gene Expression Data. *Entropy*, *22*(11), 1205. doi.org/10.3390/e22111205. 2. Das, Samarendra, McClain, C. and Rai, S.N. (2020). Fifteen Years of Gene Set Analysis for Genomic Studies: A Review of Statistical Approaches and Future Challenges. *Entropy* 22(4), 427; doi.org/10.3390/e22040427. 3. Das, S. and Rai, S.N. (2020). SwarnSeq: An improved Statistical Approach for SwarnSeq: An Improved Statistical Approach for Differential Expression Analysis of Single-Cell RNA-Seq Data. *Genomics*. (Accepted). 4. Das, S. and Rai, S.N. (2020). Differential Expression Analysis of Single Cell RNA-Seq Data: An Overview and Comparative Analysis. *Genome Biology*. (under review). 5. Das, S. and Rai, S.N. (2020). Statistical Approach for Gene Set Analysis with Trait Specific Quantitative Trait Loci for RNA-seq Data. *Plos One*. (under review). 6. Das, Samarendra., Rai, A., Mishra, D.C. Rai, S.N. (2018). Statistical Approach for Gene Set Analysis with Trait Specific Quantitative Trait Loci. *Sci Rep* 8, 2391. doi.org/10.1038/s41598-018-19736-w.
2.	*Project 1*. Modelling and construction of transcriptional regulatory networks using time-series gene expression data. Publications: 1. Das, S., Pandey P., Rai, A. and Mohapatra C. (2015). A computational systems biology approach to construct gene regulatory networks for salinity response in rice (Oryza sativa). *Indian Journal of Agricultural Sciences*. 85(12): 1546–52. 2. Das, S., Meher P.K., Pradhan U.P., Paul A.K. (2017). Inferring gene regulatory networks using Kendall's tau correlation coefficient and identification of salinity stress responsive genes in rice. *Current Science* 112(6):1257-63. 3. Das, S., Meher PK, Rai A, Bhar LM, Mandal BN (2017) Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (Glycine max L.). *PLoS ONE* 12(1): e0169605. doi:10.1371/journal.pone.0169605.

	4. Das, S. (2017). Modeling of Gene Regulatory Networks Using State Space Models. *Curr Trends Biomedical Eng & Biosci* 4(5): 555646. doi: 10.19080/CTBEB.2017.04.555646.
3.	*Project 2.* Development of gene selection approaches for classification of crop gene expression data. Publication: 1. Das, S., Rai, A., Mishra, D.C. and Rai, S.N. (2018). Statistical approach for selection of biologically informative genes. *Gene*, 655, 71-83. doi: 10.1016/j.gene.2018.02.044
4.	Other Publications: 1. Das Samarendra, Paul A.K., Wahi S.D., Pradhan U.P. (2017). Comparative performance of imputation methods for different proportions of missing data in classification of crop genotypes. *J Indian Soc Agric Stat* 71(2), 147–153. 2. Behera S.K., Paul A.K., Wahi S.D. ,Iquebal M.A., Das Samarendra, Paul R.K., Alam W. and Kumar, A. (2014). Estimation of heritability of mastitis disease using moment estimators*. Int. J. Ag. Stat. Sci.,* 10 (1), 243-247. 3. Paul A.K., Paul R.K., Das Samarendra, Behera S.K and Dhandapani A (2015). Non-parametric stability measures for analysing non-normal data. *Indian Journal of Agricultural Sciences*, 85(8):1097-1101. 4. Raman R.K., Paul A.K., Das Samarendra, Wahi, S.D. (2015). Empirical comparison of the performance of linear discriminant function under multivariate non-normal and normal data. *Int. J. Ag. Stat. Sci.* 11(2): 403-409. 5. Das Samarendra, Paul A.K., Wahi S.D., Raman R.K. (2015). A comparative study of various classification techniques in multivariate skew-normal data*. J. of the Ind. Soc. of Ag. Stat.* 69(3), 271-280. 6. Kumar, P., Bhar, L.M., Paul, A. K. Das, S., and Roy, H.S. (2018). Development of Composite Stability Measure by using Multi Criteria Decisions Making (MCDM) Techniques. *Journal of the Indian Society of Ag. Stat.* (Accepted) 7. Das S, Chhuria S, Rouchka EC, Rai SN. (2020). A Computational Network Biology Approach to Understand Salinity Stress Response in Rice (Oryza Sativa L.). *Bioinform. Int*., 1(1): 1003. 7. Das Samarendra, Bhatacharya A., Rai S.N., Kantardzic M. (2019). Early Hospital Readmission among Patients with Diabetes through Predictive Data Mining. *Current science* (Under review). 8. Das Samarendra, S Chhuria, SP Rai and AK Paul. (2019). Comparison of Descriptive Data Mining Approaches to Detect Gene Clusters using Rice (Oryza sativa L.) Gene Expression Data. *The Journal of the Indian Society of Agricultural Statistics*. (Under review). 9. Malhotra, A., Das, S. and Rai, S.N. (2020). Analysis of Single-Cell RNA-seq Data from Adenocarcinoma Cell Lines: A Stepwise Guide. *Evolutionary Bioinformatics*. (under review).