Distribution Agreement

Signature:

Kenong Su Date

Statistical Modeling and Learning in Single Cell RNA Sequencing Data

By

Kenong Su
Doctor of Philosophy

Computer Science and Informatics

Hao Wu, Ph.D.
Advisor

Peng Jin, Ph.D.
Committee Member

Zhaohui Qin, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Modeling and Learning in Single Cell RNA Sequencing Data

By

Kenong Su
B.S., University of Iowa, IA, 2016

Advisor: Hao Wu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Abstract

Statistical Modeling and Learning in Single Cell RNA Sequencing Data

By Kenong Su

The Single-cell RNA-sequencing (scRNA-seq) has emerged as a powerful tool to explore the biology at the unitary resolution of life. It has successfully deepened our understanding of various biological problems such as cell populations, gene regulations, and cellular transcriptional states. It also opens a door for investigating complex biological systems such as brain regions and immune responses. Furthermore, it leads to the discovery of new and rare cell types, which benefits for identifying drug targets and decoding disease etiologies in clinical studies.

Even though researchers are inspired by the success of the scRNA-seq, there still exist difficulties with the respect to the data analysis. Specifically, in the scRNA-seq gene expression profiles, the sparsity of excessive zero expressions, the heterogeneity across and within cell types, and confounding batch effects together contribute to the analytical challenges. To deal with these concerns, we have developed algorithms and pipelines for different research aspects in scRNA-seq data.

With the advance of high-throughput techniques, nowadays we are able to perform transcriptome sequencing for a massive number of cells experimentally. To facilitate the analysis on the large-scale scRNA-seq data, one commonly performed task is cell clustering, which enables the quantitative characterization of cell types. An essential step in scRNA-seq clustering is to select a set of most representative genes (referred as “features”) whose expression patterns will be adopted for proper cell clustering. Currently, almost all existing scRNA-seq clustering tools include a simple unsupervised feature selection step (e.g., statistical moments of gene-wise expression distribution) and uses random top number (e.g., 1000) of features for clustering. Therefore, it is more reasonable to designate a rigorous approach for better feature selection.

We created an algorithm named FEAture SelecTion (FEAST) specifically designed for selecting the most informative genes in the context of scRNA-seq clustering. We demonstrated that applying FEAST can significantly improve the cell clustering accuracy, and outperformed other feature selection methods embedded in the state-of-art scRNA-seq clustering methods such as Seurat and SC3.

Furthermore, determining the sample size for adequate power to detect statistical significance is a crucial step at the design stage for high-throughput experiments. Due to the unique sparse and heterogeneous characters presented in scRNA-seq, there are few tools explicitly designed for scRNA-seq experiments to address this topic. We developed POWSC pipeline, a simulation-based approach to provide power evaluation and sample size estimation in the context of differential expression (DE) analysis. POWSC provides a variety of power evaluations including stratified and marginal power analyses for DE genes characterized by two forms (phase transition or magnitude tuning), under different comparison scenarios. Additionally, we also devised the POWCLUST workflow as an extension of POWSC with a focus on assessing power for clustering. POWCLUST is able to recover the underlining information for cell type hierarchies and cell type proportions with a proper sample size estimation.

Overall, I designed new algorithms and pipelines including FEAST and POWSC for accurately selecting features and adequately evaluating power in scRNA-seq. We showcase that FEAST can assist to find more representative genes and POWSC can potentially be served as a guideline for scRNA-seq experiment design.

Statistical Modeling and Learning in Single Cell RNA Sequencing Data

By

Kenong Su
B.S., University of Iowa, IA, 2016

Advisor: Hao Wu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Acknowledgments

First of all, I would like to give special thanks to my advisor, Dr. Hao Wu, who has patiently and tirelessly guided me through my Ph.D. studies. Only with his solid supports, can I step into this very popular research field of single cell, which requires both knowledge of computer science and statistics. Those research topics conceived by him also intrigues me most. I have been inspired when having discussions with him from our weekly meetings and benefitted from his statistical instinct and sense of data. Additionally, I have learned a lot from his integrity and rigorous mentorship. He teaches me not only academic knowledges but also writing and communication skills, which makes him far beyond just an advisor to me.

Secondly, I am sincerely thankful for my committee members including Dr. Zhaohui (Steve) Qin, and Dr. Peng Jin. Dr. Qin led me for my first rotation project and helped me dive into bioinformatics research area. His valuable suggestions and persistence on collaboration escalated our project into another level. Dr. Jin kindly supported my research financially through the years which was such a tremendous help for me and my family. More importantly, his insightful input from a biological perspective motivates me to reveal interesting stories behind the datasets. Moreover, I would like to express my appreciation to the Computer Science Department at Emory for providing both study resources and research opportunities. I would like to thank my current and former colleagues from Dr. Wu's lab. I would like to thank all my collaborators: Dr. Yulin Jin, Dr. Zhijin Wu, Dr. Ronglai Shen, Dr. Shiyong Sun, Dr. Carlos S. Moreno, Dr. Xiaoxian Li, Dr. Tianwei Yu, and Dr. Xuyu Qian, and many others. I am privileged to contribute to the interesting biological discoveries via the close collaborative works.

Last but not least, I would like to deeply thank my parents, wife, and son who always love and support unconditionally, and back me up whatever difficulties I have encountered.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Single-Cell RNA sequencing and challenges

Since the first description of single-cell transcriptome analysis in 2009 [139], single-cell sequencing techniques have aroused people's attention for its ability to dissect individual cell, so as to profile genome, transcriptome, epigenome, and proteome. In 2013, single-cell RNA and DNA sequencing techniques were highlighted as the "Method of Year" [1]. Later, single-cell multi-omics and spatial transcriptomics were then selected as the "Method of Year" in 2019 [141] and 2020 [104] respectively. Among these single-cell related sequence research areas, single-cell RNA sequencing (scRNA-seq) is the most active study field, and it has been successfully widely used for studying complex tissues [26, 85, 91, 100, 165] and diseases [8, 14, 30, 109].

A typical workflow for scRNA-seq experiment contains the following steps. The first step is the isolation of the cells from the prepared issue of interest. Then, isolated cells are lysed for effectively capturing the mRNA molecules. Next, processed mRNA will convert to complementary DNA (cDNA) through revise transcription followed by the amplification. The amplified cDNAs usually undergo nucleotide barcode-tagging as sequencing library preparation, and lastly, they are pooled and profiled via sequencing platforms. To date, a variety of scRNA-seq techniques and protocols have been developed, which can generally fall into two categories based on the captured transcript coverage: full-length sequencers such as Smart-seq2 [113] and MATQ-seq [128], and 3'-end or 5'-end sequencers such as Chromium [178] and Drop-seq [102]. Recently, the latter droplet-based sequencers have gained popularity because they can encapsulate thousands of cells which presents unprecedent opportunities for studying cell populations. However, one caveat for scRNA-seq experiments especially from the droplet-based sequencers is the low volume of mRNA products extracted from individual cells which consequently impact the data analytical procedure.

Considering the unique characters in the scRNA-seq expression data, it presents three major analytical challenges including high sparsity, heterogeneity, and confounding batches. The sparsity meaning the excessive zero expression values observed in the expression mainly results from biological effects and technical noises. Some imputation methods have been developed to tackle the increasing sparsity issue [59]. Due to the cell-to-cell variation from between and within cell types, the cell populations demonstrate high heterogeneity especially in the complex tissues. In terms of expression patterns, bimodal and multi-modal distributions level are widely acknowledged which requires complicated statistical mixture models. Moreover, the cell types usually structure hierarchically [162] which adds to the level of heterogeneity. Even though the advancing scRNA-seq technique allows to sequence thousands of cells at one time, many experiments are performed under different conditions, platforms, patients, and timepoints. These confounding factors known as batches gain the difficulty for data analysis, numerous methods are specifically designed to address this concern [143]. Researchers need to properly handle these three unique characters when delivering data analysis with respect to different scRNA-seq research topics.

With the revolution of scRNA-seq experiments accessible to academic laboratories and commercial customers, many pipelines have been designed to address the aforementioned concerns when conducting data analysis. However, this field is still at the infancy status, currently, there are no gold standard rules widely acknowledged for data analysis. For example, since the imputation of zero expression would falsely enhance clustering signals, there is still lack of agreement of which imputation method can least pose circularity that introduce false-positive results [4, 173]. Furthermore, with the respect to modeling the gene expression distributions, because of distinct sequence platforms, there is no consensus about the best choice for the probabilistic density that can lead to the best sensitivity. Thus, researchers need to make their best judgements to properly deal with these factors for different purposes of data analysis.

1.2 ScRNA-seq clustering analysis

ScRNA-seq clustering has become the routine step for scRNA-seq analysis, which lays the foundation for identification of existing and novel cell types [69, 156], and classification of subpopulations from complex tissue systems [171, 174]). The essence of single cell clustering is to apply statistical learning approach to characterize gene expression patterns of cells and label them into putative cell types.

To date, lots of unsupervised computational algorithms have been developed for single cell clustering [36, 131, 83]. An overview of the workflow [79] involves the following components: quality control, normalization, feature selection, dimension reduction, and the core step of clustering. Different scRNA-seq clustering algorithms have at least one step difference by adopting their own strategies. However, it is noting that each strategy has its own strength as well as limitation. Thus, these established clustering algorithms could be sensitive and resulting into rather dissimilar clustering outcomes [43, 63, 168]. It is important to balance the clustering outcomes and optimize the clustering results.

Several other concerns remain for single cell clustering. First, there is no universal agreement of how to define a cell type particularly for the rare cell type based on gene expression patterns. How to reasonably combine biological aspect (e.g., known marker genes) for determining a cell type is worth exploring. Second, sometimes there is no gold-standard for evaluating the clustering outcomes especially for the cases that the truth (i.e., true cell labels and the number of cell types) is unknown. Third, the increasing size of scRNA-seq data ranging to millions of cells can hinder the computational scalability.

Except for unsupervised approach, some recently designed methods using prior knowledge of the similar tissue systems can also achieve the goal of cell clustering. The supervised approach is essentially to map the cell types from the existing dataset to the given dataset depending on the closeness of expression patterns. These approaches can be generally categorized into two ways: reference-tree-based approach [31, 78] and transfer-learning approach [61]. It is rational to choose the proper source dataset in order to correctly assign cell types in the target dataset.

Overall, scRNA-seq clustering presents promising opportunities as well as analytical challenges. We need to consider both computational and biological aspects to benefit the single cell clustering analysis.

1.3 ScRNA-seq power evaluation

At the design stage of scRNA-seq experiment, it is needed to determine the desired number of cells (known as sample size) to be sequenced with the consideration of the sequencing platforms, sequencing cost, and total read counts. With these factors, providing comprehensive and detailed power assessments is crucial for optimizing the design of scRNA-seq experiment. The power evaluations under the context of differential expression (DE) will confidently guide the researchers for detecting cell-type-specific marker genes, which is an important task for conducting the experiment.

Even though numerous methods and tools are available for sample size calculation for microarray and RNA-seq in the context of DE, this topic in the field of scRNA-seq is understudied. Moreover, the unique data characteristics present in scRNA-seq such as sparsity and heterogeneity increase the challenge. Recently, it starts to gain people's attention and a few statistical frameworks explicitly designed for scRNA-seq have just come out [134, 94]. Each pipeline recruits its own model assumption for estimating key parameters and simulating data, but there is still lack of systematic evaluation and comparison.

Currently, all most all existing power evaluation tools of DE analysis are constructed with the focus major cell types. It is more rigorous to estimate the power for sub-populations [28] and sub-cell-types. However, the unexchangeable hypotheses tests and potential batch effects gain the complexity of statistical inference. Furthermore, it is also necessary to accurately estimate power and sample size with the respect to the clustering, which is directly related to recovering the reference cell type structures.

Overall, scRNA-seq power estimation is central to the scRNA-seq experiment design. Even though many challenges presented in this topic, it is still worth exploring and the findings can potentially guide for conducting the scRNA-seq experiments.

1.4 Outline

In this dissertation, I present some statistical methods for analyzing scRNA-seq data with the different research perspectives. In chapter 2, I concentrate on scRNA-seq clustering and comprehensively review the feature selection methods included in the existing scRNA-seq clustering methods. I propose a method named FEAST [135], which is designed for selecting an optimized feature set before the core of scRNA-seq clustering. FEAST can serve as a plug-in for the well-established scRNA-seq clustering methods. I will discuss the logistic of FEAST and demonstrate its ability to improve the clustering accuracy. In chapter 3, I focus on scRNA-seq power evaluation and sample size estimation. I created a simulation-based pipeline named POWSC [134], which uniquely combines two forms of DE genes. POWSC can provide a variety of power evaluations including stratified and marginal power. In addition, it offers strategy for optimizing the tradeoffs between sequencing depth and sample size. Besides scRNA-seq, in chapter 4, I list a cancer genomics project by introducing a unique statistical model iPath. Ipath is able to identify highly predictive biomarkers for clinical outcomes, including overall survival, tumor subtypes, and tumor stage classifications. In chapter 5, I outline several potential research directions that I want to pursue in the near future.

Chapter 2

FEAST: Accurate Feature Selection Improves Single Cell RNA-seq Cell Clustering

Algorithm 1: FEAST

2.1 Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have revolutionized biological research [81, 178, 102]. Unlike the traditional bulk RNA sequencing (RNA-seq) that measures the average expression of large number of cells, scRNA-seq profiles the transcriptome of individual cells, which provides data with higher resolution for better understanding the transcriptomic regulation and variation at cellular level. It has been successfully applied to study many complex biology systems such as immune system [66], cerebral cortices [37], and tumor progressions [112]. In addition to the traditional expression analysis in bulk RNA-seq, scRNA-seq provides information to answer many new biological questions, such as discovering novel and rare cell types [89], and constructing pseudotime cell trajectories [19].

The scRNA-seq experiments usually generate expression profiles for large number of cells. For example, the 10x Genomics sequencer can profile thousands to millions of cells at a relatively low cost. One of the most important goals for scRNA-seq data analysis is the cell clustering, which is to partition cells into multiple groups via unsupervised clustering algorithms. Cell clustering provides important information for the cell composition and cell type specific transcriptome in complex tissues. It lays the foundation for downstream analyses such as differential expression, pseudotime construction, and new/rare cell type discovery. There are many methods and tools developed for unsupervised cell clustering [77, 123, 67], and they have been comprehensively reviewed and compared [114, 84, 36]. These methods usually start with a matrix of gene expression and output the grouping of cells. Many algorithmic factors can affect the performances of the cell clustering methods, including data preprocessing [155], normalization [9, 99], feature selection, dimension reduction [142], cell-to-cell similarity calculation, etc. Among them, feature selection is an important step which could have significant impact on the overall performance of cell clustering.

Although feature selection is implemented in most scRNA-seq clustering tools, it is not clear how different selection procedures will impact the results. Despite some efforts have been made to systematically compare and evaluate methods for data normalization [52, 25], dimension reduction [137], and cell similarity metrics [75] in scRNA-seq, there is no study specifically focused on the impact of feature selections. In this work, we comprehensively evaluate and compare the impact of feature selection on cell clustering in scRNA-seq. To the best of our knowledge, this is the first work to systematically evaluate and compare the impact of feature selection approaches on cell clustering accuracy. In addition, we also develop an algorithm, named FEAST (FEAture SelecTion), which selects representative genes for scRNA-seq cell clustering.

We compare FEAST with the feature selection approaches implemented in existing clustering tools through extensive benchmark tests. We demonstrate that FEAST can select more representative features than other approaches. Moreover, we demonstrate that using features selected FEAST with existing clustering tools can significantly improve the clustering accuracy.

2.1.1 Feature selection in scRNA-seq cell clustering

The scRNA-seq experiment produces expression levels for the whole transcriptome. A majority of the genes are not differentially expressed among different cell types; thus, they contain no information for cell clustering. The feature selection step selects a subset of genes best representing the structures of the dataset in a lower dimensional space, which enhances the signal to noise ratios and subsequently improves the cell clustering results. Since the cell grouping is unknown before clustering, the feature selection has to be done in an unsupervised fashion. Simple metrics based on quantities related to the statistical moments of the gene expressions are often used in most methods. We conduct a comprehensive review on the feature selection algorithms in existing cell clustering methods, summarized In Table 2.1. To be specific, both Seurat [123] and PanoView [62] first groups genes into 20 bins according to the mean expressions, and then selects the most variable genes, termed as highly variable genes (HVGs), within each bin. SC3 [77] filters out ubiquitous and rarely expressed genes to retain informative genes based on mean expression levels and dropout rates. Monocle [144]] selects genes based on minimum mean expression and variance. SCANPY [160] identifies a set of HVGs by using normalized dispersions in the preprocess across different batches. scVI [98] selects top ranked gene by variance. TSCAN [67] finds featured genes by considering both dropout rates and coefficient of variation (CV). SAIC [167] first filters out low-expressed genes and selects genes deviated from the fitted loess regression between CV and mean. SCENT [28] retrieves a set of most

scRNA-seq clustering method	**Quantities used for feature selection**
Seurat	μ and ϕ
PanoView	μ and ϕ (similar to Seurat)
SC3	μ and δ
Monocle	μ and σ^2
SCANPY	ϕ
scVI	σ^2
TSCAN	δ and CV
SAIC	loess regression between μ and CV
SCENT	SVD
SOUP	Gini index and $SPCA$
FiRE	μ, δ, and ϕ
SINCERA	μ, δ, and cell-specific index
RaceID3	Second-order polynomial between σ^2 and μ

Table 2.1: Feature selection methods implemented in different scRNA-seq clustering algorithms. Mean is denoted as μ. Variance is denoted as σ^2. Dispersion is denoted as ϕ. Coefficient of variation is denoted as CV. Dropout rate is denoted as δ. $SPCA$ means the sparse PCA algorithm. SVD means the singular value decomposition.

variable genes by singular value decomposition (SVD). SOUP [179] obtains most informative genes from two approaches: Sparse PCA (SPCA) [181] algorithm and Gini index, which is also adopted in DESCNED [154]. FiRE [69] first filters out genes with low expression levels and high dropout, and then selects top 1000 genes with the largest normalized dispersions. SINCERA [50] also first removes genes with low expression and high dropout, and then defines a cell specificity index based on the scaled expression to further filter out uninformative genes. RaceID3 [45] finds the featured genes exceeding the estimated variability from the fitted second-order polynomial functioning on the mean.

In addition to these moment-based approaches, there are other relatively more complicated methods for feature selection in high dimensional data. For example, Laplacian Scores [55] evaluate the feature importance by constructing local weighted graph. Moreover, some unsupervised approaches can be modified as the supervised approaches assuming the cell grouping is known. For instance, both Fisher Scores [38] and F-test statistics assess the efficiency of discrimination based on the fractions of between-group variance and within-group variance. If initial cell partitions are pre-determined, one can use statistical test based approaches such as Fisher Scores and F-statistics to select the significant features. When preparing the manuscript, we found a method named FEATS [147] just came out recently. FEATS uses F statistics to rank the features and optimizes a feature set by using silhouette coefficient [119] based on the initial hierarchical clustering outcomes.

2.1.2 Feature evaluation in scRNA-seq cell clustering

How to evaluate the quality of the feature set is another important problem. A straightforward assessment is the clustering accuracy if the reference labels (true classes for cells) are available. There are several metrics for clustering accuracy: adjust Rand index (ARI) [63], normalized mutual information (NMI) [133], Jaccard similarity index [88], Fowlkes–Mallows index [42], normalized information distance (NID) [12], and purity [126]. Without reference labels, it is more difficult to validate the quality of the selected features in an unsupervised manner. In this case, one can resort to a "pseudo-supervised" way, that is, to look at the "separation" of the clusters from the result based on selected features. The separation can be defined based on the average distance among the cluster centroids, or the mean squared distances between individual cells and the cluster centroids, or the combination of them. A set of features is deemed better if it leads to clusters with larger between-group and smaller within-group distances.

2.2 Method

2.2.1 Preprocess and normalization

We preprocess the raw gene expression data as the followings. First, genes with all zero read counts and low expression rates (σ) are filtered out. The default threshold for σ equals to 2 divided by the total number of cells. It is common to observe some genes are only expressed in very few (one or two) cells in 10x and inDrop data, which are not informative for cell clustering. We do not remove the ubiquitously expressed genes and use a relatively conservative threshold for σ because we intend to keep more features for further selection. Next, we normalize the count matrix by cell-specific size factors, which are calculated based on the sequencing depths, and take a log2 transformation on the normalized counts.

2.2.2 The consensus clustering

With the preprocessed gene expression matrix (Y) of G genes and N cells, FEAST utilizes the cluster-based similarity partitioning algorithm (CSPA) [133] to create a consensus matrix. Specifically, FEAST first performs principle component analysis (PCA) to obtain a sequence of principle components (PCs). For each of the top-i (i= 2,3,...) PCs, FEAST fits a Gaussian Mixture Model (GMM) to cluster the cells into k groups. Each clustering result is represented by a binary $N \times N$ matrix, where the corresponding cell unit is 1 if two cells belong to the same cluster, and 0 otherwise. By default, FEAST examines till top 10 PCs because we purposely cover a relatively large number of PCs to account variabilities from different directions in the covariance matrix. Next, FEAST construct a consensus matrix by averaging all the similarity matrices. The final clustering labels are obtained through fitting another round of GMM on the consensus matrix. Only cells with posterior probability of belonging to a cluster greater than 0.95 are kept in the final clusters.

The consensus clustering is similar to the procedure implemented in SC3. It only retains cells that are tightly clustered together and excludes the ones whose cluster membership cannot be determined with high confidence. As shown in the Result section (Figure 2.2), this step enhances the signal in the data, which subsequently helps identifying features.

2.2.3 Gene-level significance inference

After obtaining the consensus clusters, selecting the most representative features becomes a supervised feature selection step. FEAST uses F-statistics to test the feature significance because it can summarize the differences among multiple groups into a single number. F-statistics essentially calculates the fraction between between-group variance (varb) and within-group variance (varw). Noticeably, F-statistics is similar to Fisher-scores which was initially developed as the estimation of variance ratios. Mathematically, the F-statistics calculation for the g^{th} gene is denoted as in equation (2.1).

$$F_g = \frac{(varb_g)/(df_1)}{varw_g/df_2} = \frac{(varb_g)/(df_1)}{(vart_g - varb_g)/df_2} \tag{2.1}$$

Here, df_1 and df_2 are degrees of freedoms calculated as $K - 1$ and $N^{'} - K$ respectively where $N^{'}$ is the total number of cells in the consensus clusters ($N^{'} \leq N$). FEAST uses the difference between total variance ($vart_g$) and between-group variance to represent within-group variance, where $varb_g$ is calculated as $\sum_{i=1}^{K} n_i \times (\bar{Y_g} - \bar{Y_{gi}})^2$ and $vart_g$ is calculated as $\sum_{j=1}^{N'} (\bar{Y_g} - Y_{gi})^2$. $\bar{Y_g}$ is the average expression for g^{th} gene, and Y_{gi} is the expression value for the g^{th} gene and i^{th} cell. $\bar{Y_{gi}}$ and n_i denote the mean and sample size for the i^{th} cluster respectively.

2.2.4 Determine the optimized feature set

Unsupervised feature set validation is challenging without a properly predefined optimization criterion. FEAST uses the MSE to evaluate the clustering results. The MSE represents the average distances between cells and the cluster centroids, which is a good representation for the goodness of fit. To be specific, with the obtained clustering labels, FEAST fits simple linear regression between the normalized gene expression and the clustering outcomes, Then, FEAST computes the MSE from the regression residuals, which represents the mean squared distance of each data point to its assigned cluster center. For each clustering outcome with different feature set, FEAST calculates an MSE. The feature set associated with the smallest MSE is recommended as the optimal feature set.

The feature selection in clustering is similar to the variable selection problem, i.e., one tries to identify a subset of variables to best predict the classification outcomes. Since the clustering is unsupervised, it is difficult to evaluate which set of variables is the best without knowing the outcome. In this case, MSE, which represents the model fitting, is a reasonable choice for evaluating the variable selection result. It is worth noting that the MSE is calculated from all genes and all cells, even though the features are selected based on a subset of cells and the predicted cluster is based on a subset of genes (the selected features). This ensures fair comparisons for different clustering outcomes and avoids over-fitting of the data. Our real data analyses demonstrate that this approach can select an optimal set of features, i.e., the feature set with the smallest MSE usually corresponds to the best clustering results.

2.3 Result

We comprehensively evaluate several existing scRNA-seq clustering methods in a number of datasets (Supplementary Table A.1) and find that feature selection has significant impact on the cell clustering results. To better assist existing scRNA-seq clustering algorithm, we develop the FEAST framework (https://github.com/suke18/FEAST) that produces a representative feature set to improve the clustering accuracy. To provide a quick summary, FEAST first performs a consensus clustering to get initial cell clusters. Features are then ranked and selected based on the initial clusters. Optimal number of features is determined by the fitness of the clustering results from different numbers of top features. The output of FEAST is a list of features that can be fed into the existing cluster methods. We systematically compare features selected by FEAST with other unsupervised feature selection methods implemented in existing cell clustering tools. We demonstrate that the FEAST can identify more representative features and significantly improve the clustering accuracy.

2.3.1 Overview of FEAST

FEAST is a tool solely designed for scRNA-seq feature selection, and works with any existing cell clustering method. Users can use FEAST to replace the feature selection step provided in existing cell clustering methods and obtain improved results. The FEAST workflow includes three major steps, as illustrated in Figure 2.1. First, it implements a computationally efficient algorithm to obtain a consensus cell clustering (Figure 2.1A). This unique consensus clustering step allows the detection of the most confident cell clusters, which improves the feature selection in the next step. Second, based on the consensus clusters, it calculates the significance for each feature via F-test and ranks the features according to the F-statistics (Figure 2.1B). Third, it finds an optimal feature set through a feature evaluation algorithm (Figure 2.1C). We provide detailed description for each step in the Method section.

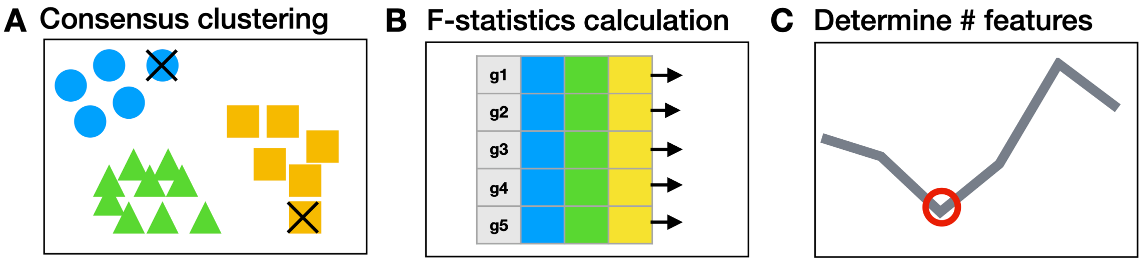


Figure 2.1: The overall FEAST workflow. FEAST includes three major steps: (A) it performs consensus clustering to find clusters with high confidence, the cell that are less correlated with the clusters are filtered out as indicated by the "×". (B) it calculates the feature significance based the initial clusters. (C) it determines the optimal size of the feature set through a validation process.

2.3.2 Datasets

We collect 12 public scRNA-seq datasets (Supplementary Table A.1) for evaluating the impact of features selection on clustering and benchmarking the performance of FEAST. These datasets are obtained from different sources, including https://hemberg-lab.github.io/scRNA.seq.datasets, https://portal.brain-map.org/atlases-and-data/rnaseq, and Gene Expression Omnibus from the National Center for Biotechnology Information (NCBI-GEO) [11]. It is noted that the cell type information for these collected datasets are either obtained by experimental validation such as fluorescence activated cell sorting (FACS) or annotated by well-known cell-type-specific marker genes. All datasets include the raw count gene expression matrix as well as the cell type labels, which enable the evaluation and comparison of methods.

2.3.3 Consensus clustering improves the signal

As discussed before, feature selection in existing methods are mostly based on first and second moments of the gene-wise expression distribution. We found that this procedure can select wrong features, for example, a gene with high marginal variance can be caused by the large within cell type variation. We design an algorithm to convert the unsupervised feature selection problem into supervised fashion. To be specific, we first cluster the cells to generate initial clusters, and then detect features based on these initial clusters. The initial clustering from this approach plays an important role. A biased cluster will obviously lead to poorly selected features. FEAST implements a consensus clustering procedure (details in the Method section) to find clusters with high confidence, and then computes the feature significance based on the cells in the consensus clusters. Here we show that this consensus clustering step can improve the signals.

Figure 2 shows the distribution of the statistical significance of all genes when comparing their expression across clusters. As a comparison, we benchmark the

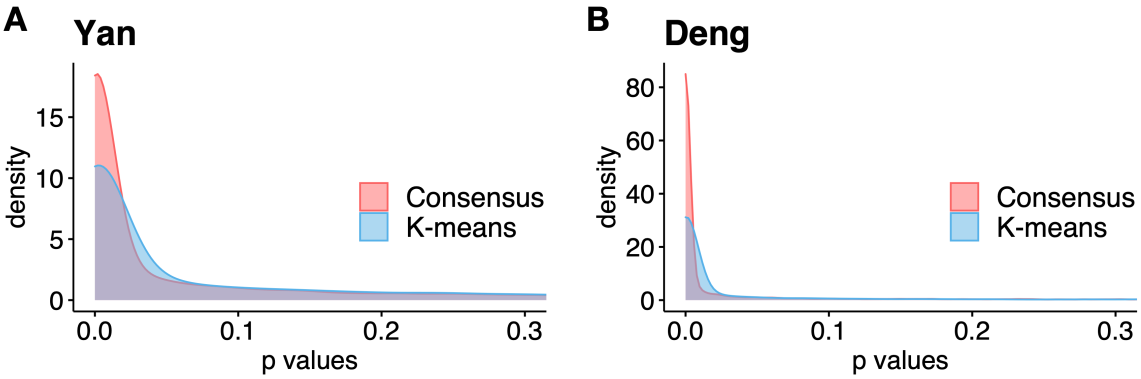


Figure 2.2: Consensus clustering improves the separation signals. Results are shown for two embryonic development datasets: Yan (A) and Deng (B). We use consensus clustering from FEAST and K-means to determine initial clusters. Then, we calculate the feature significance by F-test. The results demonstrate that the p-values from the consensus clustering are more significant.

results from using K-means to determine initial clusters. Results from two embryo development datasets Yan (Figure 2A) and Deng (Figure 2B) are shown. To be specific, we apply both K-means and consensus clustering on each dataset to obtain the clustering. Then for each gene, we perform F-test to compare the expression levels cross clusters. These figures show that the p-values from the consensus clustering in FEAST are more significant than those from K-means, that is, there are more genes with p-values closer to 0. Additionally, we investigate the distributions of F-statistics (Supplementary Figure A.1) from these two approaches, and obtain a similar finding that the consensus clustering can improve the separation signal by showing higher F-statistics values than K-means. These results demonstrate that the consensus clustering procedure provides “tighter” clusters and more distinctive features (ones that show greater difference among clusters).

2.3.4 FEAST selects features better than other unsupervised approaches

After obtaining the initial cell labels from consensus clustering, FEAST selects the top features based on F-test statistics. We systematically compare the top-m features generated by FEAST with other three feature selection procedures implemented in SAIC, SC3, and Seurat. Specifically, for SAIC, we select the genes that are most deviated from the fitted loess regression between CV and mean. For SC3, we filter out the rarely and ubiquitously expressed genes and select the top genes based on expression levels. For Seurat, we adopt the FindVariableFeature function inside the Seurat R package to select the top genes. We purposely fix the number of top features for each approach and evaluate the feature quality via cell clustering. Specifically, we select top-m (m = 500,1000,and 2000) and perform the clustering by SC3 on a series of test datasets (Supplementary Table A.1). It is noted that SC3 allows users to specify the input number of clusters. For the evaluation and comparison, we assume the number of the true cell types are known. We use the ARI value as metric to compare the cell clustering results with features selected from different methods.

These comparison results are summarized in Figure 2.3, where each panel represents a test dataset, and each group of bars corresponds to the ARI values from using a certain number of top (m =500,1000,and 2000) features. The results show that the FEAST has the best performance compared with other feature selection methods. Out of the 12 datasets, FEAST shows the highest ARI values in 11 of them. The performance gain can be substantial, for example, in Goolam, Treutlein and LGd data. Even in the Nestorowa data where FEAST result is not the best, its performance is comparable with other methods. The features selected by Seurat show the second-best performance overall. It also shows that genes selected by SAIC could lead to poor ARI values such as in Close, Treutlein, and Zheng datasets. Additionally, we also compare FEAST to the feature selection approaches implemented

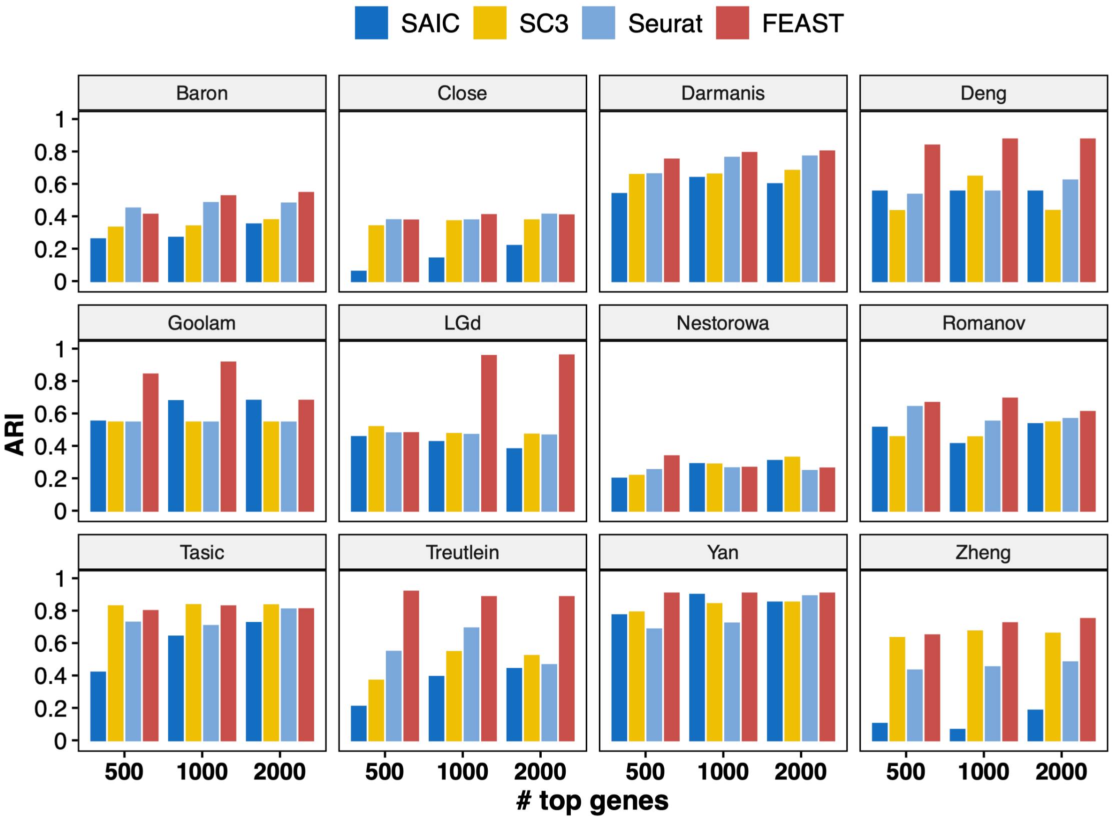


Figure 2.3: The comparison of the feature selection methods. We benchmark FEAST with other three unsupervised feature selection procedures implemented in SAIC, Seurat, and SC3. In each test dataset, we select the top 500, 1000, and 2000 features from each criterion followed by SC3 clustering. FEAST outperforms the other methods in almost all the scenarios by showing the highest ARI values in 11 out of 12 datasets.

in raceID3, scVI, and SOUP. It is demonstrated that features selected by FEAST lead to better cell clustering results compared to the features selected by the other approaches (Supplementary Figure A.2).

We further inspect the features selected by other unsupervised approaches including kurtosis and CV, and find that the top selected genes show extreme high expression in only a few cells while remaining the same (usually 0) in the rest of the cells (Supplementary Figure A.3). These are the ones with highly skewed expression distribution, and clearly not good features for clustering. These bar plots in Figure 3 also indicate that including more features does not necessarily lead to a better clustering performance; for example, the performances decrease from $m = 1000$ to $m = 2000$ in Goolam, and Romanov datasets. Overall, these results show that FEAST can select better features than the other approaches with respect to cell clustering accuracy.

2.3.5 FEAST optimize the feature set through validation

Above we show that FEAST outperforms other methods in top-m features. In addition to provide better ranking for the genes, a good feature selection method also needs to determine an optimal number of genes to be included in the final feature set. For the second part, FEAST implements a validation process to determine the number of features. Details of the method are provided in the Method section. Briefly, FEAST selects a series of top-m (m = 20, 50, 100, 200, 500, 1000, 2000, 5000, and all genes) features based on consensus clustering, and then conducts clustering using different number of features. Then, FEAST assesses the goodness of fit of the clustering results and determines the optimized number of features.

We benchmark the method on two datasets, the Zheng dataset which contains 8 well-annotated PBMC types, and Deng dataset which includes 6 adult liver cell types. In Figure 2.4 A and C, each curve represents a metric for evaluating the clustering results from SC3 under different number of top features. The conclusions from these metrics overall agree with each other. For example, in the Zheng data, with the increasing number of input features (m =50 to 1000), the clustering accuracy also increases. Specifically, the ARI increases from 0.33 to 0.74 and the NMI increases from 0.48 to 0.80. However, after reaching to the peak at m = 2000 (ARI = 0.75 and NMI = 0.81 respectively), the accuracy curve plateaus until using 5000 features, and becomes lower is using all genes. This indicates that including more features will not necessarily improve the clustering accuracy.

For many datasets where the true cell labels are unavailable, we adopt a criterion based on the mean squared errors (MSE) of clustering (details in Method section) to assess overall clustering fitness and select the optimal number of features. Figure 2.4 B and D shows the MSE values from the clusters based on different numbers of top features. We find that the MSE reaches the lowest level at m=2000 for the Zheng data, which matches the best clustering accuracy result in Figure 2.4 A. In the Deng data, we find the lowest MSE result is concordant with the best clustering accuracy at m = 1000. These results show that the MSE criteria works well in selecting the optimal number of features.

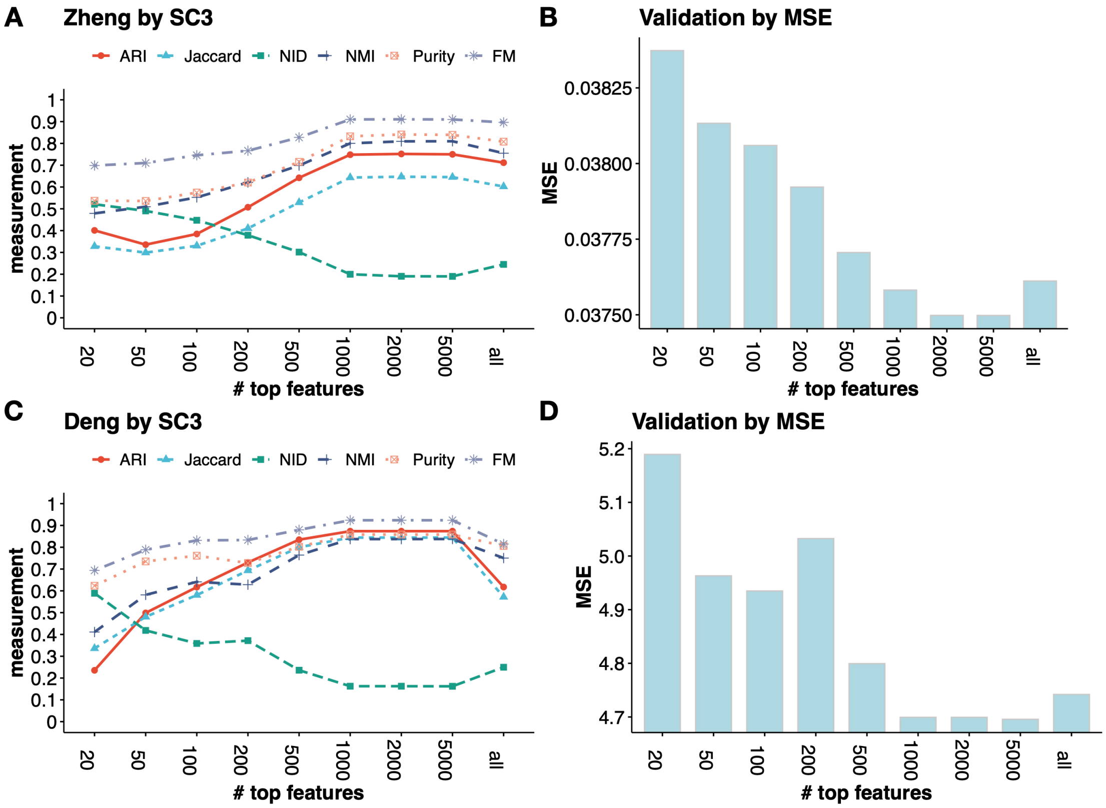


Figure 2.4: The validation process used in FEAST to determine the optimal number of features. In both Zheng and Deng datasets, FEAST selects the top-m (m = 20, 50, 100, 200, 500, 1000, 2000, 5000, and all genes) features, and performs cell clustering by SC3. For different m, (A) and (C) show the clustering accuracy measurements, (B) and (D) show the MSE which represents the goodness of fit of the clustering results. We find that the lowest MSE results (B and C) agree with the best clustering accuracy (A and C).

Additionally, we also perform the above analyses using TSCAN as the clustering method (Supplementary Figure A.4). We obtain similar findings that the optimized feature set in general matches with the validation procedure by MSE. It is noted that we utilize TSCAN or SC3 for clustering, which allow to specify the number of clusters (k). The user can also adopt their favorite scRNA-seq algorithm on the selected feature sets, but need to keep the same k for fair comparison and evaluation.

2.3.6 FEAST improves the clustering accuracy

We systematically evaluate the performance of FEAST on 12 publicly available scRNA-seq datasets (Supplementary Table A.1). These datasets cover a wide range of sample sizes (from tens to thousands of cells), as well as from different sequencing technologies such as smart-seq2 [113], 10x Genomics, and inDrop [80]. In each dataset, we utilize FEAST to select features, which are obtained through the MSE validation process of using the top-m (m =500, 1000, 2000) features. Then, we feed the optimal feature set into SC3 for cell clustering. We compare these results to the default setting in SC3, which selects features based on mean expression and dropout rates. The clustering ARI values from default SC3 and SC3 with FEAST features are summarized in Figure 2.5. For all datasets, features selected by FEAST results in better clustering ARI. In all 12 datasets, the ARI is increased by 0.19 on average, indicating a significant improvement. In some datasets, the ARI values increase dramatically with specified FEAST features. For example, in Goolam dataset the ARI values increase from 0.65 to 0.93. Similar improvements are also observed in Treutlein, LGd, and Deng datasets. To demonstrate the broad applicability of FEAST, we perform the same analyses using three other clustering methods: TSCAN, SHARP [152], and SIMLR [153]. We observe significant improvements of clustering accuracy in all methods. The results are summarized in Supplementary Figure A.5, A.6, and A.7.

Note that all above tests are well-controlled: the only difference between the blue and red bars is the feature selection procedure. Even though these clustering

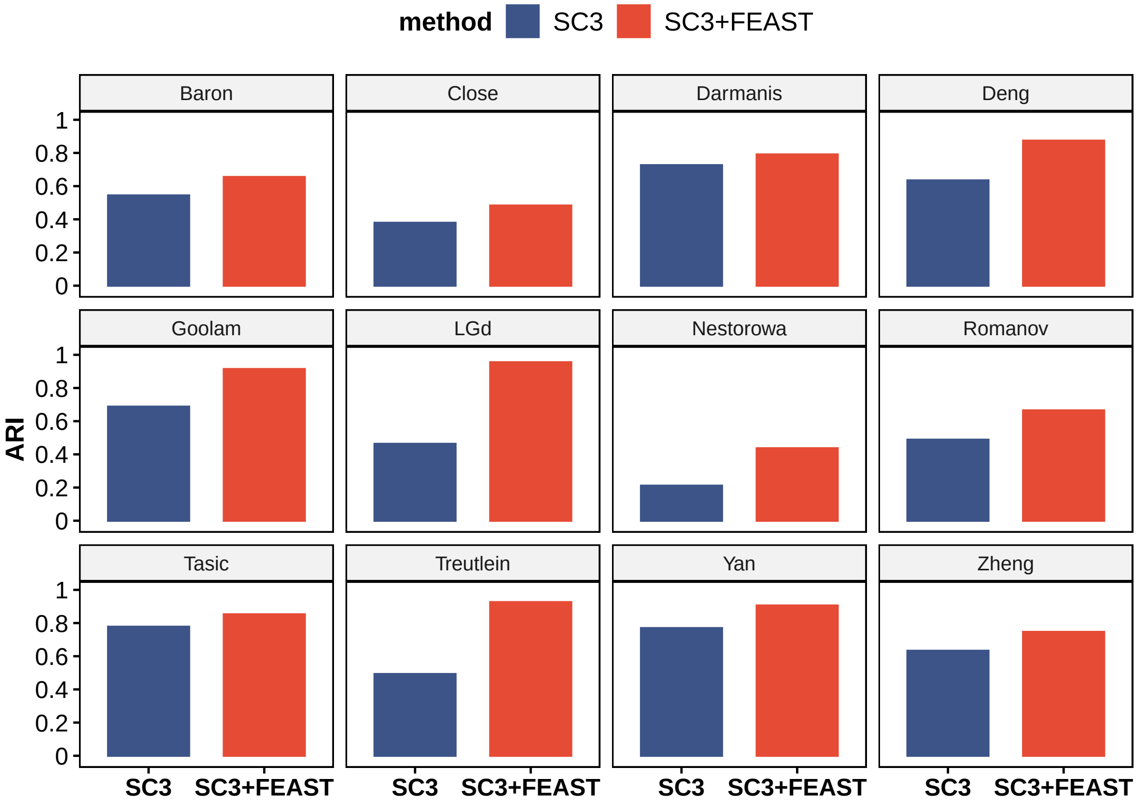


Figure 2.5: FEAST improves the clustering accuracy with existing method. The figures show ARI values for 12 public datasets. For each dataset, we compare the results from SC3 and SC3 with FEAST selected features. For all datasets, we observe significant improvement in ARI using SC3 with FEAST features.

tools implements different methods and perform differently at different datasets, we show that using features selected by FEAST can instantly improve the clustering accuracy. Taken together, we show the superior performance and broad applicability of FEAST, regardless of the clustering method, experimental protocol (full-length or 3' end sequencing), and size of the dataset.

2.3.7 Test FEAST on larger datasets

Furthermore, we test the performance of FEAST on relatively larger datasets. The purpose is to evaluate the computational scalability and the robustness of the algorithm when there are more cells and cell types. We analyzed three public datasets

(Supplementary Table A.2), which contains ∼28k cells and ∼28 cell types on average. For these tests we use SHARP as the clustering method since it's specifically designed for large dataset. Again, we observe significantly improved ARI values using the features selected by FEAST (Figure 2.6). These results suggest that FEAST is robust and efficient, and work well for large datasets.

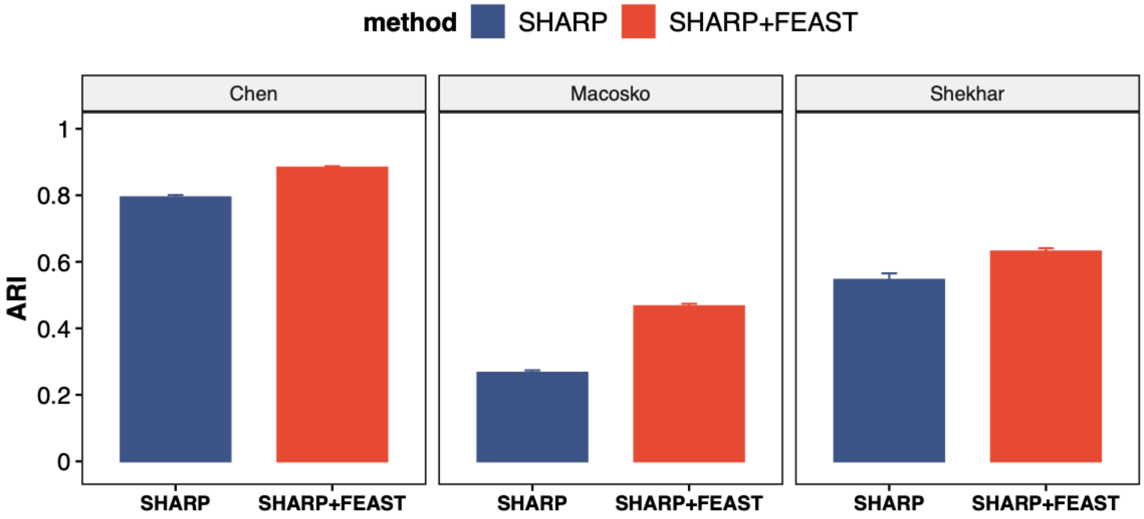


Figure 2.6: FEAST improves the clustering accuracy on the larger datasets. We investigate three datasets with ∼28k cells and ∼28 cell types on average. For each dataset, we compare the results from SHARP and SHARP with FEAST selected features. For all datasets, we observe significant improvement in ARI.

FEAST is implemented as an open-source R package and freely available at https://github.com/suke18/FEAST. As a feature selection tool, it can serve as a plug-in for established scRNA-seq clustering methods. FEAST offers excellent computational performance. We profile the computational performance of FEAST for a wide range of sample sizes (100 to 50,000 cells). Results are shown in the Supplementary Figure A.8. It is important to note that the computational burden does increase exponentially with the increasing number of cells, due to the first step of consensus clustering in the algorithm. However, with an efficient implementation, FEAST still provides excellent computational performance and will handle a majority of the tasks. For example, the feature selection step takes less than one minute for 10,000 cells and takes less than four minutes for 50,000 cells. The validation process requires running clustering for different number of top features; thus, its performance depends on the clustering method itself.

2.4 Discussion

In scRNA-seq clustering, selecting a desirable feature set before performing clustering is very important because the features will have significant impact on the clustering outcomes. Particularly, a feature set including excessive non-informative genes or lacking marker genes will result in poor clustering accuracy. Even though numerous clustering algorithms tailored for scRNA-seq have been developed and widely used in the community, the importance of feature selection step has not been thoroughly investigated. Currently, almost all clustering methods include a feature selection step, mostly based on thresholding some simple statistics, for example, to use the top 2000 highly variable genes, or to choose genes with low dropout rate and high average expression. It is unclear how much the feature selection will impact the cell clustering accuracy, and whether better selected features can improve the cell clustering result.

The major contribution of this work is two-fold. First, we carefully evaluate and compare the impacts of feature selection on cell clustering by comprehensive data analysis. Secondly, we design a new algorithm named FEAST for selecting an optimal set of features. FEAST can work as a plug-in tool for existing clustering methods. We systematically compare FEAST with other common feature selection methods, and demonstrate that FEAST outperforms other methods in selecting more representative features, which subsequently improves clustering accuracy. We show that the improvement brought by the FEAST features is not limited to the clustering method, i.e., we observe significant improvements using a number of existing cell clustering tools including SC3, TSCAN, SHARP, and SIMLR. These results show that researchers can first run FEAST to obtain a set of features then feed them to established scRNA-seq clustering algorithms, which will likely improve the clustering accuracy. Moreover, based on our experiences, selecting top 1000 or 2000 features from FEAST usually give satisfactory results. So, if computational time is a concern, we recommend users take top 1000 features as the final feature set.

Determining the number of clusters (K) is an important step in cell clustering. Some clustering software tools such as SC3, TSCAN, and CIDR provide function for estimating K, but the clustering functions in these tools require users to specify a fixed K. FEAST does not provide function for estimating K. It works merely as a feature selection tool for cell clustering, and the users need to provide K. On the other hand, users can use methods implemented in current software tools or prior knowledge to estimate K.

The current FEAST frame, similar to most other clustering methods, selects features based on the given dataset. It is possible to incorporate existing biological knowledge on marker genes into the feature selection algorithm. For example, we can impose a prior on the features and formula the problem in a Bayesian framework. In addition, even though the clustering put cells into several distinct, exchangeable groups, the cell types form a hierarchical tree in reality. With the consideration of such hierarchical structure, it might be better to use a different set of features at each branching point, and perform clustering in a top-down, step-wise manner. Furthermore, a new paradigm of cell type identification has recently gained much attention [31, 78]. Those methods don't cluster the cells. Instead, they assign each cell to a particular cell type, based on a reference panel. We believe feature selection will also play an important role for those methods, and FEAST can potentially be used to improve those methods. These interesting questions are all on our future research plan.

Chapter 3

POWSC: Simulation, Power Evaluation, and Sample Size Recommendation for Single Cell RNA-seq

Algorithm 2: POWSC

3.1 Introduction

Single cell RNA-sequencing (scRNA-seq) has emerged recently as a powerful technology to investigate transcriptomic variation and regulation at the single cell level [47, 151]. The traditional "bulk" RNA-seq pools RNA from a large number of cells and measures the average expression in a sample. scRNA-seq, on the other hand, profiles the transcriptome at individual cell level, which reveals cell to cell heterogeneity in transcription and provides more insights into understanding many important biological processes and disease etiologies such as early embryonic development [23], immunology [108], and tumorigenesis [60]. The scRNA-seq technology has drawn tremendous attention recently, and an enormous amount of data have been generated. These data present many opportunities as well as challenges to the developments of analytic methods. One fundamental challenge, similar to many other high-throughput genomics data, is to identify genes that are differentially expressed (DE) under distinct biological or clinical conditions. Over the last several years, a number of methods and tools for scRNA-seq DE analysis have been developed [73, 41, 163, 123], and comprehensively compared [131].

With the DE tools available, the sample size estimation has become an important question at the experimental design stage. The investigators would like to know the required number of cells and the sequencing depth, within a certain budget, in order to achieve the desired level of statistical power to detect the DE genes. Traditional power evaluation and sample size calculation methods often serve studies with a single primary end point, thus these deal with a single hypothesis test. They are not applicable to high-throughput data sets, which involve testing many unexchangeable hypotheses simultaneously. Recently, power evaluation and sample size recommendation for high-throughput data have attracted much attention along with the increasing application of the technology, and a number of methods and software tools have been developed for gene expression microarray [97, 157, 33] and RNA-seq [39, 54, 90].

Compared with the traditional single-hypothesis test setting, the DE problem in high-throughput data such as RNA-seq involves many parameters, including the baseline expression levels, within group variances, effect sizes, sequencing depths, type I error control with multiple testing adjustment, and more complicated testing procedures. These complexities make it very difficult to derive an analytical solution for power calculation. Thus, scientists often resort to simulation-based procedures to evaluate the power and provide sample size recommendation. To imitate real biological situation, most simulations introduce DE with various levels of effect sizes, ranging from near zero to substantial differences. The statistical power in detecting DE genes with near zero effects, as expected, is very low. The genes with minimal effects, though non-null by classical definition, are of little biological interest. Thus the concept of "targeted power", which is the probability of detecting DE with effect sizes exceeding a user-defined level, is introduced and applied to bulk RNA-seq experiments [161]. Another characteristic of RNA-seq data is that nuisance parameters, such as mean expression level, may affect statistical power. Inspecting stratified power may inform not only experimental design, but analysis plan as well, as filtering out genes in some strata may result in a higher true discovery rate.

Compared with bulk RNA-seq, scRNA-seq presents even more challenges and unique characteristics in DE test. First, bimodal [127] and multi-modal [82] expression at single cell level are widely observed thus DE may include a discrete transition of expression status as well as a continuous change in expression level. To reflect this observation, methods for detecting DE in these two forms, phase transition and magnitude tuning, are developed [41, 163]. The first form tests the differences in proportions of cells expressing a certain gene, and the second form tests the quantitative changes given the gene is expressed. Second, the cells profiled from a sample often contain a mixture of multiple cell types, and DE test can be carried out for each cell type in the mixture. This can be done as the following: first clustering cells into multiple groups and then performing DE. The cell mixture further complicates the power evaluation since the mixing proportion directly influences the statistical power: rarer cell types have fewer cells, hence with lower power detecting the same level of DE. Moreover, researchers might be interested in comparing different cell types in a mixture to identify marker genes [102, 171, 58, 110], and it is desirable to provide power assessment in such tests.

The power assessment for scRNA-seq experiment designs has gained some interests recently. To the best of our knowledge, there are two power assessment methods for scRNA-seq data: powsimR [149] and scDesign [94]. Both methods are simulation-based approaches: they simulate scRNA-seq expressions based on some template data, and then evaluate power in DE tests. Each method, however, has its own data model and power evaluation criteria. PowsimR assumes a negative binomial distribution for sequencing counts, while scDesign uses a gamma-normal mixture model for log counts. Both methods perform DE detection using existing methods, and then compare the results with the truth to evaluate power in different ways. PowsimR produces power stratified by expression levels, while scDesign outputs a number of power-related quantities (precision, recall, true negative rate, etc.) for detecting the top (with the default set at 1000) DE genes. Both methods concentrate on two-group comparison designs. In addition, there is another data simulator Splatter, which adopts complex step-wise procedures in simulating gene-wised means, enforcing dropouts, and removing outliers.

In this work, we developed a method, named POWSC, to provide comprehensive functionalities for power evaluation and sample size recommendation in scRNA-seq DE analyses. Note, unlike traditional RNA-seq where the number of biological replicates is regarded as sample size, POWSC refers the number of cells as the sample size because a single cell is the unit for scRNA-seq experiment. Compared with powsimR and scDesign, POWSC computes stratified targeted power for two forms of DE tests. In addition, POWSC considers the cell type mixtures in the data and provides comprehensive power evaluations for two DE test scenarios: comparing the same cell types from two biological conditions, and comparing different cell types under the same condition. Moreover, POWSC investigates the relationship between sequencing depths and cell numbers under the same total sequencing depth, and offers an optimal strategy.

3.2 Method

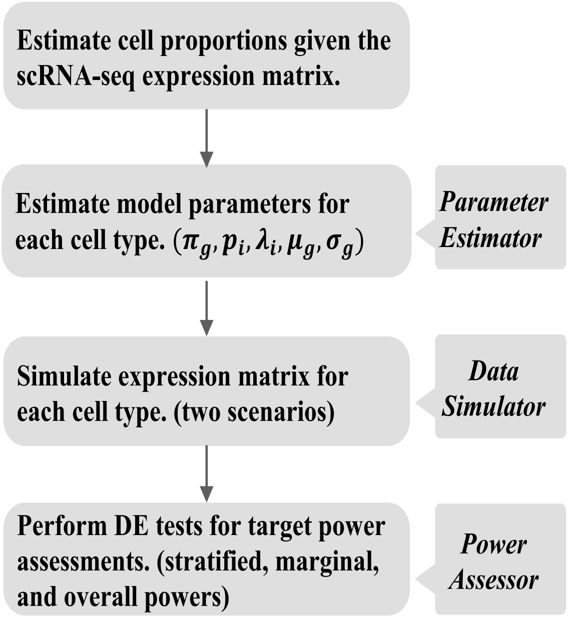

Estimate cell proportions given the scRNA-seq expression matrix.
Estimate model parameters for each cell type. ($\pi_g, p_i, \lambda_i, \mu_g, \sigma_g$)
Parameter Estimator
Simulate expression matrix for each cell type. (two scenarios)
Data Simulator
Perform DE tests for target power assessments. (stratified, marginal, and overall powers)
Power Assessor

Figure 3.1: The schematic overview of the POWSC pipeline.

POWSC relies on simulation to evaluate the relationship between the sample size and statistical power in scRNA-seq DE analysis. It simulates scRNA-seq data with known DE status, then runs DE analysis and evaluates the statistical power. POWSC contains three modules: *Parameter Estimator*, *Data Simulator*, and *Power Assessor*. First, a number of pre-specified model parameters, including the marginal distributions of gene expression, dispersion, sequencing depth, and etc. are accurately estimated by the *Parameter Estimator*. To mimic real scRNA-seq data, POWSC estimates these parameters from a pilot dataset provided by the user or selected from public databases. Users who do not have in-house pilot data may choose from one of several tissue types for which the model parameters are pre-calculated, including blood, brain, immune system, and tonsil. Given the parameters, the *Data Simulator* then generates scRNA-seq counts under different sample sizes. Last, *Power Assessor* will evaluate and report different types of power. Below we provide detailed description for each module. The pipeline diagram for POWSC is listed in Figure

3.1.

3.2.1 Parameter estimator

Given a preliminary scRNA-seq dataset, this module will estimate the parameters required by the simulator. The preliminary dataset is in the form of a matrix of sequence read counts. Denote the expression matrix as Y, which is a $G \times N$ matrix for G genes and N cells. Let Y_{gi} represents the count observed on g^{th} gene and i^{th} cell. Note that Y often contains cells with highly heterogeneous states such as different cell types, differentiation stages, disease progressions, etc. Estimating parameters directly from the data will bias the results; for example, over-estimate the gene-specific variances. Thus, the first step in the estimator module is to cluster the cells, and then perform parameter estimation for each cluster. In our implementation, POWSC can choose either Seurat [123] or SC3 [77] as the cell clustering tool. With cells properly clustered, POWSC estimates a number of model parameters to characterize the data in each cluster.

One special characteristic in scRNA-seq is the excessive observation of zero gene counts. It could be caused by two factors: biologically some genes are not expressed, or technically the gene expression levels are too low to be detected. To account for these zeros, we follow the data model presented in SC2P [163], which is a mixture of zero inflated Poisson (ZIP) and log-normal Poisson (LNP) distributions. This mixture captures the expression heterogeneity among cells for a particular gene. Specifically, ZIP describes the inactive transcription (Phase 1) from the background, and LNP represents the active transcription (Phase 2) from the foreground. This mixture model is written as the following Equation 3.1.

$$P(Y_{gi} = y_{gi}) = (1 - \pi_g)ZIP(y_{gi}|p_i, \lambda_i) + \pi_g LNP(y_{gi}|\mu_g, \sigma_g^2) \tag{3.1}$$

The parameters to be estimated in the data model include the following: p_i is the point mass for zero-inflation in the background signal; λ_i is the Poisson rate from background signals in unexpressed genes; μ_g and σ_g^2 are parameters in LNP for the distributions of active transcription levels; and π_g is the mixture proportion for Phase 2. The parameter estimation procedure is adopted from SC2P. Given a pilot dataset, the *Parameter Estimator* module outputs the cell-specific parameters in ZIP distribution, gene-specific parameters in LNP distribution, and the mixture proportion. These parameters characterize the distribution of the input data. In addition, we have pre-calculated the model parameters for certain cells for several tissue types, and these parameters can be easily accessed and used in the downstream simulation.

3.2.2 Data simulator

With model parameters, we are in place to simulate scRNA-seq data for DE detection and power evaluation. A real scRNA-seq experiment is often conducted on bulk tissues, where investigators randomly pick a number of cell and measure their expressions. In DE analysis, scientists are usually interested in two different scenarios: (1) within cell type: comparing the same cell types across biological conditions such as case vs. control, which reveals the expression change of a particular cell type under different contexts. (2) between cell types: comparing different cell types under the same condition, which identifies biomarkers to distinguish cell types. In either case, the experiment starts from a number of cells randomly picked from a tissue sample consisting of a mixture of different cell types. The only factor one can control is the total number of cells. In the first scenario, the numbers of cells for a particular cell type under different biological conditions are often similar, barring significant changes in cell composition. In the second scenario, the numbers of cells for distinct cell types can be very different, so the power for DE highly depends on the mixing proportions. For both scenarios, another experimental design question is how to optimize the tradeoff between the number of cells and average sequencing depth in order to maximize power, under the constraint of total sequencing depth.

For the first scenario (comparing cells of the same cell type between two biological conditions), the simulation starts from a given cell number in each group. POWSC generates expressions for one condition according to the ZIP-LNP mixture distribution in Equation 3.1, using estimated parameters from one condition in the pilot data. To be specific, for gene g, we first generate an indicator variable for its phase according to Bernoulli(π_g). If the gene is in Phase 1, its expression will be generated from $ZIP(y_{gi}|p_i, \lambda_i)$; otherwise, the expression is generated from $LNP(y_{gi}|\mu_g, \sigma_g^2)$. Next, POWSC simulates DE genes, and generates expressions for the other condition. By default, POWSC randomly select 5% genes to be Form I DE (phase transition), and 5% genes with non-zero average expressions to be Form II DE (quantitative difference in Phase 2). For DE of Form I, POWSC perturbs the gene-specific mixture proportions π_g to π_g', as follows:

$$\pi_g' = \begin{cases} \pi_g + \delta & : \pi_g < 0.5 \\ \pi_g - \delta & : \pi_g \geq 0.5 \end{cases}$$

By default, $\delta \sim$ Uniform(0.1,0.3). π_g' will be capped to be between 0 and 1. For DE of Form II, POWSC randomly generates log fold changes, denoted as κ_g, from a user-specified distribution. By default, we use the following mixture distribution $0.5 \times N(-1, 1) + 0.5 \times N(1, 1)$ for κ_g. Then the means of the LNP distribution for DE genes are generated as $\mu_g' = \kappa_g \mu_g$. After obtaining model parameters for the other condition, sequencing counts are generated according to Equation 3.1 again.

For the second scenario (comparing different cell types within the same condition), POWSC starts with a given total number of cells N in an experiment. The first step is to generate the numbers of cells for each cell type from a multinomial distribution, where the mixture probabilities are cell type proportions estimated from real data. Next, it applies *Parameter Estimator* to obtain the model parameters for each cell type. Last, it generates cell-type-specific gene expression matrix based on Equation **??**. For this scenario, DE is not generated by perturbing model parameters. This is because with multiple cell types, it is very difficult to assume reasonable effect sizes for DE genes in all pair-wise comparisons. Instead, we generate all expressions based on parameters estimated from real data, and then determine the true DE status from these parameters. This strategy allows the simulation to mimic real data situation.

In both scenarios, POWSC produces a series of expression matrices for downstream DE analysis, corresponding to different total cell numbers. The number of total cells can be specified by the user. By default, they are set to be a range of numbers 50, 100, 200, 500, and 1000. The average sequencing depth can also be specified for investigating how sequencing depth influences the power estimation. Additionally, POWSC can adjust the sequencing depth and total number of cells simultaneously to explore how to balance these two factors in a scRNA-seq experiment. For instance, if the researchers have a fixed total sequencing depth, POWSC compares the powers from sequencing more cells with lower depth to fewer cells with higher depths, and suggests an optimal strategy.

3.2.3 Power assessor

POWSC utilizes either MAST or SC2P Bioconductor packages to perform DE analysis. The results from these packages include p-values and false discovery rates (FDR) for each gene in two forms of DE tests (Form I: phase transition; and Form II: magnitude turning in Phase 2). Genes with FDR less than a user-specified threshold (with default set to 0.1) are the "called DE" genes, among which the true discoveries are referred to as the "Recovered DE" (RD) genes. POWSC compares the RD genes with the known truth to evaluate power for the two forms of DE. For each form, we focus on the “targeted power” as introduced in PROPER [161], which refers to the power in detecting DE with effect sizes considered biologically relevant. Genes with none-zero but trivial DE may not be “null” genes in the conventional definition with the null hypothesis being $\delta = 0$, but they could be biologically meaningless. Thus, we report the power for detecting DE with effect size beyond a user-specified threshold.

To define the gold standard, the default threshold for relevant DE effect sizes are set to be 10% for the difference in π_g for Form I DE, and log fold change κ_g of 0.5 for Form II DE. These thresholds can be adjusted by users. Furthermore, POWSC reveals the impact of other factors, including zero percentage (sometimes referred to as dropout rate in other publications) and mean expression level, by stratifying genes into different categories. For each gene, we compute its zero fraction $\bar{Z}_g = \sum_i (Y_{gi} = 0)/N$, and average expression $\bar{Y}_g = \sum_i Y_{gi}/N$. For Form I test, genes are stratified by the zero fractions from 0 to 1 into 5 equal-sized intervals. For Form II DE test, the genes are stratified by average expression. The average expression $\bar{Y}_g$ is divided from 0 to infinite into intervals such as (0,2], (2, 4], (4, 8], etc. These strata can be specified by the users. Within a stratum, the stratified target power is calculated, e.g., as $Power = RD/(RD + FN)$. Here, RD and FN represent the number of genes with high enough DE that are recovered or missed by the DE detection, respectively (Table 3.1).

stratified power		true DE status			
		high DE	low DE	Non- DE	Total
DE test result	Positive	RD	RD'	FD	CD
	Negative	FN	FN'	TN	CN

Table 3.1: The stratified power calculation. RD: Recovered DE; FD: false discovery; CD: called DE; FN: false negative; TN: true negative. CN: called non-DE.

3.3 Result

3.3.1 Overview

We use a real scRNA-seq dataset as blueprint to demonstrate the functionalities of POWSC. This dataset is obtained from GEO under accession number GSE29087. It is generated to profile the transcriptomes for 92 single cells consisting of mouse embryonic fibroblast (MEF) and embryonic stem (ES) cells [65]. The average sequencing depth for the dataset is around half a million.

We first evaluate the performance of the *Parameter Estimator* and *Data Simulator* on this blueprint dataset, and compare with two alternatives: scDesign [94] and Splatter [170]. We then use the *Power Estimator* to obtain the stratified targeted power vs. sample sizes relation under the two aforementioned scenarios: comparing the same cell types between conditions, and comparing different cell types within the same condition. The results demonstrate the general trends of improved power with the increase of cell numbers. In addition, we investigate the relation between sequencing depths and sample sizes under the constraint of total sequencing depth. Moreover, we provide estimated parameters for a list of datasets from common tissue types such as brain, blood, immune, and tonsil (Table 3.2). These estimated parameters are distributed with the software package, which can be easily accessed by the users to compute desired sample sizes on similar biological systems without providing pilot data.

3.3.2 POWSC accurately simulates scRNA-seq data

Parameter estimation and data simulation play essential roles in the whole simulation process because these lay the foundation for power assessment. Given a pilot dataset, simulators such as scDesign, Splatter, and POWSC will first estimate a set of model parameters based on their own statistical models, and then produce a synthetic ex-

Tissue	GEO Accession ID	# Cells	# Genes	Average Sequencing Depth (reads)	Platform
Peripheral blood mononuclear cells	GSE94820	1140	26593	∼1 M	Smart-Seq2
Brain cortex tissue	GSE67835	466	25287	∼2.8 M	Fluidigm C1
Immune	GSE65528	192	37315	∼1.1 M	Hiseq-2500
Tonsil	GSE70580	648	25219	∼2.3 M	Smart-Seq2

Table 3.2: Estimated model parameters for different biological systems. These estimated parameters are pre-stored in POWSC package

pression matrix. Splatter includes six simulation models, here we use Splat (which is their own simulation model) in our comparison. To assess how well the simulated expression matrix mimics the real data, we follow similar strategies described in Splatter and scDesign to compare the simulated to the real data. Given an expression matrix, we compute six parameters to characterize the real scRNA-seq distributions. The parameters include four gene-wise variables: mean (μ), variance (σ^2), coefficient of variation (cv), and zero fraction (ρ_1); two cell-wise variables: library size (l), and zero fraction (ρ_2). A good simulated dataset should have similar empirical distributions for the six parameters to the corresponding real scRNA-seq data. To evaluate the similarity between the simulated and real expression matrices, Splatter proposes to calculate the median absolute deviation (MAD) for the six parameters, while scDesign suggests using Kolmogorov-Smirnov (KS) distances. Smaller quantities for these two measurements indicate higher similarity between the simulated and real data.

We adopt these two metrics to compare the estimates of these six parameters from three methods. As shown in Figure 3.2A, POWSC has the best performance in the MAD measurement for half of the parameters investigated. The improvements can be substantial: the MAD for mean and gene zero ratio from POWSC is less than one third of the other two methods. For the performance by KS measurement, POWSC has the best performance in three parameters: CV, gene zero ratio, and cell zero ratio. Splat performs slightly better than POWSC in mean variance and library size

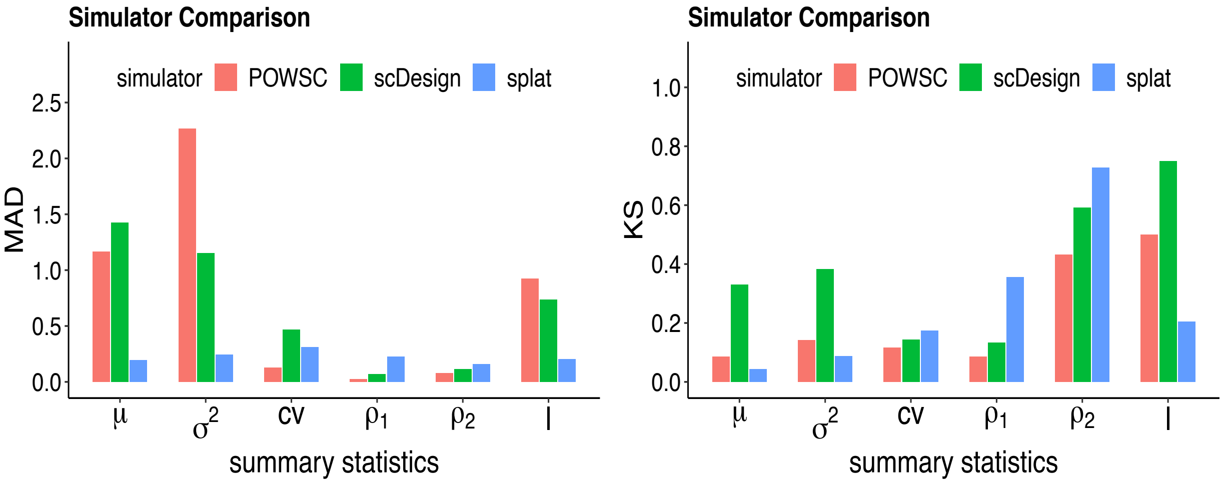


Figure 3.2: The comparison of data simulators in POWSC, scDesign, and splat. Two metrics **A**: Median absolute deviation (MAD) and **B**: Kolmogorov-Smirnov (KS) distance, are used to quantify the fidelities of a number of gene- and cell-wise parameters from simulated data. In **A**, MAD value are scaled by 10 for gene-wise mean and 1000 for gene-wise variance for better visualization. We found in the blueprint data (GSE29087), POWSC outperformed the other two simulators for MEF cell type.

estimates (Figure 3.2B). Overall, POWSC significantly outperforms scDesign and Splat in imitating the real scRNA-seq data in this comparison. We also run the comparison in 11 additional datasets in (Table 3.3). Again, POWSC shows the best performance among all methods.

Both scDesign and Splatter use stepwise procedures in simulating expressions and then introducing dropout. POWSC uses a more unified statistical model for the data: a ZIP distribution for zero-inflated background and an LNP distribution for expressions. The merit of POWSC mainly comes from the the proper data model to capture bimodal expression, and a rigorous mechanism in simulating the two expression (inactive and active) phases.

accession	**species**	**cell types**	**genes**	**cells**	**protocol**
GSE73121	Human	Primary renal cell carcinoma	25219	48	HiSeq
GSE73121	Human	Metastatic renal cell carcinoma	25219	73	HiSeq
GSE94820	Human	CD1C dendritic cells	26593	192	Smart-Seq2
GSE94820	Human	CD141 dendritic cells	26593	192	Smart-Seq2
GSE94820	Human	Monocyte	26593	372	Smart-Seq2
GSE75748	Human	Definitive endoderm cells at 96 hours	19189	188	HiSeq
GSE75748	Human	Neuronal progenitor cells	19097	173	HiSeq
GSE45719	Mouse	Middle blast	22431	60	HiSeq
GSE45719	Human	Early blast	22431	43	HiSeq
GSE70758	Human	Cell cluster with the highest proportion	25892	162	Smart-Seq2
GSE67835	Human	oligodendrocyte	25892	37	Fluidigm C1

Table 3.3: A list of scNRA-seq datasets that we curated to investigate the simulation accuracy

3.3.3 POWSC provides recommended sample size for two-group comparison

In the context of detecting DE genes that undergo two biological conditions, e.g., control versus treatment, POWSC uses the *Data Simulator* to produce a series of expression count matrices with different numbers of total cells (e.g., 50, 100, 200, 500, and 1000). Subsequently, POWSC applies *Power Assessor* on each matrix, and obtains the stratified targeted power. The reported power curves are demonstrated in Fig. 3.3. As expected, we observe that larger sample sizes lead to higher power in both forms of DE tests. For Form I DE, 500 cells in total are needed in order to achieve 80% power of detecting DE genes overall except that (0.4, 0.6] stratum is associated with the 71% power (Fig. 3.3A). For Form II DE, the sample size requirement is

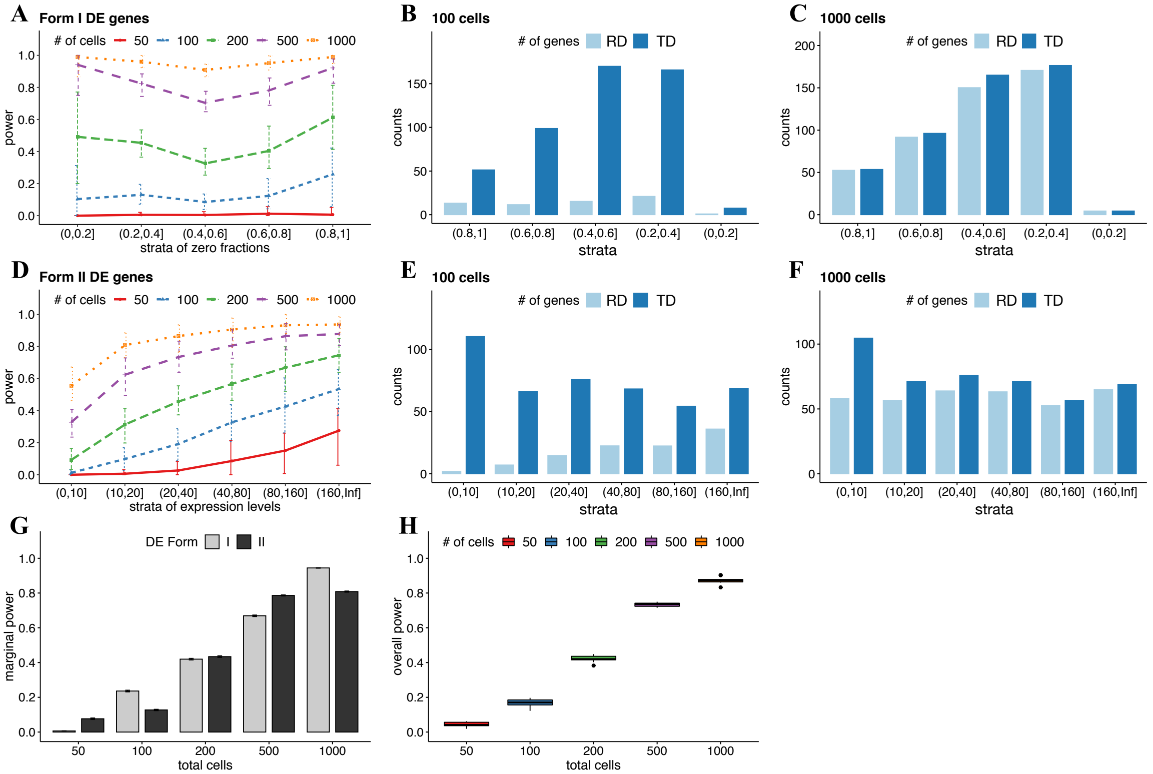


Figure 3.3: The stratified powers, marginal powers, and overall powers for two-group comparison. **A** and **D** show the stratified power curves under different number of cells, for Form I and II DE respectively. The power is stratified by cell-wise zero fractions for Form I DE, and average expression levels for Form II DE. The means and the confidence intervals are obtained by repeating 50 simulation runs. **G** and **H** show the increase of the marginal and overall power with the increase of the total cell numbers from 50 to 1000. **B-C** and **E-F** demonstrate the distributions of the recovered and true DE genes for Form I and II respectfully, in the cases of 100 and 1000 total cells.

higher. It shows that even if we discard the genes with average read counts less than 10, we still need 1000 cells to reach 80% power (Fig. 3.3D).

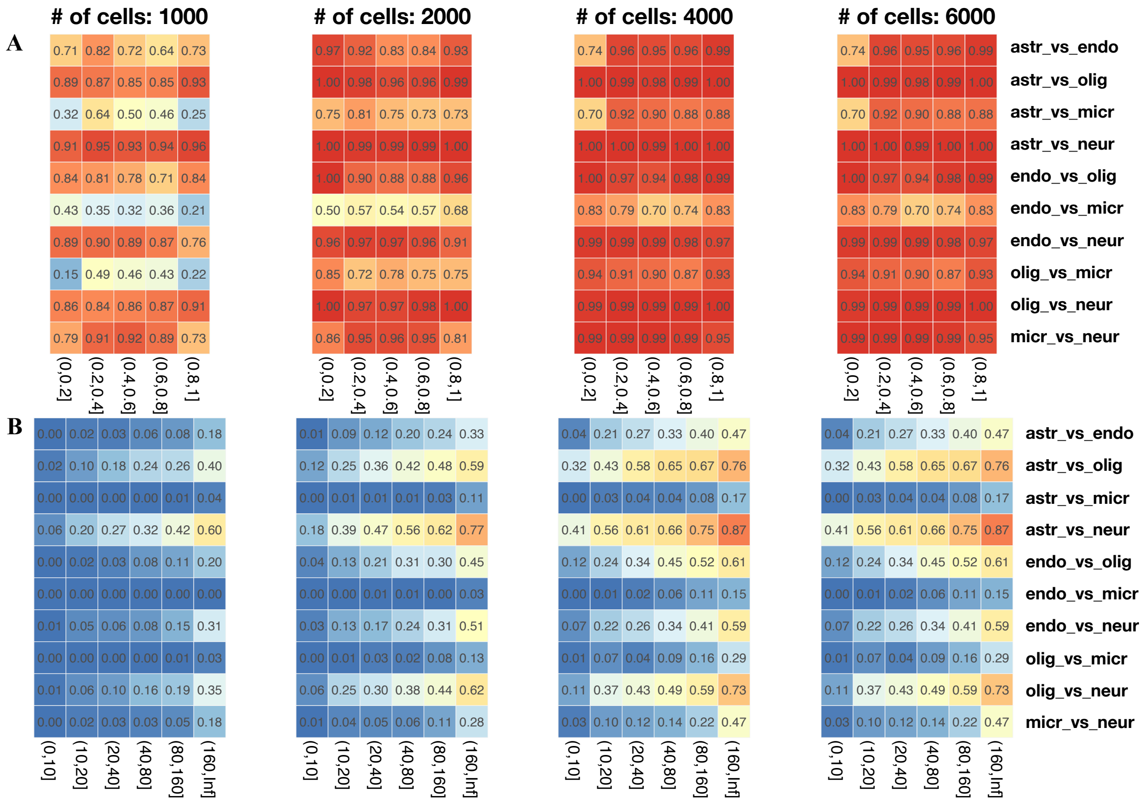
In Figure 3.3, the power curve for Form I DE has a V-shape, which indicates that DE genes with low and high zero fractions (dropout rates) are easier to be detected than DE genes with fractions close to 0.5. This is because the estimated proportions from binomial distribution have the highest variance at 0.5, thus the statistical test is not as sensitive. In Fig. 3.3D, it shows the power curves for Form II DE genes. Horizontally, the power curve inclines up as the average expression level increases, indicating that Form II DE is easier to detect for genes with higher expression levels. The same phenomenon is observed in bulk RNA-seq data [140, 3]. For genes with very low expression (average counts less than 10), the power for detecting Form II DE is low (less than 56% for 1000 cells). Thus, it might be desirable to ignore those genes before performing Form II DE test, which will reduce the testing space and improve the FDR estimate and power. Besides, we also investigate how the targeted power for Form II DE changes with respect to the zero fractions. The result (Supplementary, Figure B.1) illustrates that higher targeted power is associated with lower zero fractions, which implies that if a Form II DE gene has high zero fraction (many dropouts), it is less likely to be detected.

Furthermore, we also evaluate the marginal power, which is computed by considering genes in all strata. The marginal power is defined as $Power_{marginal} = \sum_{i=1}^{n} RD_i / \sum_{i=1}^{n} TD_i$, where TD and RD represent the number of true DE genes $(TD = RD + FN)$ with meaningful fold change and the number of Recovered DE in each stratum. We show the numbers for CD and TD for one simulation in Figure 3.3B, C for Form I test, and Figure 3.3E, F for Form II test. The marginal power for both forms is shown in Figure 3.3G. For Form I, the marginal power can reach to 94.4% with 1000 cells. For Form II, the marginal power can reach to 81.1% with 1000 cells, 90.8% if ignoring the genes with average counts less than 10. Last, we calculate the overall power in Figure 3.3H which combines the Form I and II DE genes together. As expected, higher overall power is associated with larger sample sizes. It is important to note that the sample size here is for one particular cell type. To get the total cell numbers required, one should divide this number by the estimated proportion of this cell type in the cell population.

3.3.4 POWSC provides recommended sample size for cross cell type comparison

In addition to comparing the same cell type under different conditions, another interest is to identify differences between cell types. For that, the number of cell types and their proportions in the cell population affect the power, in addition to the total number of cells. POWSC provides power analysis for this situation. Again, the *Parameter Estimator* acquires cell-type-specific model parameters, and the *Data Simulator* produces a series of expression count matrices with different numbers of total cells, each simulated dataset contains a mixture of different cell types. Next, the power assessor performs DE analysis for each pair of cell types and obtains the stratified targeted powers.

This analysis starts a template scRNA-seq dataset as blueprint for simulation. Here, we demonstrate the functionality using a human brain dataset (GSE67835) as template. From this dataset, five cell types including astrocytes, endothelial, oligodendrocytes, microglia, and neurons with proportions of 0.23, 0.08, 0.14, 0.06, and 0.49 are considered. The reported stratified targeted power is illustrated in Figure 3.4: Figure 3.4A shows the power for Form I DE, and Figure 3.4B is for Form II DE. In each plot, the columns are for different strata and the rows are for different comparisons. For example, `astr_vs_endo` means the comparison between cell types astrocyte and endothelial. As expected, more cell leads to improved power for both forms of DE. Rows in Figure 3.4B follow similar trends as the curves shown in Figure 3.3D, which implies genes with higher expression levels are more likely to be detected as Form II DE. Moreover, cell types with higher proportion are associated with higher power: the power for cell type `astr_vs_neur` comparison is higher than the other comparisons, since their effective sample sizes are greater due to their higher proportions in the cell population. These results provide detailed information for researchers to choose a proper sample size. For example, if one wants to focus on detecting the



A

# of cells: 1000	(0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1]
astr_vs_endo	0.71	0.82	0.72	0.64	0.73
astr_vs_olig	0.89	0.87	0.85	0.85	0.93
astr_vs_micr	0.32	0.64	0.50	0.46	0.25
astr_vs_neur	0.91	0.95	0.93	0.94	0.96
endo_vs_olig	0.84	0.81	0.78	0.71	0.84
endo_vs_micr	0.43	0.35	0.32	0.36	0.21
endo_vs_neur	0.89	0.90	0.89	0.87	0.76
olig_vs_micr	0.15	0.49	0.46	0.43	0.22
olig_vs_neur	0.86	0.84	0.86	0.87	0.91
micr_vs_neur	0.79	0.91	0.92	0.89	0.73

# of cells: 2000	(0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1]
astr_vs_endo	0.97	0.92	0.83	0.84	0.93
astr_vs_olig	1.00	0.98	0.96	0.96	0.99
astr_vs_micr	0.75	0.81	0.75	0.73	0.73
astr_vs_neur	1.00	0.99	0.99	0.99	1.00
endo_vs_olig	1.00	0.90	0.88	0.88	0.96
endo_vs_micr	0.50	0.57	0.54	0.57	0.68
endo_vs_neur	0.96	0.97	0.97	0.96	0.91
olig_vs_micr	0.85	0.72	0.78	0.75	0.75
olig_vs_neur	1.00	0.97	0.97	0.98	1.00
micr_vs_neur	0.86	0.95	0.96	0.95	0.81

# of cells: 4000	(0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1]
astr_vs_endo	0.74	0.96	0.95	0.96	0.99
astr_vs_olig	1.00	0.99	0.98	0.99	1.00
astr_vs_micr	0.70	0.92	0.90	0.88	0.88
astr_vs_neur	1.00	1.00	0.99	1.00	1.00
endo_vs_olig	1.00	0.97	0.94	0.98	0.99
endo_vs_micr	0.83	0.79	0.70	0.74	0.83
endo_vs_neur	0.99	0.99	0.99	0.98	0.97
olig_vs_micr	0.94	0.91	0.90	0.87	0.93
olig_vs_neur	0.99	0.99	0.99	0.99	1.00
micr_vs_neur	0.99	0.99	0.99	0.99	0.95

# of cells: 6000	(0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1]
astr_vs_endo	0.74	0.96	0.95	0.96	0.99
astr_vs_olig	1.00	0.99	0.98	0.99	1.00
astr_vs_micr	0.70	0.92	0.90	0.88	0.88
astr_vs_neur	1.00	1.00	0.99	1.00	1.00
endo_vs_olig	1.00	0.97	0.94	0.98	0.99
endo_vs_micr	0.83	0.79	0.70	0.74	0.83
endo_vs_neur	0.99	0.99	0.99	0.98	0.97
olig_vs_micr	0.94	0.91	0.90	0.87	0.93
olig_vs_neur	0.99	0.99	0.99	0.99	1.00
micr_vs_neur	0.99	0.99	0.99	0.99	0.95

B

# of cells: 1000	(0,10]	(10,20]	(20,40]	(40,80]	(80,160]	(160,Inf]
astr_vs_endo	0.00	0.02	0.03	0.06	0.08	0.18
astr_vs_olig	0.02	0.10	0.18	0.24	0.26	0.40
astr_vs_micr	0.00	0.00	0.00	0.00	0.01	0.04
astr_vs_neur	0.06	0.20	0.27	0.32	0.42	0.60
endo_vs_olig	0.00	0.02	0.03	0.08	0.11	0.20
endo_vs_micr	0.00	0.00	0.00	0.00	0.00	0.00
endo_vs_neur	0.01	0.05	0.06	0.08	0.15	0.31
olig_vs_micr	0.00	0.00	0.00	0.00	0.01	0.03
olig_vs_neur	0.01	0.06	0.10	0.16	0.19	0.35
micr_vs_neur	0.00	0.02	0.03	0.03	0.05	0.18

# of cells: 2000	(0,10]	(10,20]	(20,40]	(40,80]	(80,160]	(160,Inf]
astr_vs_endo	0.01	0.09	0.12	0.20	0.24	0.33
astr_vs_olig	0.12	0.25	0.36	0.42	0.48	0.59
astr_vs_micr	0.00	0.01	0.01	0.01	0.03	0.11
astr_vs_neur	0.18	0.39	0.47	0.56	0.62	0.77
endo_vs_olig	0.04	0.13	0.21	0.31	0.30	0.45
endo_vs_micr	0.00	0.00	0.01	0.01	0.00	0.03
endo_vs_neur	0.03	0.13	0.17	0.24	0.31	0.51
olig_vs_micr	0.00	0.01	0.03	0.02	0.08	0.13
olig_vs_neur	0.06	0.25	0.30	0.38	0.44	0.62
micr_vs_neur	0.01	0.04	0.05	0.06	0.11	0.28

# of cells: 4000	(0,10]	(10,20]	(20,40]	(40,80]	(80,160]	(160,Inf]
astr_vs_endo	0.04	0.21	0.27	0.33	0.40	0.47
astr_vs_olig	0.32	0.43	0.58	0.65	0.67	0.76
astr_vs_micr	0.00	0.03	0.04	0.04	0.08	0.17
astr_vs_neur	0.41	0.56	0.61	0.66	0.75	0.87
endo_vs_olig	0.12	0.24	0.34	0.45	0.52	0.61
endo_vs_micr	0.00	0.01	0.02	0.06	0.11	0.15
endo_vs_neur	0.07	0.22	0.26	0.34	0.41	0.59
olig_vs_micr	0.01	0.07	0.04	0.09	0.16	0.29
olig_vs_neur	0.11	0.37	0.43	0.49	0.59	0.73
micr_vs_neur	0.03	0.10	0.12	0.14	0.22	0.47

# of cells: 6000	(0,10]	(10,20]	(20,40]	(40,80]	(80,160]	(160,Inf]
astr_vs_endo	0.04	0.21	0.27	0.33	0.40	0.47
astr_vs_olig	0.32	0.43	0.58	0.65	0.67	0.76
astr_vs_micr	0.00	0.03	0.04	0.04	0.08	0.17
astr_vs_neur	0.41	0.56	0.61	0.66	0.75	0.87
endo_vs_olig	0.12	0.24	0.34	0.45	0.52	0.61
endo_vs_micr	0.00	0.01	0.02	0.06	0.11	0.15
endo_vs_neur	0.07	0.22	0.26	0.34	0.41	0.59
olig_vs_micr	0.01	0.07	0.04	0.09	0.16	0.29
olig_vs_neur	0.11	0.37	0.43	0.49	0.59	0.73
micr_vs_neur	0.03	0.10	0.12	0.14	0.22	0.47

Figure 3.4: Stratified power for cross cell type comparisons. Five cell types including astrocytes (astr), endothelial (endo), oligodendrocytes (olig), microglia (micr), and neurons (neur) are considered to imitate the situation of multiple cell types. Each panel shows the power under a certain cell number. Inside each panel, each row is for one comparison between two cell types e.g., cell type astrocytes versus endothelial denoted as astr_vs_endo; each column is for one stratum. **A** and **B** correspond to Form I and II DE.

marker genes among abundant cells, the required cell number will be smaller. If, however, one wants to identify markers for a rarer cell type, one may have to measure more cells.

We also test POWSC on a Glioblastoma (GBM) dataset (GSE57872) to demonstrate how the power changes under different biology context (Supplementary Figure B.2). This scRNA-seq dataset includes 5 individual tumor samples (MGH26, MGH28, MGH29, MGH30, and MGH31). By using MGH31 as the template, we obtain four cell types (denoted as 1 to 4) with proportions of 0.66, 0.15, 0.1, and 0.09 by SC3. We find that the cell types with higher proportion lead to higher power evaluations, consistent with the previous finding. It is also noticeable that the stratified power for both forms of DE is generally lower than that in simulation based on GSE67835. This is caused by the larger within-cell-type gene expression variability Supplementary Figure 3.5, which makes DE detection more challenging. In another GBM dataset (GSE84465) (Supplementary Figure B.3) which contains 4 patients (BT_S1, BT_S2, BT_S4, BT_S6), we perform POWSC on BT_S1 by considering 4 cell types: Astrocytes(HEPACAM), Endothelial(BSC), Microglia(CD45), and Oligodendrocytes(GC), with proportions of 0.36, 0.25, 0.23, and 0.16. The power results from this dataset is more similar to that from the human brain data in Fig. 3.4. Thus, power analysis is case sensitive and scientists should be cautious about choosing proper template data. For example, within-cell-type variability can play an essential role in the power assessment: larger variation is indicator of relatively lower power.

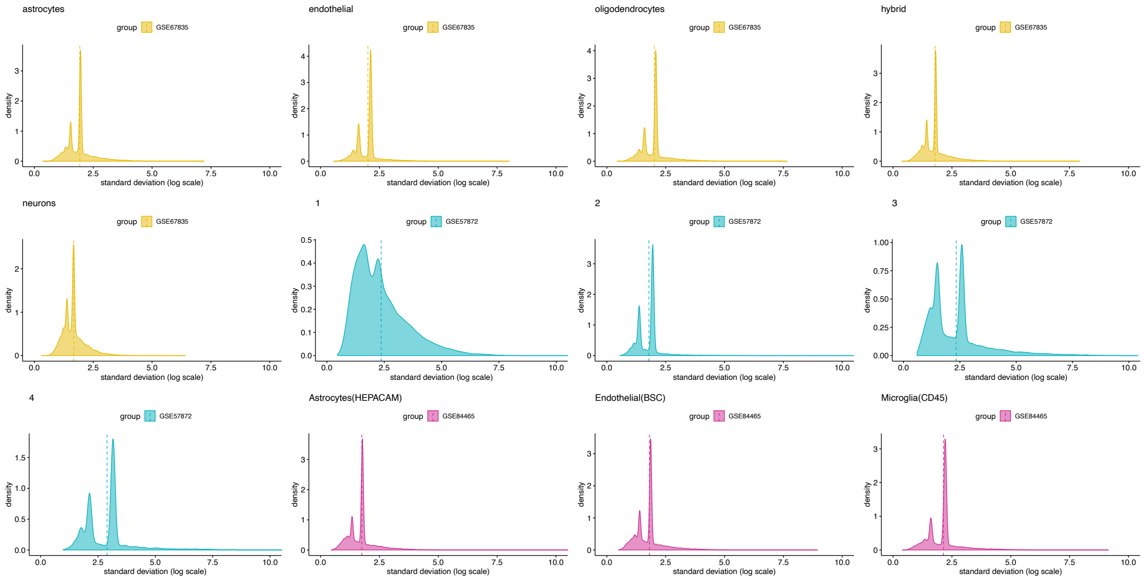


Figure 3.5: The gene-wise standard deviation for the various cell types in GSE67835 GSE57872, and GSE84465. Different colors represent different dataset. Each panel includes the density plot for gene-wise sds for one specific cell type.

Note that the cell composition information, including the cell types and the corresponding proportions, is very important for power analysis and sample size estimation. We perform additional evaluation on the impact of biases of cell proportions on sample size estimation. Overall, the power (and sample size) estimation can become unrealistic if the proportions are inaccurate, especially when the biases are large. Detailed simulation procedures and results are provided in the Section **3.3.6**. In practice, POWSC recommends two approaches for users to obtain and specify the cell proportions: (1) POWSC provides pre-estimated model parameters for a few tissue types (brain, tonsil, immune, and blood), which include cell type proportions. The researchers can directly use these if they work on the same (or similar) biological systems. (2) The users can specify the cell proportions for their experiment. This usually requires a pilot dataset, from which the cell type proportions can be estimated through existing clustering tools.

3.3.5 POWSC offers a strategy to balance sample size and sequencing depth

The results above show that DE genes with higher sequencing counts are easier to detect, which is tempting for researchers to consider deeper sequencing to assist the DE detection. However, they often also face a budget limit, which is directly related to the total sequencing reads they can afford. Thus, an interesting question is how to optimize the tradeoff between sample size (denoted by N) and sequencing depth (denoted by S) with fixed total sequencing reads, in order to maximize the power.

POWSC uses 500 cells with an average sequencing depth of 0.5 million as the baseline setting, and varies the values of N and S while keeping the total sequencing depth NS constant. The targeted power curves for Form I DE are illustrated in Figure **??**A, Form II DE in Figure **??**D, for different combinations of S and N. When utilizing the blueprint dataset of GSE29087 for simulation, we find that measuring more cells at a shallower sequencing depth leads to higher power. For example, doubling the total cells to 1000 and reducing the average sequencing depth to 0.25 million improves the power in all strata significantly. It implies that larger sample sizes are preferable to deeper sequencing when the total sequencing reads are fixed, because the positive impact of larger cell numbers on power is stronger than the negative impact of shallower sequencing depth. The overall power plot in Figure **??**H also demonstrates the preference of larger N rather than deeper S, and this finding can potentially guide the researchers to conduct the scRNA-seq experiment when facing a limited budget or fixed total sequencing reads.

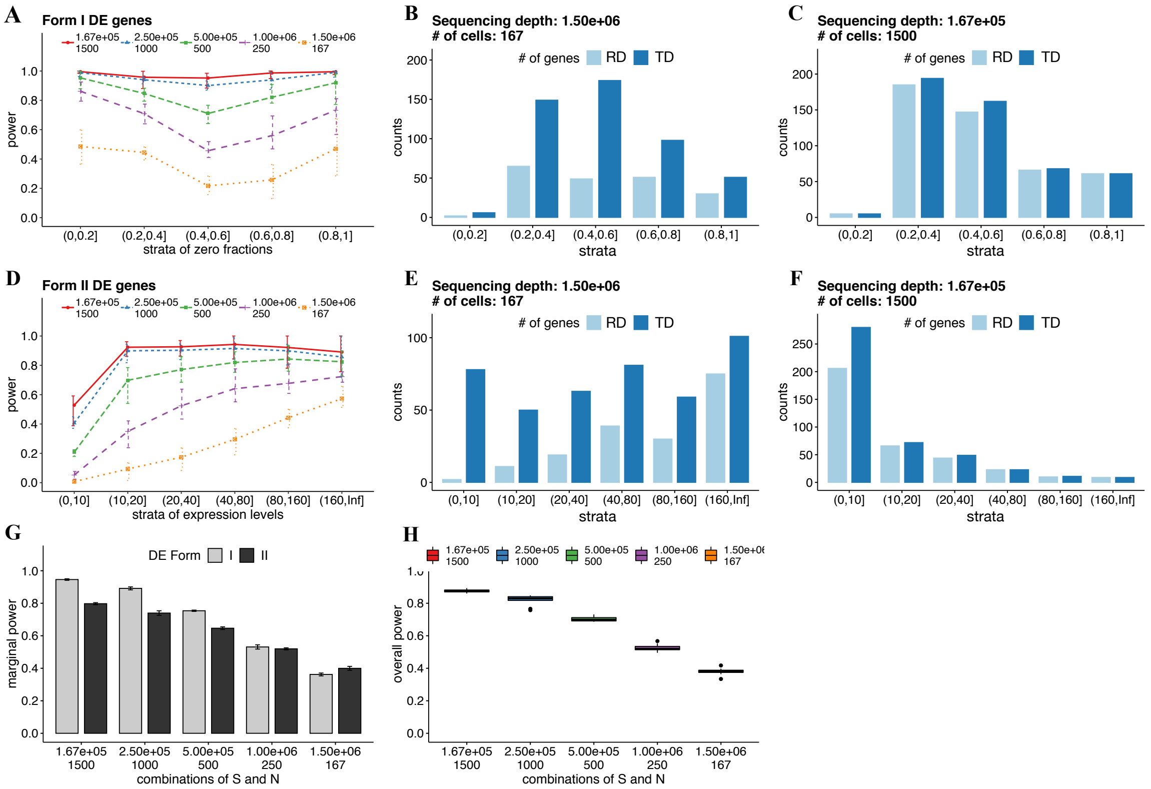


Figure 3.6: The tradeoffs between average sequencing depth (S) and sample sizes (N). We set up the case of 500 cells with the average sequencing depth 0.5 million reads as the baseline. The pre-determined total read counts is from the baseline. Then, we shrink or expand the S to (1/3, 1/2, 2, 3 times), and expand or shrink N to (3, 2, 1/2, 1/3 times) subject to a fixed cost. We found larger sample sizes are more preferable to deeper sequence depths (**A** and **D**). **G** demonstrates the marginal power (Form I and II DE) changes for each combination of S and N. **H** shows the overall power changes for each combination of S and N. Both **G** and **H** are averaged by 50 simulation runs. **B-C** and **E-F** demonstrate the distributions of the recovered DE (RD) and true DE (TD) genes for Form I and II respectfully, in the cases of 1.5e6 sequencing depth with 167 cells, and 1.67e5 with 1500 total cells.

The marginal power for both DE forms under different scenarios is illustrated in Figure 3.6G. For example, if one decreases the sequencing depth to about 167 thousand reads per cell but sequences 1500 cells in each condition, the marginal power for the two forms of DE tests becomes close to 0.95 and 0.8, respectively. On the other hand, if one spends 1.5 million reads for each cell but only sequences 167 cells in each condition, those two marginal power suffers a huge reduction to only around 0.35 and 0.39. Notice that with lower sequencing depth for each cell, more DE genes will locate in the (0,10] interval, so successfully detecting these DE genes at low magnitudes will contribute the marginal power evaluation specifically for Form II DE. Each panel in Figure 3.6B,C, and Figure 3.6E,F representing a combination of S and N, plots the distributions of recovered and true DE genes for both forms.

Even though more cells in general leads to higher power, it is not true that infinitely increasing the cell numbers can improve power, because the sequencing depth will eventually become too shallow have adequate coverage. We investigate the extreme cases where the N is extended by factors such as 20 and 30 and S is shrunk accordingly. The targeted power for Form I and II DE becomes unstable especially in the stratum (0.8, 1) for Form I DE in (Figure 3.7A), and (0, 10), (160, ∞) for Form II DE in (Figure 3.7B). A similar trend is observed in data from 10X Genomics platform, where the sequencing depth is extremely shallow (~3-10k per cell).

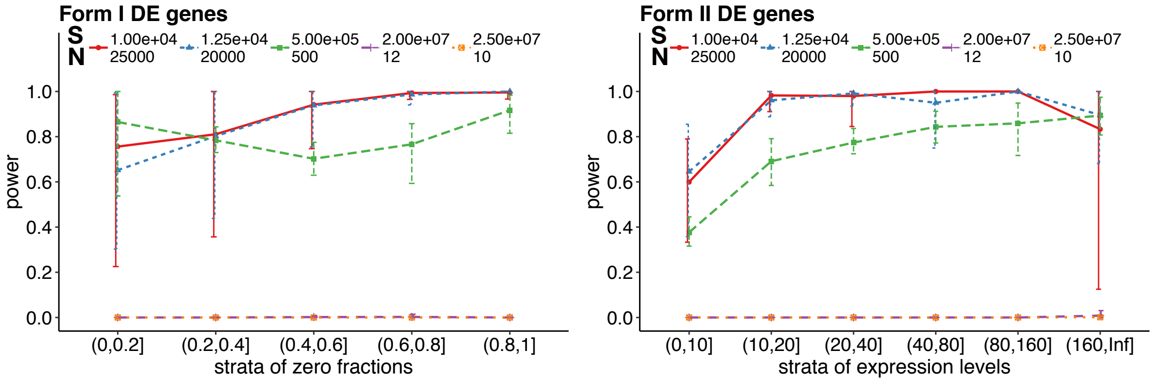


Figure 3.7: Simulation to extreme cases. We multiple the sample size to larger factors such as 30 and 20 to mimic the extreme cases. We found that it is not entirely true that infinitely increasing the sample sizes will lead to higher targeted powers..

Thus the power analysis not only provides information for experimental design, but also guidance for the data analysis plan.In our simulation studies, we hold S reasonably high, which is important if one wishes to perform DE analysis. We find that as long as S is at reasonable level, our data model and the main conclusion of more cells leading to higher power generally hold well. It is possible that the data characteristics can diverge from our model when S becomes very small. For that, we will develop new statistical model to characterize the data, which is our future research plan.

3.3.6 POWSC handles the perturbation of cell compositions

Cell composition information, including the cell types and the corresponding cell proportions, are important for power analysis and sample size estimation in scRNA-seq. In practice, the cell proportions sometimes can be inaccurately estimated. Here, we discuss the robustness of POWSC when there are biases in the cell proportions. We assess the power evaluation results after purposely perturbing some cell type proportions, where the cell types are predetermined. The specific simulation settings are as followings: (1) Simulations are based on the brain cortex data (GSE67835) which includes five 5 cell types including astrocytes, endothelial, oligodendrocytes, microglia, and neurons with proportions of 0.23, 0.08, 0.14, 0.06, and 0.49; (2) We increase the cell percentage for astrocytes by 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6, and decrease proportions of other cell types; (3) Total number of cells is fixed at 2000. The stratified power for Form I DE and Form II DE are illustrated in Figure 3.8A and Figure 3.8B. Each panel represents one simulated case where the cell proportions are manually perturbed. For a pairwise comparison for Form I and II DE, we report the sum of absolute change of all stratified powers against the sum of absolute change of the cell proportions in Figure 3.8C and Figure 3.8D. With increasing biases of cell fractions, the overall power changes become more drastic. For example, the sum of absolute changes of all stratified powers in the endo_vs_micr comparison for Form I DE is increased from 0.24 to 1.57, and endo_vs_olig comparison for Form II DE changes from 0.29 to 0.85, Overall, these results show that the biases in cell proportions are directly related to the accuracy in power estimation. The power (and sample size) estimation can become unrealistic if the proportions are inaccurate.

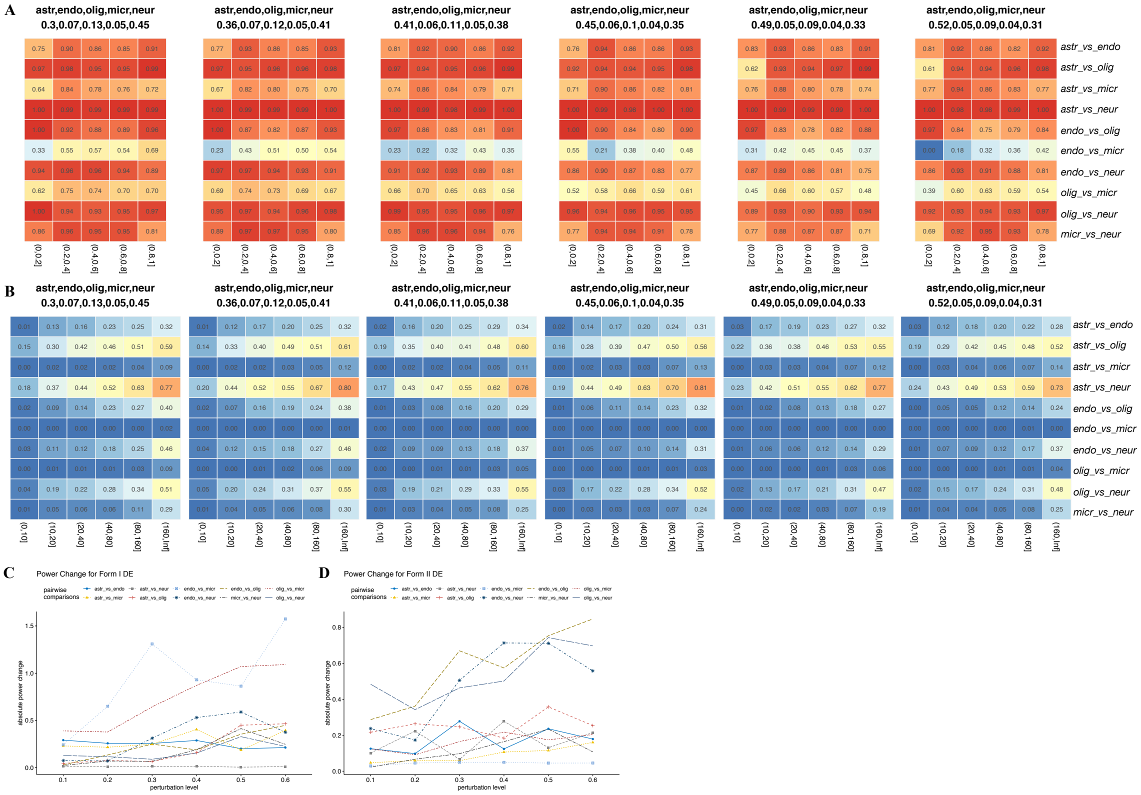


Figure 3.8: The stratified powers change with the change of perturbation levels from 0.1 to 0.6. A and B illustrate the power change for Form I and II DE. For a pairwise comparison in Form I (C) and II DE (D), we report the sum of the absolute change of all stratified powers with the sum of absolute change of the cell proportions.

3.3.7 Extend POWSC to the context of clustering

So far, we have demonstrated the success of POWSC in accurately predicting the sample size and providing power evaluation under the context of the DE analysis. Further, we develop another pipeline named POWCLUST as an extension of POWSC which assesses power with respect to properly recovering the cell type information under the context of clustering.

POWCLUST also includes three steps of parameter estimator, data simulator, and power assessor. Considering the low sequencing depth observed from the droplet-based scRNA-seq experiments, we adopt the Dirichlet-multinomial Bayesian framework to directly model the count expression at the individual cell level. The model we propose is as the following: $\vec{y_i}$ is the count vector for all the genes from the i^{th} cell. Each element in the probabilistic vector $\vec{p_j}$ corresponds to a gene expressing a certain level of reads. r_i presents the total read counts from the i^{th} cell. $\vec{\alpha_i^k}$ denotes the parameter of the k^{th} cell type specifying the conjugate prior i.e., Dirichlet distribution. To fast and accurately estimate the model parameters, we use the Method of Moment approach described by Dr. Narayanan [107].

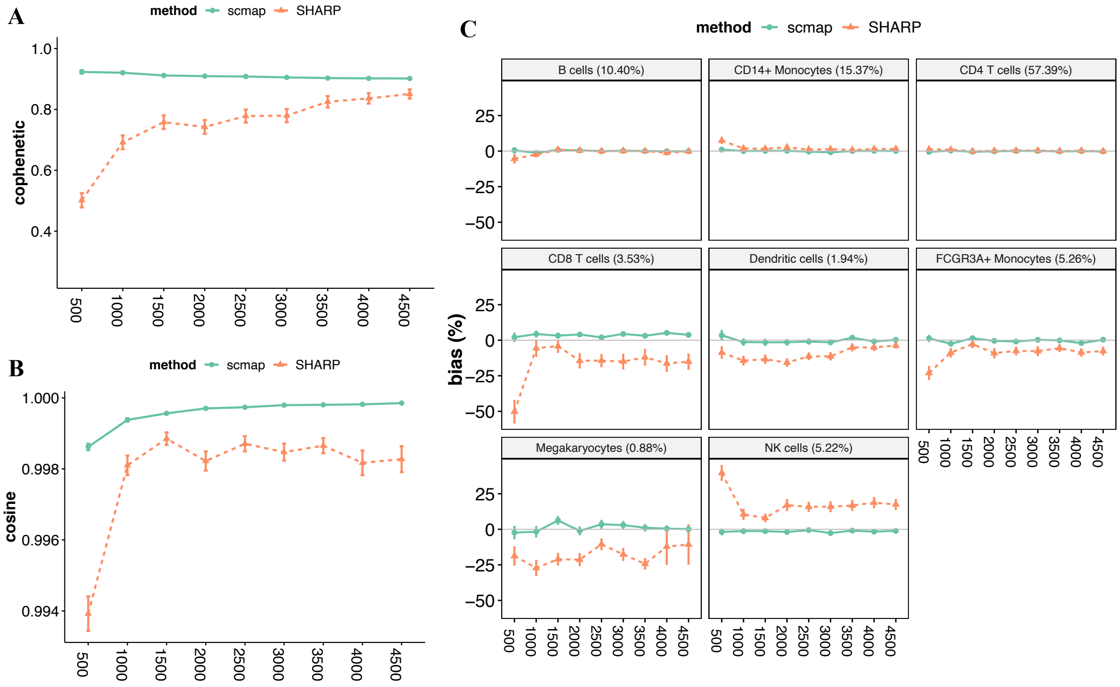


Figure 3.9: Apply POWCLUST on PBMC datase (GSE96583)

To evaluate whether the simulated the data can properly recover the cell type information, we use the following strategies: 1. Estimate the similarity of the cell type hierarchy between the reference and the obtained tree structure from the generated data. The hierarchy of the cell types are obtained by fitting the traditional hierarchical clustering on the cell type centroids. We use cophenetic correlation [129] to measure the similarities between two structured trees. 2. Evaluate the cell type proportion between the reference and the computed cell fraction from the simulated data. We use the cosine similarity to measure the closeness between two cell type proportion vectors. 3. Calculate the specific cell type bias.

To obtain the classified cell type information, we investigated both supervised approach and unsupervised approaches. For supervised approach, we used the scmap [78] to assign the cell type in the simulated data by comparing to the reference data. For unsupervised approach, we used the SHARP [152] for clustering the cells and iteratively map the cell types with the reference cell types by constructing the similarity matrix.

We used one dataset under the GEO accession ID GSE96583 about peripheral blood mononuclear cells (PBMCs) as the template to estimate model parameters. The data includes 8 cell types of B cells, CD4 T cells, CD8 T cells, and etc. For the computation process, we first simulate a series of datasets with different sizes ranging from 500 to 4500. Then, we applied both scmap and SHARP for predicting the cell type labels for each dataset. Next, we compare with reference by reporting the cophenetic correlation, cosine similarity, and bias. As demonstrated in the Figure 3.10, it shows that larger size will associate with higher cophenetic statistics in (Figure 3.9A) and cosine similarity in (Figure 3.9B), which indicates including more cells will be more likely to recover the true information of cell type hierarchical structure and cell type proportion. It also tells that using supervised approach for assigning cell types is more sensitive than unsupervised approach. Furthermore, we showcase that the cell type bias (Figure 3.9C) for the abundant cell types will likely converge with large sample sizes such as CD4 T cells (57.39%), and CD14+ Monocytes (15.37%), but the trend is not obvious for cell types with relatively smaller proportions such as CD8 T cells (3.83%), and Dendritic cells (1.94%).

3.4 Discussion

For the scRNA-seq experiment designed for identifying differential expression, statistical power evaluation plays an essential role for sample size recommendation. Multiple factors affect the ability of detecting DE genes in scRNA-seq. These factors include but go beyond the typical size effect and sample size in traditional single hypothesis testing, and beyond those encountered in bulk RNA-seq including sequencing depth and mean expression level. The unique nature of scRNA-seq datasets includings the high percentage and large variation of zero counts, the multiple cell types measured simultaneously, and the different forms of DE. These unique characteristics, adding the high dimensionality with unexchangeable tests, make theoretical samples size "calculation" or "determination" impractical, and the power assessment even more complicated than in bulk RNA-seq. Rather, we follow the strategy in bulk RNA-seq and provide sample size recommendation via simulation so that the scientists can have a comprehensive view of what can be expected in various sample sizes.

Another consideration is the diverse sequencing methods with different protocols. There are currently a number of scRNA-seq technologies, including CEL-seq2, Drop-seq, MARS-seq, SCRB- seq, Smart-seq2 (the improved version of Smart-seq), and Seq-Well. Many considerations such as the cost and capture accuracy will influence the choice of sequencing platforms. It is reported that Drop-seq is more effective in transcriptome quantification for large numbers of cells, while MARS-seq, SCRB-seq, and Smart-seq2 provide improved transcripts detection and coverage [180]. Recently, 10X Genomics system has become a popular choice for researchers. The technology is designed to sequence a large number of cells at shallow coverage, which is ideal for cell clustering and rare cell type discovery. To test whether POWSC can still perform on 10X datasets, we applied POWSC based on a Peripheral blood mononuclear cells (PBMCs). This dataset (https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k) includes 4,340 cells with average sequencing depth at around 3000. We find that enlarging sample size can lead to higher power even with such shallow sequencing depth, as expected. However, the power becomes saturated after including a certain number ($\sim$1000) of total cells (Supplementary, Figure B.4-B.5). For Form II DE, the majority of the DE genes have less than 10 counts on average (Supplementary, Figure B.5A-E). This leads to low marginal power: with 2000 cells, the marginal power is barely over 0.4 (Supplementary, Figure B.6). Overall, 10X is not recommended for DE analysis.

There are some discussions and debates on the cause of sparsity (the abundance of zeros) in scRNA-seq data, e.g., whether the sparsity is due to biological fact or technical limit or a combination of the two. The sparsity mechanism will have some impact on the data modeling strategy; for example, whether a zero-inflation component is necessary. It has been reported that the raw read counts from Smart-seq have zero-inflation, while Unique Molecular Identifiers (UMIs) counts can be adequately modeled by a negative binomial distribution (NB) [46, 142, 138]. We model the scRNA-seq counts as a mixture of a zero-inflated Poisson (ZIP) for background and a lognormal-Poisson (LNP) for foreground. This provides the flexibility to accommodate either case (dropout or dropdown) when the user choose different parameters. For example, the zero-inflation can be removed (then the background simply becomes a Poisson) if the mixing proportions for the point mass at zero are specified as 0, or the whole ZIP component for background can be removed if proportions for ZIP are set to be 0. The LNP model captures the same mean-variance relationship as the negative binomial model does but is more flexible when the dispersion is greater (as typical for scRNA-seq data) [163]. Most importantly, our modular software implementation keeps the tool flexible such that users may elect to simulate scRNA-seq data from other data generative models (either published data generator or their own in-house simulation) and still benefit from the POWSC's power evaluation functionalities.

It's important to point out that the current scheme in POWSC starts from simulating read count matrix. For more realistic simulation, it is ideal to consider additional factors related to data processing, such as read mapping efficiency and read count summarization. This requires simulating scRNA-seq data at the sequence read level, and there are existing works for it [122]. To incorporate that component in POWSC is our research plan in the near future.

In cross cell type comparisons, our results in Section 3.3 and Figure 4 are based on the true cell type labels. In real data analysis, cell types may be identified by surface markers (such as flow cytometry cell sorting) in which case the cell types are usually considered as identified without mislabeling. Another type of analysis is also common: cell clusters are identified using unsupervised clustering algorithms [77, 123], followed by an inspection of the highly expressed genes in each cluster to infer their cell types. In this kind of analysis, since the cell types are inferred and are subject to error, the DE analysis comparing the clusters may find spurious DE that is only associated with these given clusters [172]. The typical multiple testing adjustment procedures do not guard against this type of false discovery, because some of these false positives represent true difference between the clusters. The impact of the potential mistakes in the clustering step on the DE analysis depends on how distinct the cell types are, hence how accurately the cells are clustered and cell types inferred. This is beyond the scope of this manuscript, but certainly an important direction for our continuous development of POWSC. Moreover, with a modular structure, POWSC makes it easy for users to adopt other DE analysis to the ones we currently include, such as the one proposed in [172] that provides a solution for clustering-then-DE type of analysis.

Right now, POWSC is designed for assessing powers under the context of two-group comparison DE analysis. In the near further, we will extend the power evaluations for DE in more complex experimental design (multiple-group comparison, with continuous covariates, etc.). Another important point is related to scRNA-seq data from multiple subjects. To date, the analysis method for scRNA-seq data from multiple subjects is not yet well-developed. A common practice for that is to combine data from all subjects, which implicitly ignore the biological variation among subjects. We acknowledge that this could potentially underestimate the number of cells required. Continuous development of POWSC with consideration of inter-subject variation is our research plan in the near future. Other important utilities of scRNA-seq such as cell clustering and rare cell type discovery also require sample size recommendation. In these questions, there is no clear definition of “power” in the traditional statistical sense. Scientists need to use other metrics and relate them to sample size. These are certainly interesting questions worth careful exploring and may demand new metrics for assessment.

Chapter 4

Besides scRNA-seq: the design of iPath

```
Algorithm 3: iPath
```

4.1 Introduction

Cancer is a leading cause of morbidity and mortality worldwide and its prevalence is rapidly increasing, primarily due to the aging of the population. Given this, there is an urgent need for understanding the molecular mechanisms of tumorigenesis in order to develop effective treatments. It has long been recognized that dramatic transcriptome alteration is a hallmark of cancer [53]. Detecting gene signatures in transcriptome profiling data has been an essential step for many cancer studies [18, 29, 166, 182]. Using microarray or RNA sequencing (RNA-seq), many important discoveries have been made using differential expression (DE) detection techniques [115, 130, 177]. For example, important biomarker genes in breast cancer have been identified using high-throughput technologies [146, 70].

Despite the successes and importance of DE gene detection, significant challenges limit its utility. First, the expression level of many genes is rather dynamic and is affected by many factors that may or may not relate to the disease. Second, most high-throughput technologies produce data with substantial uncertainties: a long list of DE genes is usually produced, with many of them potentially being false positives. The low reproducibility of high-throughput technologies has long been acknowledged [92]. To overcome this challenge, scientists have developed gene set enrichment analysis (GSEA) [136]. Instead of individual genes, GSEA focus on pre-defined gene sets and use rankings instead of actual expression levels, to determine whether a given gene set shows concordant and statistically significant changes between two conditions. GSEA is specifically designed to analyze inherently noisy data produced from high-throughput assays, such as microarray and RNA-seq. Operationally, GSEA first ranks all genes in the genome based on the level of expression changes between two conditions (e.g., treatment and control). Then, it focuses on whether the genes from predefined functional gene sets locate towards the top or bottom of the sorted list by calculating a Kolmogorov-Smirnov version of enrichment score (ES). GSEA has been shown to be a powerful method, especially for cancer research. Besides GSEA, a dozen of methods designed for pathway analysis around the same time, and Atul et al. systematically reviewed these pathway analytic approaches during the past ten years in 2012 [74]. Recent studies demonstrate that alterations in multiple genes tend to accumulate in pathways central to the control of cell growth and cell fate determination [5, 124, 175].

However, cancer is characterized by tremendous phenotype heterogeneity, which is also reflected at the molecular level. The new precision-medicine philosophy advocates for a treatment plan that targets the unique characteristics of the tumor. Therefore, it is critically important that one focuses on the unique pattern shown in the individual tumor sample in order to identify the most promising treatment strategy for the patient. Despite its success, GSEA is predominantly carried out as a follow-up to DE analysis. GSEA looks for those gene sets that have gone through significant systematic changes between two groups of samples. Therefore, significant pathway changes that occur only in a small number of samples will likely be missed by GSEA.

Cancer is a disease of the genome. Multiple types of genomic or epigenomic alterations have been linked to human malignancies, including mutations, translocations, and changes in DNA copy number, gene expression and CpG methylation patterns. Given the vast heterogeneity among disease prognoses, it is of great interest to identify biomarkers that can predict clinical progression and outcomes. In a recent study [145], Uhlen et al. comprehensively and systematically correlated gene expression differences with patient survival. Using data from The Cancer Genome Atlas (TCGA), they identified multiple candidate prognostic genes whose expression level strongly correlated with the patients' overall survival.

Despite identifying many prognostic genes, the substantial variation and uncertainties that are ubiquitous in high-throughput technologies may raise concerns of robustness when using a single gene as the biomarker. Additionally, cancer is a complex disease: tens, or even hundreds, of genes are interactively involved and together play an important role in tumorigenesis and progression. Therefore, we hypothesize that gene sets—especially pathways and pre-defined, biologically meaningful gene sets—could serve as better biomarkers than individual genes to predict clinical outcomes for cancer patients in terms of robustness and interpretability. We acknowledge that a pathway is much more than just a gene set since how genes interact with each other is exceedingly important. However, in this work, we only focus on the gene membership part of the pathway, for simplicity consideration, we use the two words interchangeably. Given that whole transcriptome profiling has become increasingly affordable in the clinic, in this study, we explored the feasibility and efficacy of using the expression profiles of pathways or pre-defined gene sets as biomarkers, and compared them with individual gene biomarkers.

Here, we introduce iPath, or individual-level pathway analysis, to quantify the magnitude of alteration occurring for a particular pathway at the individual sample level. Our goal is to understand cancer one tumor sample at a time. Since tens or hundreds of genes are required to work together harmoniously in order to achieve even a simple biological function, and because high-throughput assays are known to produce data with a substantial amount of noise and artifacts, we believe it is more effective and robust to study genes in a pathway or gene set collectively, as a group, rather than one by one. To achieve this, for each pathway we calculate a pathway-based individual-level Enrichment Score (iES) (see Methods) to classify tumor samples into two groups—normal-like or perturbed—and then conduct a formal statistical test (reporting a log-rank p-value) to check whether such grouping has any implication on clinical outcomes such as overall survival.

The idea of conducting personalized GSEA has appeared in the literature. For example, Barbie et al. introduced single sample GSEA (ssGSEA) [10], which internally integrates the calculation of GSEA with a modified weighting factor. Gundem and Lopez-Bigas introduced Sample-level enrichment analysis (SLEA) [49]. Both methods produce a score for every pathway and sample. However, in ssGSEA, genes are ranked by their absolute expression values and the ESs are based on their ranks. In SLEA, genes are randomly permuted, and a pathway is scored by comparing the expression levels of its member genes before and after permutation. In both methods, pathways consisting of genes with constitutively high levels of expression (for example, housekeeping genes) will score higher. Other tools use a relative complicated approach for calculating the individual-level pathway ES. For example, GSVA [64] obtains the gene ranks by fitting the gene-specific kernel functions, and computes a Kolmogorov-Smirnov statistics like ES. Pathifier [35] computes the distance between each individual and a fitted principal curve in the low-dimension space for each pathway. iPS [38] computes ESs for tumors by summing up the correlation perturbations with the reference of normal samples. In contrast, iPath ranks genes based on the magnitude of their departure from the overall expression levels across the tumor and normal samples which improves quantification of the changes induced by experimental condition or disease status. As a result, iPath is better at identifying disrupted pathways as prognostic biomarkers which we are able to demonstrate in the present study.

We applied iPath to perform a pan-cancer analysis using well-established pathways and gene sets cataloged in the Molecular Signature Database (MSigDB) [95]. Our results suggest that pathways are better options than single genes in terms of predicting clinical outcomes. Thus, we believe that prognostic pathways are promising and reliable biomarkers for precision oncology. Additional analyses further reveal that many of these prognostic biomarker pathways can be linked to frequently mutated cancer driver genes in a cancer-specific manner, illustrating the intricate interactions between somatic mutations, abnormal gene expression, and tumorigenesis.

4.2 Method

4.2.1 Data sources

Transcriptomics data

The iPath program takes expression data and clinical data as input. We utilize the RTCGA—an R package, to retrieve level 3 RNA-Seq data and clinical data from TCGA maintained by Broad Institute GDAC Firehose.

Pathways and gene sets

Pathway information is obtained from the Molecular Signature Database (MSigDB). MSigDB stores eight different collections of biologically relevant pathways, enabling the discovery of biomarker pathways from different biological perspectives. From

MSigDB, we downloaded the C2 collection (4,726) curated from pathways sources, biomedical literature, and expert knowledge, as well as the GO collection (5,917) annotated by GO terms.

4.2.2 Overview of the iPath approach

The goal of iPath is to identify pathways that show unusual patterns at single sample-level. To achieve this, we defined a novel statistics, namely iES. For each pathway and each given patient, iPath first computes iES, a single value that reflect the overall expression profile of the pathway in this sample relative to the population average of all samples. Such a method allows us to quantify the level of irregularity for a set of genes in a single sample. Next, for each pathway, using normal samples' distribution of the iESs (Figure 4.1c), we come up with an iES threshold which we use to classify all tumor samples into either the category of normal-like or perturbed. Last, we compare the survival difference between these two groups, and designate a pathway as a prognostic biomarker pathway if the two groups of patients show significant difference in overall survival (Figure 4.1b).

4.2.3 Calculation of iESs

For each cancer type, we denote the RNA-seq expression matrix as $Y = y_{ij}$, with rows corresponding to the patients and columns corresponding to the genes, and i=1,. . . ,M, and j=1,. . . ,N, M is the total number of samples and N is the total number of genes in the genome. The expression levels Y are assumed to have already been normalized, for example, measured by FPKM or RPKM values. We first use all the samples of this cancer type to construct a transcriptomic homeostasis, calculate the mean ($\bar{y_j}$) and standard deviation (s_j) of the expression level for every gene in the genome. Then for each sample, assuming sample i, calculate iES for every pathway as follows.

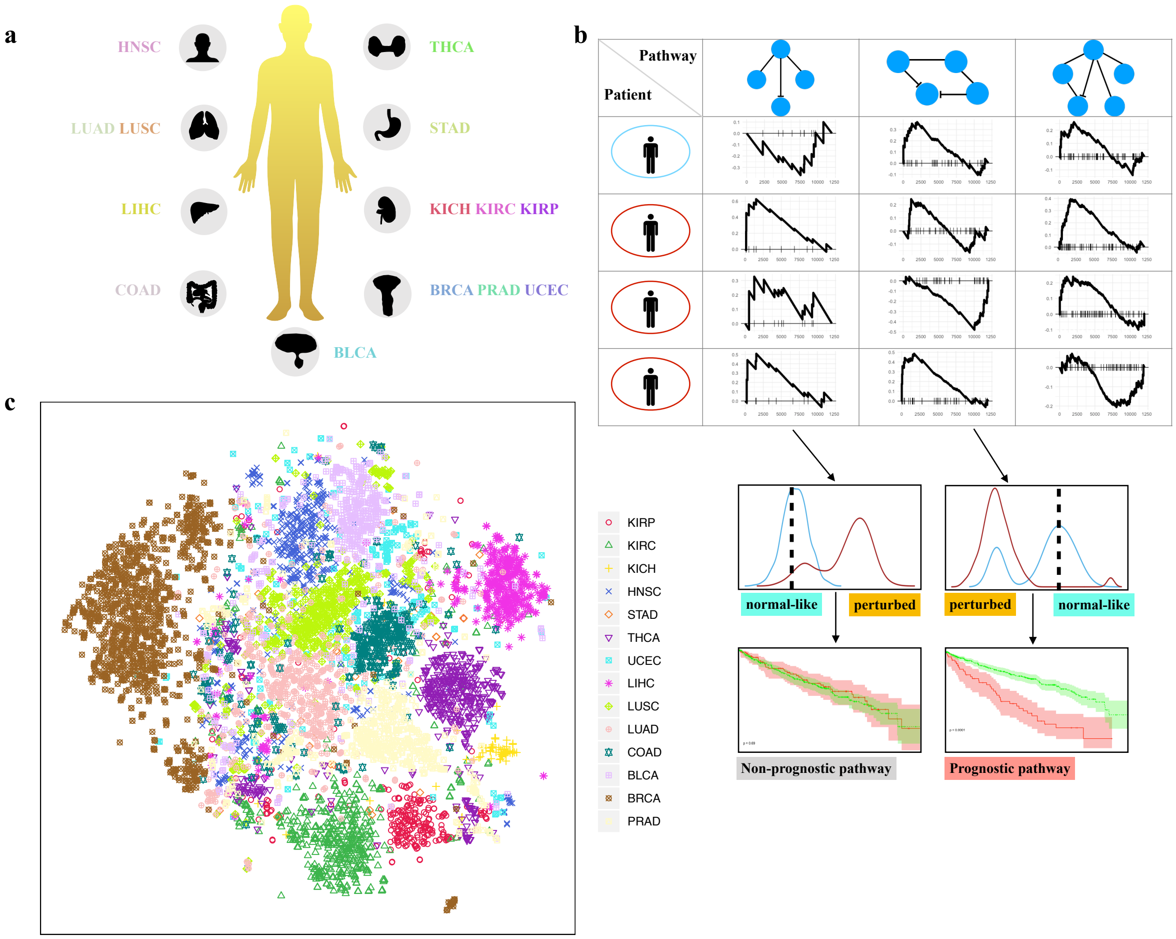


Figure 4.1: **a**.The 14 cancer types analyzed in this pan cancer study. **b**. The workflow of iPath, as demonstrated in the table with rows representing patients and columns representing pathways, iPath first calculates an iES score for each pathway and each patient (norms in the blue circle and tumors in the red circle). Then, for each pathway, iPath divides tumor samples as either normal-like or perturbed groups based on the iES scores from the normal patients. Last, iPath performs survival analysis for two tumor groups and determines prognostic pathways based on significant survival difference. **c**. The t-SNE data visualization of the iES scores from all samples of the 14 cancer types.

1. Calculate z-score $z_{ij} = (Y_{ij} - \bar{y_j})/s_j$ for every gene, here z_{ij} represent the level of deviation from the norm for gene j in the ith sample, i=1,. . . ,M, and j=1,. . . ,N.

2. Next, sort the absolute value of z_{ij}, denoted as $|z_{ij}|$, in descending order to obtain the ranks of all genes in the genome, denoted as $g_i1, g_i2, \ldots, g_{iN}$ such that

$|z_{ig_{i1}}| \geq |z_{ig_{i2}}| \geq \cdots \geq |z_{ig_{iN}}|$.

3. Subject the sorted gene list $g_i1, g_i2, \ldots, g_{iN}$ to the GSEA analysis: given one pathway (S) including R genes, iPath loops through the sorted gene list $g_i1, g_i2, \ldots, g_{iN}$ and calculates a running sum (Kolmogorov–Smirnov) statistics iES_i for i^{th} sample in the following manner: if the g_j is not in S, then subtract a penalty score $\frac{1}{N-R}$; If the g_j is in S, then add a n incremental score $\frac{|Z_{ij}|}{\sum_{j\in S}|Z_{ij}|}$. By aggregating the scores from each position, it computes the iES_{ip} value at the p^{th} position in L^i as:

$$P_{increments}(S,p) = \sum_{\substack{g_i \in S \\ j \leq p}} \frac{|Z_{ij}|}{S_R},\ where\ S_R = \sum_{g_i \in S} |Z_{ij}|$$

$$P_{penalities}(S,p) = \sum_{\substack{g_i \notin S \\ j \leq p}} \frac{1}{N-R}$$

The iES score fori^{th} sample acquires the maximum deviation from zero of $P_{increments} - P_{penalities}$. It is worth noting that utilizing $|Z_{ij}|$ for the i^{th} sample allows for the estimation of the leading contribution of the most perturbed genes.

4.2.4 Definition of perturbed tumor samples

For each pathway, we classified each tumor sample as either normal-like or perturbed. Perturbed means a significant departure from the expression homeostasis observed for this group of genes in normal samples. To achieve this classification, we used the distribution of the normal samples' iESs as the benchmark (obtained their mean and standard deviation). Specifically, we labeled a tumor sample as “perturbed” if its iES was more than two standard deviations away from the normal samples' mean, in the direction along the normal samples' mean towards the tumor samples' mean. Otherwise, the sample is labeled “normal-like”. In cancer studies, especially for solid tumors, “normal” samples typically refer to tissues adjacent to the tumor site, hence the level of heterogeneity in the normal samples is usually quite high. This is evidenced by frequently observing more than one mode in the distribution of the iES values among the normal samples. In order to best estimate the mean and standard deviation of the bona fide normal samples, we fit a Gaussian mixture model for these iES values to account for heterogeneity, and selected the mean and the standard deviation for the subgroup of samples with the highest posterior probability. This can be achieved by specifying the modelName parameter to "V" inside the Mclust function (mclust R package), which is able to automatically determine the number of the modes and assign samples to clusters.

Using pathway "FARMER BREAST CANCER APOCRINE VS LUMINAL" in BRCA as an example. In Figure 4.3, from the density plots, we observed that the overall iESs for tumor samples were higher than the normal samples (first column: waterfall plot, and second column: density plot), so we used the mean + 2sd as the cutoff to determine whether a tumor sample was perturbed. Figure 4.3a shows enrichment plots of three normal-like samples in the first column. Figure 4.3b shows that of three perturbed samples. Figure 4.3c shows a random normal sample. After classifying all tumor samples into either normal-like or perturbed, survival analysis indicated that this was a prognostic biomarker pathway (see the Kaplan-Meier plot in the fourth column of Figure 4.3e). The same trend is found in another biomarker pathway "PEDERSEN METASTASIS BY ERBB2 ISOFORM 3".

4.2.5 Performance comparison among sample-level gene set analysis methods

Clustering. We adopt the following steps: (1) randomly choose 50 normal and 50 tumor samples from the BRCA cohort; (2) for each method, we calculate an ES matrix with rows corresponding to pathway/gene sets and columns corresponding to samples. (3) conduct DE analysis on the ES using limma (67). (4) select the top 10 gene sets according to the adjusted p-values and perform the hierarchical clustering. (5)

bipartition the hierarchical tree into two classes and compare the clustering results with sample labels using the adjusted rand index (ARI). (6) repeat the above process 1000 times and summarize the average ARI for each method.

Survival analysis. We randomly sample 70% patients as the training set and use the rest of the data as the test set. Using training data, we fit individual Cox proportional hazards model for each BIOCARTA pathway and select the pathway that best correlates with the survival. Then using the test data, we assess the predictive ability of the selected pathway by computing the concordance index (c-index). We repeat the random samplings for training and test data 1000 times. The distributions of c-indices are summarized using boxplots.

4.3 Result

4.3.1 Overview

We systematically explored the relationships between biological pathways or gene sets (referred simply as “pathways” hereafter for the sake of simplicity) and clinical outcomes in 14 solid cancer types (Figure 4.1a), using data available from TCGA (Supplementary Table C.1). These cancer types were selected because we require at least 20 matching normal samples in each cancer type. These normal samples are either normal or adjacent-normal tissues in the tumor patients.

We studied two major collections of pathways: C2 curated gene sets from MSigDB and Gene Ontology (GO) [6]. There are 4,762, and 5,917 gene sets (Supplementary Table C.2) in these categories, respectively. Unlike most of the existing pathway-based studies [93, 121, 125, 150] that identify pathways with significant differences between the group of tumor samples and the group of normal samples, we intended to develop a method that focuses on pathway behavior at the individual patient level, and to identify pathways in which departure from its norm has significant implication for patients' clinical outcomes. To achieve this, we developed a new computational approach named iPath. There are three major steps in iPath: First, for each individual patient and pathway, we calculate an individual-level ES (iES), analogous to the ES used in GSEA. Then, based on the iES, we dichotomize all tumor samples into two groups: normal-like and perturbed. Finally, we conduct survival analyses to compare whether the two groups of patients show differences in terms of their overall survival. Figure 4.1b illustrates the main workflow of iPath. We demonstrate that pathways identified by iPath have intimate connections with other biological and clinical properties, including somatic mutations, cancer subtypes, and pathology imaging features.

Furthermore, we investigated whether the expression pattern reflected in the pathway's iES values could illuminate the heterogeneity among different cancer types. Using the 4,762 gene sets from the C2 category, we plotted t-distributed stochastic neighbor embedding (t-SNE [101]) for all samples across 14 cancer types (Figure 4.1c). From the t-SNE plot, we observed that samples from the same tumor type (dots with the same color) tend to cluster together, indicating that iES values are highly informative in terms of the distinct pattern in their expression profiles. As expected, we found that three clusters of kidney cancer types— Kidney renal papillary cell carcinoma (KIRP), Kidney renal clear cell carcinoma (KIRC) and Kidney Chromophobe (KICH)—are located together, and two clusters of lung cancer types— Lung squamous cell carcinoma (LUSC) and Lung adenocarcinoma (LUAD)—are located next to each other. Breast invasive carcinoma (BRCA) shows the greatest spread, and Prostate adenocarcinoma (PRAD) shows multiple cluster formations indicating potential subtypes.

4.3.2 Identifying perturbed pathways

For a specific cancer type and a specific pathway, we classify each tumor sample as either normal-like or perturbed. The latter means the gene expression pattern of this pathway is significantly deviated from that of a healthy, normal sample. We hypothesized that in any given tumor sample, multiple key pathways were perturbed. An important consideration is how many pathways are perturbed in a tumor sample, and whether these numbers vary by tumor types. From our comprehensive survey on pathways belonging to the C2 category of MSigDB, we found that there was remarkable diversity among the 14 tumor types in terms of the average percentage of perturbed pathways per patient (Figure 4.2a). LUSC shows the highest proportions (32%) of perturbed pathways whereas PRAD shows the lowest proportions (9.6%). Interestingly, for the 14 tumor types, the proportions of tumor samples showing perturbation averaged across pathways follow a similar order, but with much less variation among different tumor types (Figure 4.2b).

The MSigDB Hallmark gene set is a collection of 50 “refined” gene sets, curated from numerous “founder” sets, each representing a specific biological process or state and demonstrating coherent expression [96]. The Hallmark set contains numerous well-known signaling pathways that have long been implicated in tumorigenesis and tumor progression, including the p53 pathway, Wnt, Notch and PI3K pathways. It is of great interest to examine the expression pattern of these pathways at the individual tumor sample level. To achieve this, we applied iPath to the 50 pathways in the Hallmark category. For each of the 14 cancer types, we calculated the percentage of tumor samples that are perturbed for each Hallmark pathway. As expected, we found that some pathways such as apoptosis and myogenesis a perturbed in more than half of the samples across multiple cancer types, while some other pathways, including PI3K and KRAS and MTORC1, are perturbed in more than half of the samples in selected cancer types.

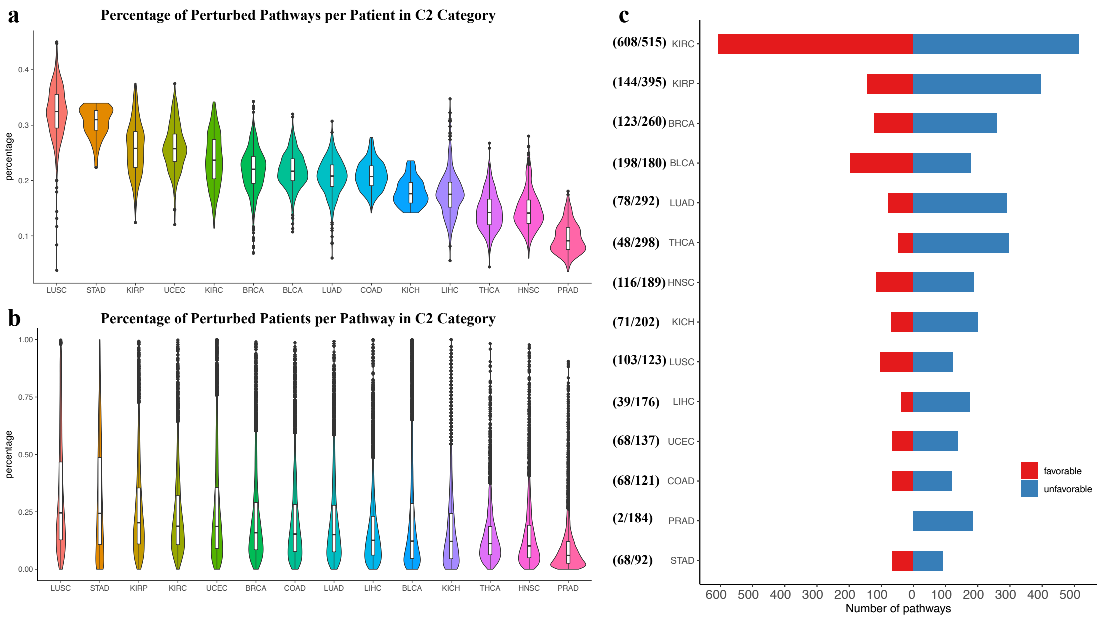


Figure 4.2: **a**.Survey of the proportions of perturbed pathways in the 14 cancer types. All analyses are performed using the C2 category pathways which includes 4,729 gene sets. a. The violin plot of parentage for perturbed pathways: the average proportions of the perturbed C2 category pathways among all tumor samples within each of the 14 cancer types are ranked. **b**. The violin plot of percentage for perturbed patients: the average proportions of tumor samples across C2 category pathways for each of the 14 cancer types are ranked. **c**. The breakdown of favorable/unfavorable prognostic biomarker pathways in these 14 cancer types.

4.3.3 Identifying prognostic biomarker pathways

In this study, we applied iPath using 10,679 gene sets to 6,198 tumor samples across 14 different cancer types. A pathway is named a prognostic biomarker pathway for a given cancer type if the Kaplan-Meier survival analysis yields a significant log-rank p-value less than 0.05. Here we used the same significance threshold used by Uhlen et al. to identify candidate prognostic genes [145]. We later applied more stringent criteria to focus on the most promising prognostic biomarker pathways. Out of these 149,506 gene set / cancer type combinations, 10,592 of them (7.1%) are deemed prognostic: 4,898 (7.3%) in the C2 category, 5,694 (6.9%) in the GO category.

Among all the identified prognostic biomarker pathways, we further classified them by clinical outcomes into two subclasses: favorable prognostic biomarker pathways and unfavorable prognostic biomarker pathways. Favorable prognostic biomarker pathways imply that higher iES values relative to normal samples are correlated with better patient survival outcomes and vice versa. Unfavorable prognostic biomarker pathways designate the opposite. Among the 4,898 C2 pathway-cancer type combinations that deemed significant in predicting patient outcome. 1,734 (35.4%) are favorable prognostic biomarker pathways and 3,164 (64.6%) are unfavorable prognostic biomarker pathways, respectively. The ratios of favorable to unfavorable prognostic biomarker pathways varied among the 14 different types of cancer. Figure 4.2c illustrates the number of prognostic biomarker pathways and the two subtypes for the 14 cancer types.

In order to concentrate on the most promising results from this long list, we here present the most significant gene sets identified by iPath, using a combination of stringent criteria including the q-value (false discovery rate (FDR)) being less than 0.15 and the number of genes in the gene set being less than 100 in order to focus on more specific pathways. Excluding KIRC which showed much more prognostic biomarker pathways than others, on average, about 70 prognostic biomarker pathways (out of total of 10,679 pathways, less than 1%) were found for each cancer type.

4.3.4 Pan-cancer view on prognostic biomarker pathways identified

We examined the number of significant prognostic biomarker pathways identified among different cancer types. We found that there was remarkable imbalance among these cancer types in terms of the number of such pathways identified. Most of the significant pathways were found in three kidney cancer types: KIRC, KIRP and KICH. A few occurred for LUAD, PRAD, THCA, BLCA, and BRCA. Almost none were found in other cancer types. This could be because the clinical outcomes of different cancer types are quite diverse. It is also of interest to discover what proportions of the prognostic biomarker pathways overlap across cancer types. To find out, we calculated the Jaccard similarity between two lists of prognostic biomarker pathways for every pair of cancer types. We found that the similarity level is very low, except for the three kidney cancer types (KICH, KIRC and KIRP), meaning most cancer types have very few shared pathways. In other words, the majority of prognostic biomarker pathways are cancer-type-specific. Our findings are consistent with the results presented in Uhlen et al. and highlight the extensive diversity in different types of human malignancy.

Compared to other cancer types, very few prognostic biomarker pathways were identified with breast cancer. This is somewhat surprising, since multiple well-established pathways are known to play critical roles in the tumorigenesis and progression of breast cancer [2, 27, 44, 71, 76, 106, 159]. One possible reason for this is the substantial pathological differences among the four major subtypes of breast cancer: Luminal A, Luminal B, HER2+, and Basal like. Supporting this hypothesis is the fact that the proportion of patients with such pathway alterations in these four breast cancer subtypes (third column) varies greatly (Figure. 3e, f). Given this observation, we were prompted to explore whether the disruption of a particular pathway preferentially occurs in a particular subtype of breast cancer. We then applied iPath to the four BRCA subtypes separately and identified 8, 10, 3, and 16 significant biomarker pathways (using FDR cutoff q-value ≤ 0.15) in the four subtypes, respectively.

4.3.5 Selected prognostic biomarker pathways identified

There were many interesting prognostic biomarker pathways identified by iPath. For example, in various kidney cancer types, including KIRP, KIRC, and KICH, many prognostic biomarker pathways from the GO collection in MSigDB were found to be related to the cell cycle (Supplementary Figure C.3). Recent studies have shown that cell cycle progression gene signatures are significant, independent predictors of long-term outcomes for patients with renal clear cell carcinoma [105] or related biomarkers [22]. Smaller studies on TCGA KIRC datasets have substantiated this [7, 48]. Our findings are also consistent with reports of cell cycle-related biomarkers for KIRP [56] and KICH [169].

In BRCA, multiple REACTOME pathways were identified by iPath as prognostic biomarker pathways. For the *REACTOME P38MAPK EVENTS* pathway, our results are consistent with studies showing that p38 MAPK signaling drives resistance to key breast cancer drugs including trastuzumab resistance in HER2+ breast cancer [34] and tamoxifen resistance in luminal breast cancer [68]. Identification of the *REACTOME RAF MAP KINASE CASCADE* pathway as a biomarker is supported by a recent study that found that a transcriptional signature called the MAPK Pathway Activity Score (MPAS) is associated with patient outcome in *ERBB2-positive breast cancer* [150]. The prognostic nature of the gene set *FARMER BREAST CANCER APOCRINE VS LUMINAL* (Figure C.2) is logical, given the fact that this signature discriminates between AR+ basal breast cancers with poor outcomes and AR+ luminal breast cancers with much better outcomes [40].

Besides the C2 category gene set database, we also identified GO term "GO CELLULAR RESPONSE TO THYROID HORMONE STIMULUS" (Supplementary Figure C.2c), which contains 13 genes, as a prognostic biomarker pathway for KIRP (Supplementary Fiure C.2d). Thyroid hormone has long been linked to the pathophysiology of various cancer types [83]. While this pathway is not one of the top enriched pathways according to classical GSEA analysis (p = 0.2112), iPath determined that a small subset of 22 KIRP patients with much reduced expression in this pathway led to significant poor clinical prognosis, suggesting that any intervention that increases the impression of this pathway may benefit this group of patients. Another GO term that has been identified as a prognostic biomarker pathway is " GO

ATP DEPENDENT MICROTUBULE MOTOR ACTIVITY" in KICH, Supplementary Figure C.2d). Cell proliferation is a hallmark of almost all tumors, and it is well known that microtubules play an important role [21] in mitosis. Interestingly, for this pathway we found that individuals with reduced expression levels have much better clinical prognoses, thus it is an unfavorable prognostic biomarker pathway. Given this, it is likely that antimitotic therapies that impede mitosis-specific microtubule functions through inhibiting motor proteins [120] may benefit patients with high expression of this gene set.

4.3.6 Links to distinct patterns shown in pathology imaging

Pathology imaging has long been regarded as the gold standard diagnostic tool in clinical oncology. We conjectured that individual-level expression profiles of a pathway could help to distinguish subtle tumor characteristics hidden in pathology imaging. To investigate, we used the gene set "FARMER BREAST CANCER APOCRINE VS LUMINAL", one of the most significant prognostic biomarker pathways identified in BRCA, as an example. We selected three tumor samples from the far end of both the normal-like group and the perturbed group, and obtained their corresponding pathology images from the cancer digital slide archive [51]. The image of the three normal-like samples and three perturbed samples are shown in the second column in Figure 4.3a and Figure 4.3b respectively. Among the six pathology images, the luminal type tumor shows well differentiated morphology with well-formed tumor lumen, low to intermediate nuclear grade and low mitotic features. The androgen type shows higher grade, with poorly-formed tumor lumen, intermediate to high nuclear grade, and focal tumoral necrosis. To confirm this observation, we obtained the ICD-O-3 codes (8500/3 Infiltrating duct adenocarcinoma; 8520/3 Lobular carcinoma) of the top ten and bottom ten samples patients quantified by their iESs. The breakdown of these codes shows a distinct distribution between normal-like and perturbed samples

(Figure 4.3d).

4.3.7 Comparison with GSEA

The core function of iPath is to identify perturbed pathways in every individual tumor sample. In contrast, the classical GSEA method identifies pathways that show differences when comparing two groups of samples, hence only one ES is calculated for each pathway, no matter how many samples there are. Given their differences, a pathway identified by iPath may not have been picked up by GSEA and vice versa. This is possibly because a pathway is perturbed only in one individual sample, and thus unlikely to display a significant difference when tested by GSEA. In other words, iPath is good at identifying perturbed pathways for a small minority group of cancer patients. To illustrate the point, we used breast cancer (BRCA) as an example. We first calculated iES for each pathway in each individual. Using iESs, we applied a Wilcoxon signed-rank test (Wilcoxon, 1945) to each pathway, compared iES values between tumor and normal samples, and used the p-values of the test to rank all pathways. For comparison, we also ran GSEA to obtain a different list of ranked pathways. The top ten pathways that differentiate the iES values of tumor and normal samples are listed in (Supplementary Table C.3), along with their significant levels. The top ten differentiated pathways identified by GSEA are listed in (Supplementary Table C.4), along with the corresponding ranking in the Wilcoxon signed-rank test comparing iES values. We found that two pathways (bold) in the two top ten lists are identical; for the remaining eight pathways, four pathways in the GSEA list are not cancer-related (red), while only two pathways in the iPath list seem not immediately cancer-related (red).

4.3.8 Comparison with other sample-level gene set analysis methods

We compared iPath against existing methods that are capable of measuring expression of a pathway at individual level, namely ssGSEA [10], SLEA [49], Pathifier [35] and GSVA [64]. We adopted the performance comparison study design used in GSVA study inside which the effectiveness of clustering a mixture of tumor and normal samples is compared. In such a study, sample-level ES scores were used to select the most differentiated pathways which in turn were used in the clustering. The details of the performance comparison procedure are presented in the material and methods section.

The performance comparison results are shown in Figure 4.4a. We use adjusted Rand Index (ARI) to measure the clustering performance. Higher ARI indicates better clustering, which can be attributed to better pathways selected by each individual method that calculates sample-level ES scores. Figure 4.4a indicates that iPath approach results in the highest ARI among all methods tested. Pairwise comparison between iPath and the three competing methods using t-test indicates that all the differences are statistically significant.

Additionally, we compared these methods in terms of their ability to consistently detect prognostic biomarker pathways. Briefly, for each method, we selected the most significant pathway in the training data and tested its ability to predict survival in the test data by reporting the c-index. Higher c-index indicates better correlation with the survival outcomes. The results demonstrate the consistency of iPath for identifying the most informative prognostic biomarker pathway across the training and test data. The details of the performance comparison procedures are presented in the material and methods section. The side-by-side boxplots shown in Figure 4.4b again demonstrate the superior performance of iPath. Pairwise tests show that iPath produces significantly higher mean and median c-index values than competing methods.

4.3.9 Comparison with the Human Pathology Atlas

In a recent study, Uhlen et al. developed the Human Pathology Atlas (HPA), in which they adopted a system-level strategy to analyze 17 major cancer types with a focus on mining characteristic genes with respect to clinical outcomes. This method is based on genome-wide transcriptomic data and searches for prognostic genes whose top 20% or bottom 20% expression values, measured in FPKM, can stratify patient cohorts with significant survival differences (p ¡ 0.001). Both HPA and iPath aim to identify prognostic biomarkers from transcriptome data. However, HPA relies on individual genes, while iPath focus on pathways. Hence, it is of great interest to compare their performance. Due to the substantial noise that is ubiquitous in high throughput technologies, we hypothesized that a pathway-based approach would be more robust and effective. To test our hypothesis, we applied both HPA and iPath to renal papillary carcinoma (KIRP). First, we used the p-value threshold of 0.05 to determine whether a pathway or a gene would be considered prognostic by either approach (Figure 4.4c and Figure 4.4d). Then, when using a more stringent threshold (q value = 0.05), we found no significant prognostic biomarker genes (Figure 4.4f), but lots of significant prognostic biomarker pathways Figure 4.4e). Tests conducted on KIRC gave similar results. These results indicate that the pathway-level biomarkers are more sensitive than the gene-level biomarkers.

A related question is whether member genes of a prognostic biomarker pathway are also prognostic biomarker genes. We found that this is not true in most cases. For some significant prognostic biomarker pathways identified by iPath, none of their member genes are prognostic genes according to HPA. In other words, at the individual gene level, many genes are not prognostic biomarkers themselves, but their expression pattern as a whole can accurately predict a patients' clinical outcome.

“REACTOME RAF MAP KINASE CASCADE”, for instance, is one of the significant biomarker pathways identified in BRCA (Figure 4.4g), but no gene inside this pathway correlates well with survival outcome (Figure 4.4h). This is reminiscent of the scenario in which a pathway is identified by GSEA as significant but none of its member genes show differential expression. Taken all together, we believe that pathway-based biomarkers are more robust and effective than single-gene based biomarkers.

4.3.10 Connection with the mutations in cancer driver genes

Progressive accumulation of somatic mutations over time in crucial oncogenes or tumor-suppressor genes has been implicated in many cancer types [103, 72, 86, 176]. Recently, the somatic mutation statuses of 127 genes have been shown to have significant effects on patient survival [72]. With the identification of prognostic biomarker pathways using iPath, a natural question is whether the perturbed state of prognostic biomarker pathways is linked to somatic mutations occurring in cancer driver genes. To answer this, given a pathway and a cancer driver gene, we first constructed a contingency table dividing samples according to their normal-like/perturbed status for the pathway, and the mutation profile (present or absent) in the cancer gene. We then conducted a Fisher’s exact test to identify incidence of co-occurrence of the two events. A binary heatmap indicating whether a significant (p ¡ 0.05, marked in the red block) connection between the top selected pathways and top mutated gene is shown in Figure 4.5. We found that indeed somatic mutation in key cancer driver genes and perturbed prognostic biomarker pathways are often co-occurring events. In breast cancer (BRCA), we observed that NOTCH1 and E-cadherin (CDH1) are associated with metastasis-related gene sets (Figure 4.5a), which is consistent with findings reported in the literature on NOTCH1 signaling [87] and CDH1 [32, 118]. In lung adenocarcinoma (LUAD) (Figure 4.5b), we identified a couple of histone-lysine

N-methyltransferase genes (MLL2 and MLL4) that are related to the top significant pathways found by iPath, and these genes are reportedly clustered in LUAD [72]. We showed that PIK3CA is correlated with one early cell cycle pathway, which demonstrates that PIK3CA deregulation serving as an early event precedes genome doubling in BRCA [13] and colorectal adenocarcinoma [20].

4.4 Discussion

We here describe iPath, a computational tool to identify perturbed pathways found in individual tumor samples. Unlike individual genes, the collection of functionally related genes in a pre-defined pathway provides a more robust assessment of the changes that affect key biological functions in tumor samples. The advantages of using pathways over individual genes have been well documented in the analysis of noisy high-throughput data [117] and more recently, as biomarkers [38]. What makes iPath unique is its ability to provide such an assessment one sample at a time. This is significant, because substantial heterogeneity among tumor genomes suggests that it is common for a critical pathway to be perturbed in only a few tumor samples. As a result, it is highly unlikely that these pathways will be identified by traditional GSEA. On the other hand, iPath can identify perturbed pathways, even if such disruption only occurs in a small subset of tumor samples. In short, iPath promises to improve patient care by enabling oncologists to develop more effective personalized treatment strategies with fewer side-effects.

To demonstrate the effectiveness of iPath, we conducted a comprehensive pancancer study across 14 different cancer types with more than 6,000 tumor samples. For each cancer type, iPath identified about 70 prognostic biomarker pathways on average, many of them showed promising biological interpretations. We also validated the top prognostic biomarker pathways using SurvExpress, an online biomarker validation tool. There are two types of prognostic biomarker pathways: favorable and unfavorable. Favorable pathways account for one third of all the biomarker pathways. These pathways can be used to identify patients with better prognostics so they can be spared for unnecessary adjuvant therapies.

Our pan-cancer study using iPath yields two interesting results. First, we found quite a few pathways or gene sets are potential prognostic biomarkers for most of the cancer types we studied. However, for the vast majority of these biomarker pathways, they are perturbed in only a small fraction of all the patients. Second, for any given pair of cancer types, there is little overlap among the two lists of prognostic biomarker pathways. Our findings highlight the fact that cancer is a highly heterogeneous disease therefore personalized treatment strategy is key for effective care for cancer patients. The present study is conducted on RNA-seq gene expression profiles but iPath can be applied to other omics data such as microarray data.

The core of iPath is the iES, a single continuous value between -1 and 1, calculated for every pathway, which may contain hundreds of genes, in each individual sample. We believe this is a powerful way to summarize the status of a pathway, or provide a big-picture view of pathway changes at single sample resolution. Our analysis has shown that iES is informative and sometimes predictive of patients' clinical features and prospects. Because it measures the level of the pathway's deviation from the norm, we find it to be more sensitive than scores calculated based on the actual expression level of member genes, used by ssGSEA and similar tools [35, 49, 98, 148].

Scientists have already identified many biomarker genes for various cancer types, for example, thousands of prognostic genes have been identified in a recent study of Uhlen et al., why it is important to identify prognostic biomarker pathways? In the present study, we found that compared to single gene biomarkers, pathway-based biomarkers are more robust with better separation power, which gives clinicians more confidence in separating patients to different risk groups, and assign treatment strategies accordingly. Furthermore, given that they represent well-curated biological pathways, easier to interpret, and hence more likely to be informative and meaningful to clinicians. Another key advantage of pathway-based biomarkers is that there are drugs that are specifically targeting specific pathways. For example, it is likely that MAPK perturbed patients will benefit more from MAPK inhibitor drugs. This is beyond the scope of our current study and we plan to pursue this in future works.

IPath can be applied broadly to other types of cancer, for any given individual sample, as long as there are corresponding normal samples that can be used to establish homeostasis. Thus, iPath is a formidable resource for unraveling the large-scale changes that occur in a small minority of patients, even a single patient. Therefore, it is an ideal tool for personalized or precision oncology. To illustrate its potential: some drugs have been developed to specifically target a kinase and its downstream genes [17, 111]. Using iPath, we can group the drug target and its downstream genes together and identify patients with elevated expression in this gene set; such patients may benefit the most from this targeted therapy. We believe iPath can potentially provide fresh perspectives on patient selection and prognostic prediction.

In this study, we only examined individual pathways to try to establish whether a given pathway is predictive of a clinical outcome. For prediction purposes, we could consider multiple pathways jointly, which may produce better prediction performance. This represents one potential future research direction for the continuous development of iPath.

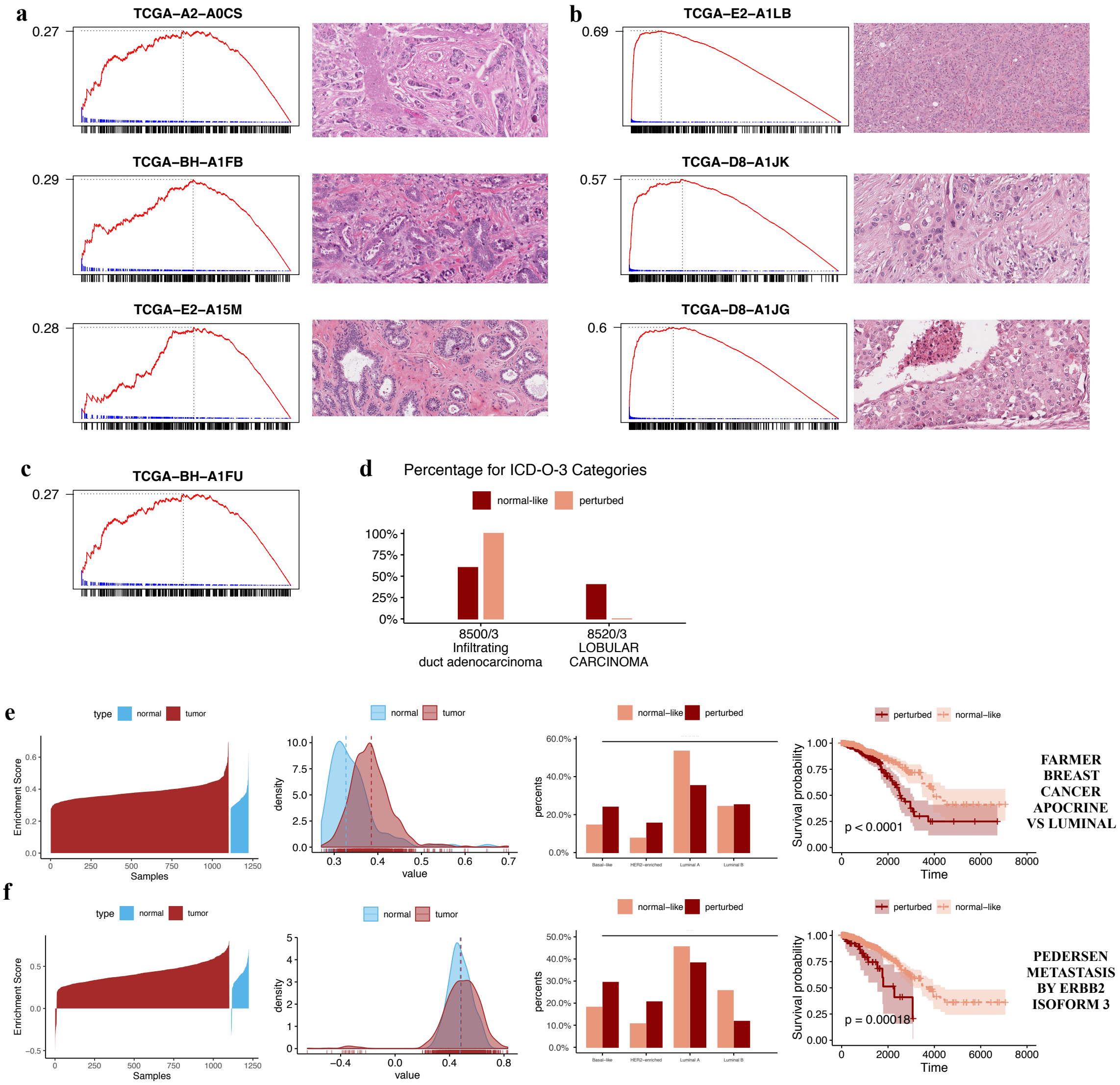


Figure 3

Figure 4.3: Demonstration of an example prognostic biomarker pathway (FARMER BREAST CANCER APOCRINE VS LUMINAL) in BRCA. **a**. Enrichment plots of the pathway and corresponding pathology images of three samples labeled "normal-like". **b**. Enrichment plots and corresponding pathology images of three samples labeled "perturbed" **c**. Enrichment plot of the pathway of a normal sample. **d**. Breakdown of the ICD-O-3 categories for the top ten perturbed (highest iES value) and bottom ten normal-like (lowest iES values) patient samples. **e-f**. Visual summary of two example pathways including: the waterfall plot shows that the iES in tumor samples marked in red and normal samples marked in blue; the density plot shows that overall tumor samples are up-regulated, because the mean of the tumor sample GSEA scores is higher than normal sample iES. The distribution of perturbed and normal-like tumors across the four subtypes of breast cancer is listed in the third column. The Kaplan-Meier plot indicates a significant survival difference for the perturbed and normal-like tumor samples.

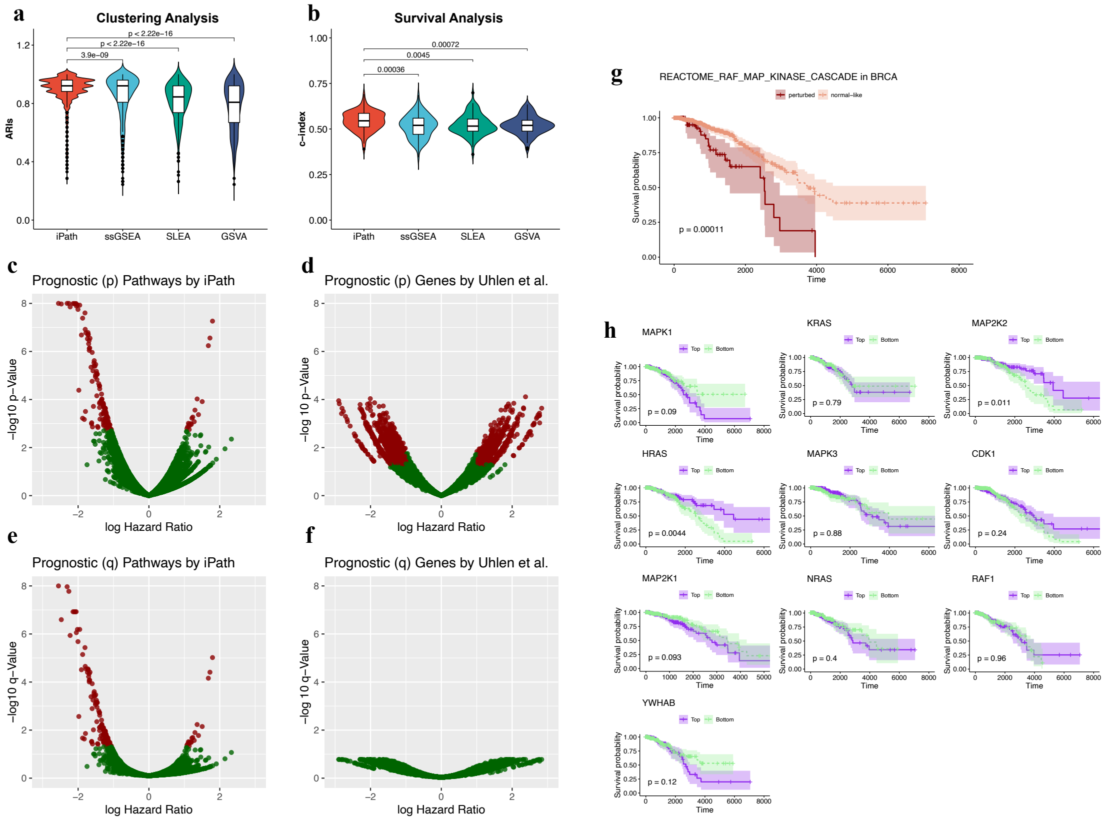
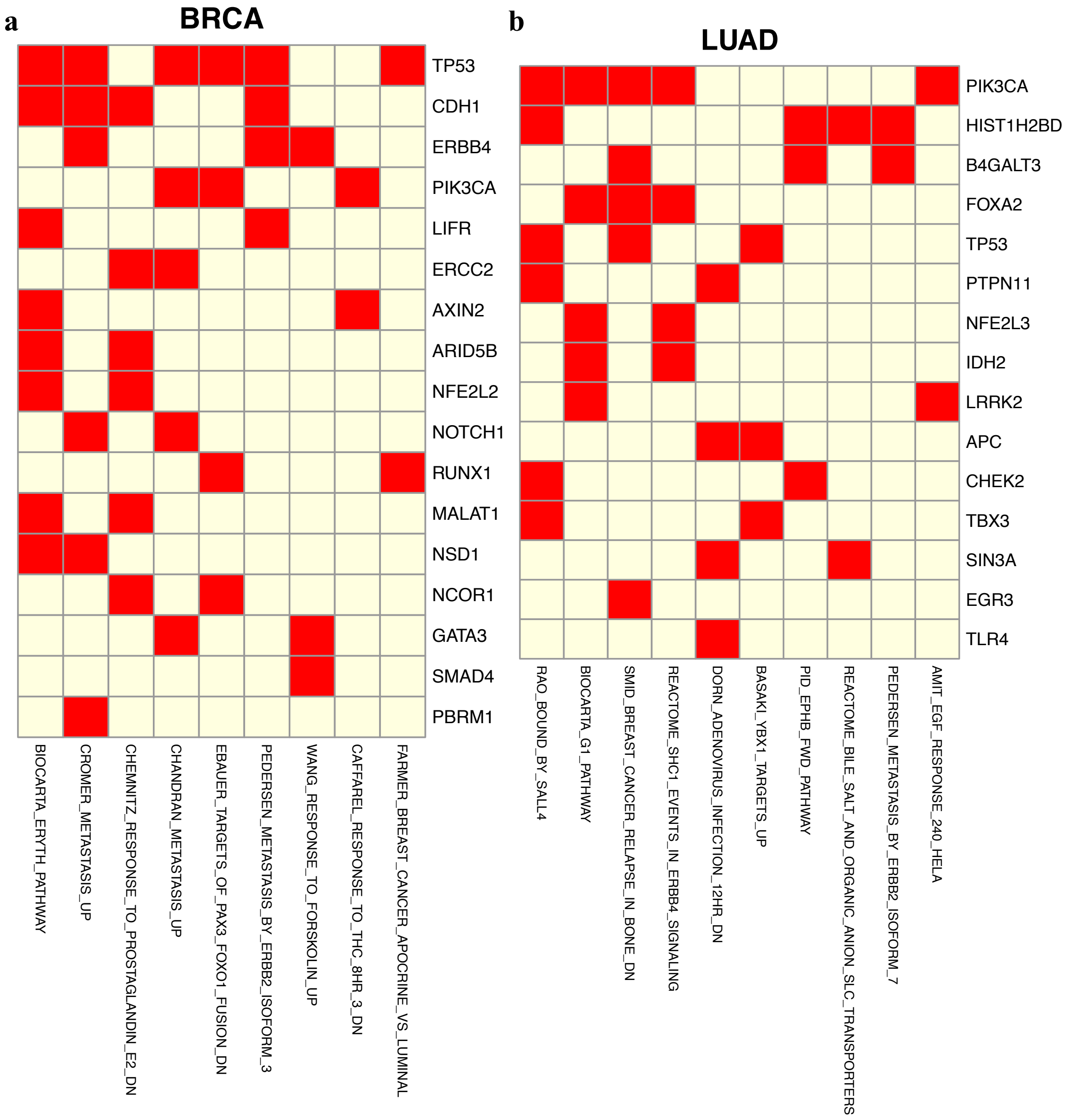


Figure 4.4: Comparisons between iPath and other sample-level gene set analysis methods including ssGSEA, SLEA, and GSVA, and comparisons between pathway biomarkers and individual gene biomarkers. **a**. Comparison of hierarchical clustering results in terms of separating tumor and normal samples from the enrichment score matrix. The hierarchical clustering accuracy is measured by ARI values. As demonstrated in the violin plot, the clustering accuracy from iPath is significantly higher than the other methods. **b**. Comparison of survival analysis results using concordance index. It shows that iPath can sleet the most significant pathways that lead to the highest concordance in the violin plot. **c**. The volcano plots for the prognostic biomarker pathways. The significance threshold is set at p-value 0.05 (log10 (p-value) = 1.4). The prognostics and non-prognostic biomarkers are marked by red and green dots respectively. **d**. The volcano plots for the prognostic biomarker genes. The significance threshold is set at p-value 0.05 (log10 (p-value) = 1.4). **e**. The volcano plots for the prognostic biomarker pathways. The significance threshold is set at q-value 0.05 (log10 (q-value) = 1.4). **f**. The volcano plots for the prognostic biomarker genes. The significance threshold is set at q-value 0.05 (log10 (q-value) = 1.4). **g**. The Kaplan-Meier plot of prognostic biomarker pathway "REACTOME RAF MAP KINASE CASCADE" in BRCA. **h**. The Kaplan-Meier plots of the member genes of the "REACTOME RAF MAP KINASE CASCADE" pathway in BRCA.



	BIOCARTA_ERYTH_PATHWAY	CROMER_METASTASIS_UP	CHEMNITZ_RESPONSE_TO_PROSTAGLANDIN_E2_DN	CHANDRAN_METASTASIS_UP	EBAUER_TARGETS_OF_PAX3_FOXO1_FUSION_DN	PEDERSEN_METASTASIS_BY_ERBB2_ISOFORM_3	WANG_RESPONSE_TO_FORSKOLIN_UP	CAFFAREL_RESPONSE_TO_THC_8HR_3_DN	FARMER_BREAST_CANCER_APOCRINE_VS_LUMINAL
TP53	🟥	🟥		🟥	🟥	🟥			🟥
CDH1	🟥	🟥	🟥			🟥			
ERBB4		🟥				🟥	🟥		
PIK3CA				🟥	🟥			🟥	
LIFR	🟥					🟥			
ERCC2			🟥	🟥					
AXIN2	🟥							🟥	
ARID5B	🟥		🟥						
NFE2L2	🟥		🟥						
NOTCH1		🟥		🟥					
RUNX1					🟥				🟥
MALAT1	🟥		🟥						
NSD1	🟥	🟥							
NCOR1			🟥		🟥				
GATA3				🟥			🟥		
SMAD4							🟥		
PBRM1		🟥							

	RAO_BOUND_BY_SALL4	BIOCARTA_G1_PATHWAY	SMID_BREAST_CANCER_RELAPSE_IN_BONE_DN	REACTOME_SHC1_EVENTS_IN_ERBB4_SIGNALING	DORN_ADENOVIRUS_INFECTION_12HR_DN	BASAKI_YBX1_TARGETS_UP	PID_EPHB_FWD_PATHWAY	REACTOME_BILE_SALT_AND_ORGANIC_ANION_SLC_TRANSPORTERS	PEDERSEN_METASTASIS_BY_ERBB2_ISOFORM_7	AMIT_EGF_RESPONSE_240_HELA
PIK3CA	🟥	🟥	🟥	🟥						🟥
HIST1H2BD	🟥						🟥	🟥	🟥	
B4GALT3			🟥				🟥		🟥	
FOXA2		🟥	🟥	🟥						
TP53	🟥		🟥			🟥				
PTPN11	🟥				🟥					
NFE2L3		🟥		🟥						
IDH2		🟥		🟥						
LRRK2		🟥								🟥
APC					🟥	🟥				
CHEK2	🟥						🟥			
TBX3	🟥					🟥				
SIN3A					🟥			🟥		
EGR3			🟥							
TLR4					🟥					

Figure 4.5: The association between prognostic biomarker pathways and somatic mutations of key cancer genes. For each pathway, we classified each tumor sample as either normal-like or perturbed. For each gene, we classified each tumor sample as either mutated or not mutated. Then a Fisher's exact test of association is carried out on the two-by-two contingency table. **a**. Matrix of gene/pathway association in BRCA. Red color indicates significant association. **b**. Matrix of gene/pathway association in LUAD.

Chapter 5

Future research plan

5.1 Single-cell multi-omics data integration

Currently, single-cell biotechnology becomes the tremendous drive for many biological discoveries. With the scale of the dataset increasing, more robust algorithms need to be developed to accommodate with both prediction accuracy and computational efficiency. With multi-omics data (DNA, mRNA, protein) available at single-cell resolution, it becomes more interesting to integrate multiple data source to gain more insights because more data potentially conveys more information. I firmly believe that one of the most important research goals in single cell is still to study the cell type information. That is, accurately knowing the cell type information is essential for conducting the analytical process for single cell research.

One interesting research topic is to design a clustering algorithm specifically for large dataset, which includes more than millions of cells. One addressable approach is to partition large dataset into smaller batches and parallelly perform the clustering for each batch. Then, assembling each piece of the clustering results. So far, particular scRNA-seq clustering algorithm for millions of cells is still understudied. Only a handful tools are available including SHAPR [152], scAIDE [164], mbkmean [57], and fastPG [15]. However, these tools did not provide the specific solution for clustering sub-cell-types which could include hundreds of detailed cell-types. The numbers of cell-types investigated in these existing tools are all less than 40.

With recent largescale pilot studies such as the Human Cell Atlas [116], it is necessary to study each cell based on the characterized clustering labels. Knowing the subtype identity of each cell is helpful to study cell functionalities. A possible and direct approach is to first clustering millions of cells into major classes and further continue clustering within each major class into subtypes. This stepwise clustering thought can be beneficial.

To accommodate multi-omics data, another interesting research topic is to integrate the regulatory information such as single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) [16] and single-cell protein sequencing to better predict the gene regulatory networks. Some tools such as scM&T-seq [24] and CITE-seq [132] provide the feasibility of generating these multi-omics data, and some mathematical tools such as liger [158] and Seurat3 allow to analytically integrate multi-omics datasets.

5.2 Single-cell spatial transcriptomics

By adding another layer of spatial information to scRNA-seq data, the spatial transcriptomics has transformed our understanding of tissue functional organization and cell-to-cell interactions. The space information of the spot location (coordinates) corresponding each cell on the histology image will help us to better understand the cellular functionalities and cell type purity. Many traditional research topics in original scRNA-seq such as differential expression (DE) analysis and trajectory inference have been modified to incorporate the spatial information. It is worth exploring whether recruiting this spatial information rather than purely scRNA-seq data can improve or alleviate signals in terms of DE analysis and trajectory inference.

Appendix A

Appendix for Chapter 2

A.1 Test datasets

Abbreviation	reference	#classes	Brief Introduction
Baron	GSE84133	13	Map of the human and mouse pancreas reveals Inter- and Intra-cell population structure.
Close	GSE93593	41	human interneuron differentiation
Darmanis	GSE67835	9	Cellular complexity of the adult and fetal human brain.
Deng	GSE45719	6	Adult liver: from zygote to late blastocyst
Goolam	E-MTAB-3321	5	Transcriptional heterogeneities in pre-implantation mouse embryos
LGd	Allen Brain Map	4	Mouse dorsolateral geniculate complex
Nestorowa	GSE81682	9	PBMCs (10X genomics)
Romanov	GSE74672	7	distinct dopamine neuronal subtypes
Tasic	GSE71585	18	cellular taxonomy of one cortical region, primary visual cortex.
Treutlein	GSE52583	5	Mouse lung epithelium cells
Yan	GSE36552	6	human preimplantation embryos and embryonic stem cells
Zheng	10x Genomics	8	PBMCs (10X genomics)

Table A.1: Test datasets for comparing different feature selections, and demonstrating that features selected by FEAST can assist scRNA-seq. For Nestorowa, we use the samples from one batch. For LGd dataset, we use the most abundant 6 cell types (1592 cells).

A.2 Test datasets with large number of cells

Dataset	# of cells	# of cell types	Introduction
Chen	14437	45	adult mouse hypothalamus
Shekhar	27499	19	neurons and mouse retinal bipolar cells
Macosko	44808	19	mouse retinal cells

Table A.2: Test large datasets for comparing feature selection. For Chen dataset, we use 5 most abundant cell types including 9596 cells.

A.3 The comparison of F-statistics distributions

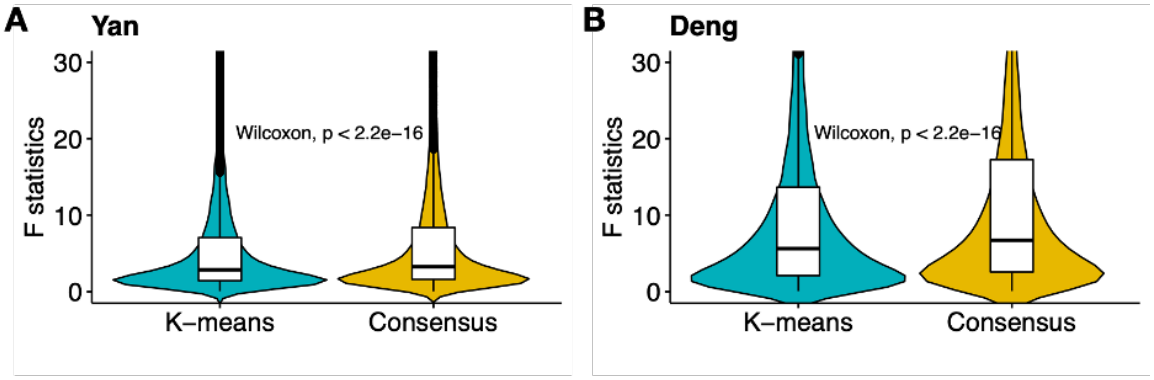


Figure A.1: The distributions of F-statistics from two approaches: k-means and consensus clustering implemented in FEAST. The F-statistics from FEAST are significantly higher than those from k-means, indicating stronger signal to noise ratio from FEAST. The significances (p-values) are obtained from Wilconxon test.

A.4 Compare FEAST to other feature selection approaches

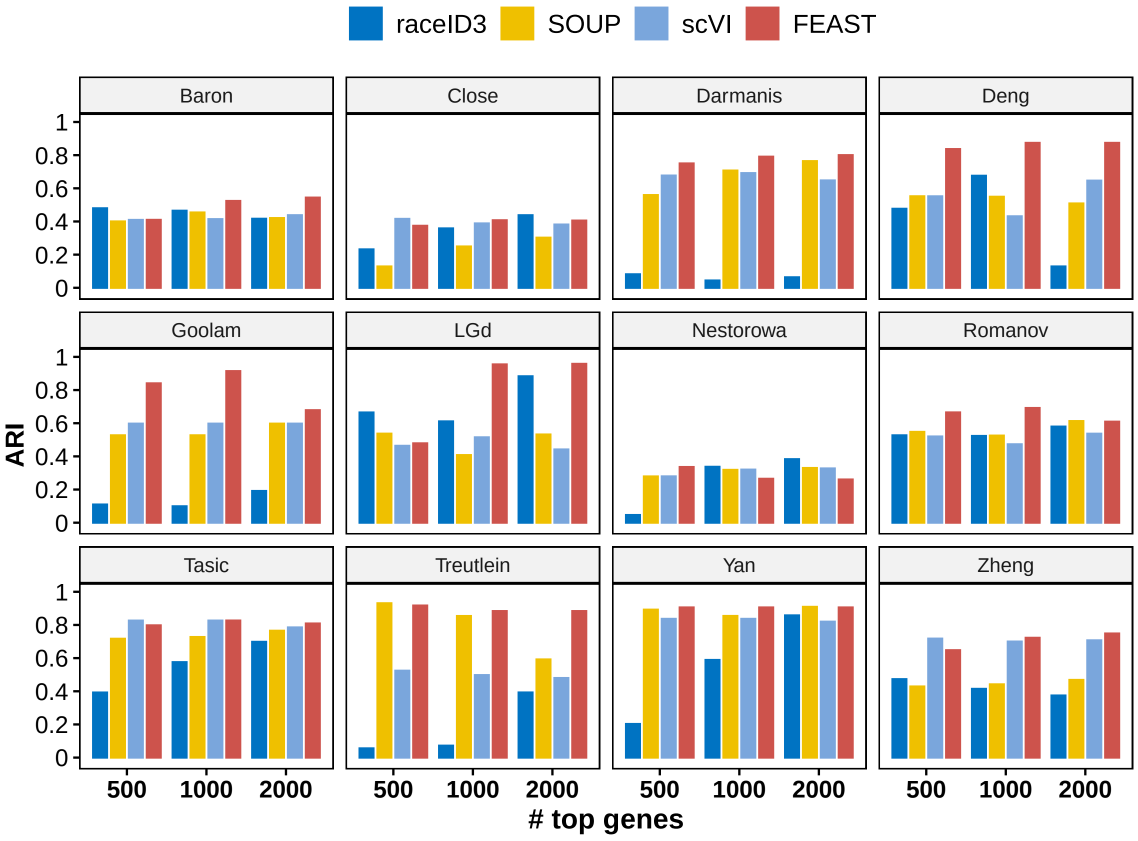


Figure A.2: The comparison of the feature selection methods. We benchmark FEAST with other three unsupervised feature selection procedures implemented in raceID3, SOUP, and scVI. Specifically, we adopted the fitbackground function inside raceID3 package to rank the features. For SOUP, it combines the a feature set by Gini index and Sparse PCA. To rank the features, we adopted the SPCAselect function inside SOUP. For scVI, we order the features based on variance. In each test dataset, we select the top 500, 1000, and 2000 features from each criterion followed by SC3 clustering. FEAST outperforms the other methods by showing the highest ARI values in almost scenarios and datasets.

A.5 Top features selected by CV and Kurtosis

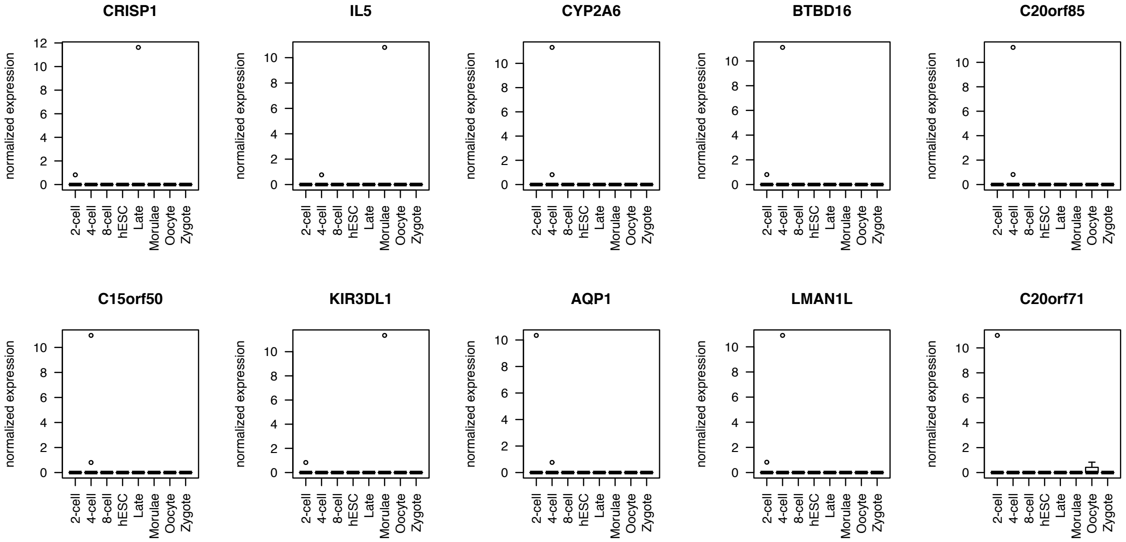


Figure A.3: Top 10 genes selected by CV and Kurtosis. It is interesting to observe that the top 10 feature genes selected by these two approaches are the same in the Deng dataset. These top features have one common character that they only show expression in one or two cells but remain very low expression (usually 0) in the rest of the cells. From the benchmark comparison, it shows that these top features are too sparse to contribute to a better clustering accuracy.

A.6 Feature set validation by TSCAN clustering outcomes

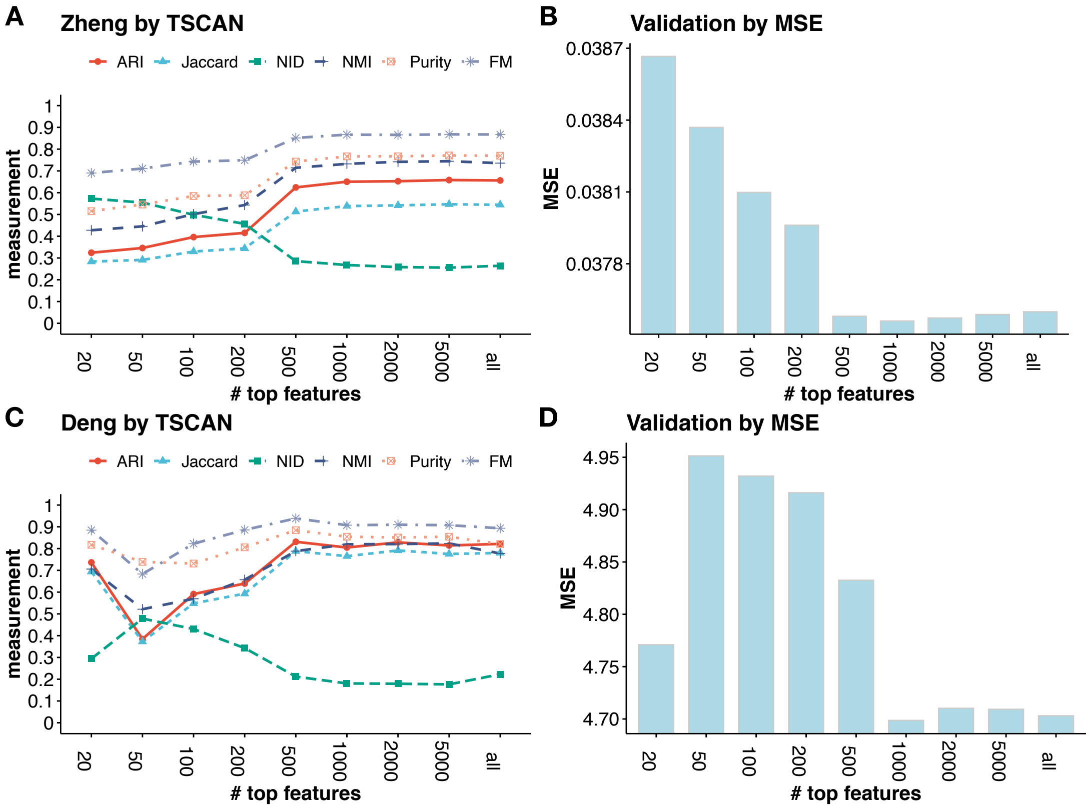


Figure A.4: Test feature set validation on Zheng (A and B) and Deng (C and D) datasets. This time, we use TSCAN to obtain the clustering groups as demonstrated in A and C. Then, we use MSE criterion to determine an optimal feature set (B and D). In Zheng dataset, we found that the feature set (1000 genes) associated with smallest MSE matches with the best clustering accuracy result. Similarly, in Deng dataset, the optimal feature set determined by MSE closes to the best feature set verified by clustering accuracy.

A.7 Features selected by FEAST Improves the clustering accuracy for TSCAN, SIMLR, and SHARP

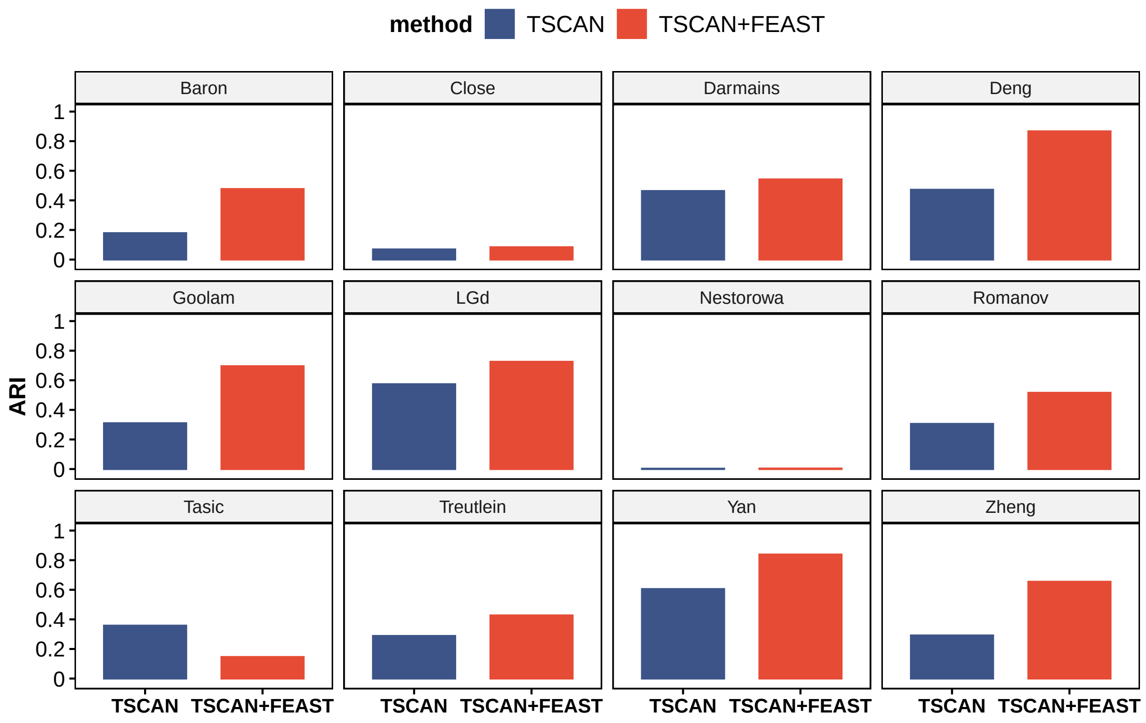


Figure A.5: Test TSCAN with selected feature on the collected datasets. We use FEAST to obtain an optimal feature set. The feature set is determined by the validation process across top-m (m=500, 1000, 2000) feature cases. Then, the selected features are fed into the TSCAN algorithm. To compare the original TSCAN and the TSCAN with specified features selected by FEAST, we calculate the adjusted rand index (ARI) to measure the clustering accuracy. Bars show the ARI values for original TSCAN (blue) and TSCAN using FEAST features (red).

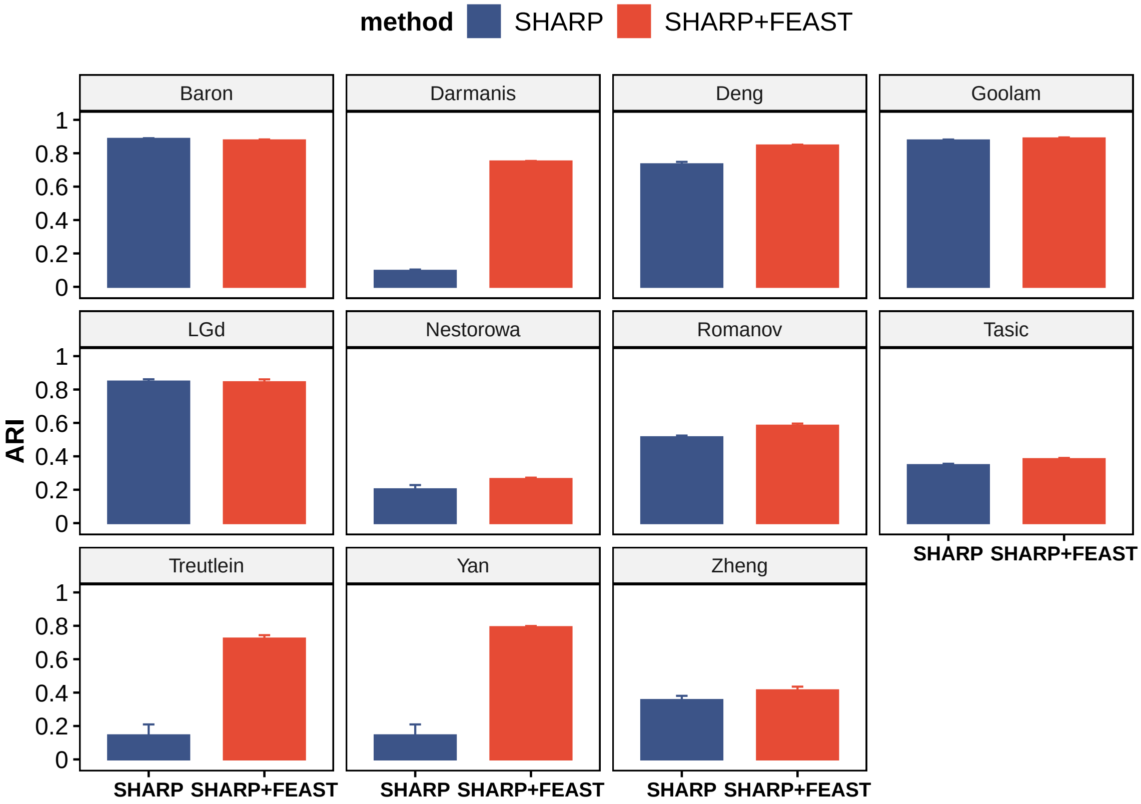


Figure A.6: Similar to Figure A.5, but using SHARP as clustering method. SHARP is based on random projection algorithm, which will lead to different clustering results from different runs. Thus, for each dataset, we run 50 times and report the mean of clustering accuracies and the standard errors.

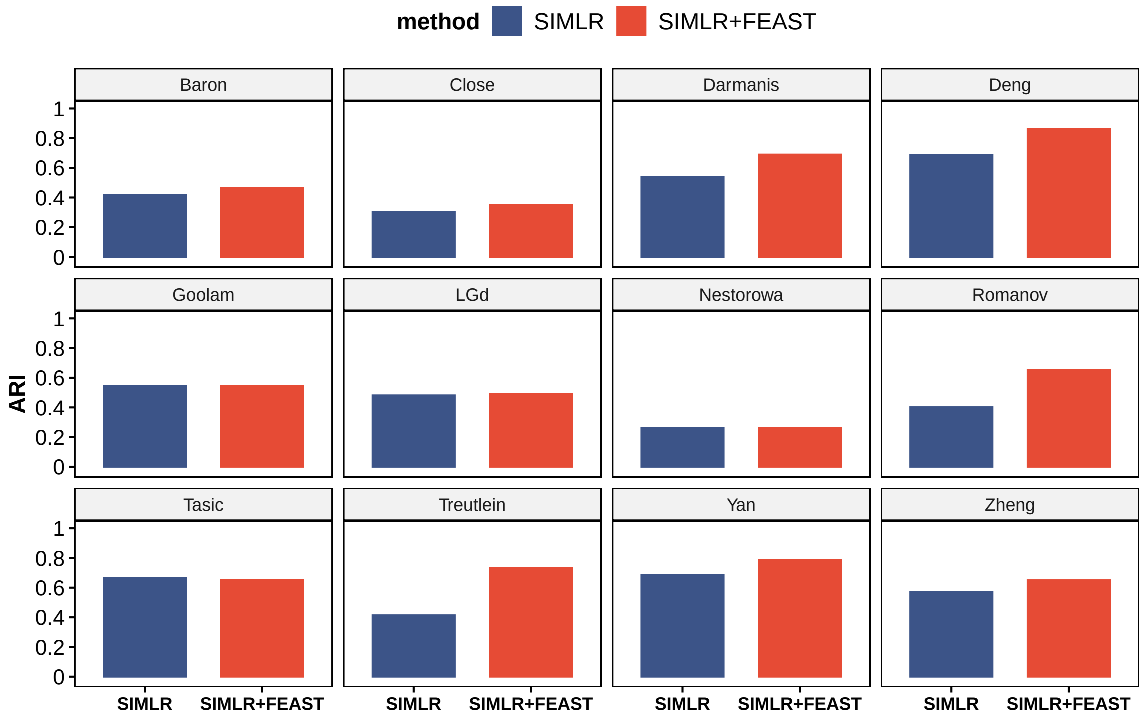


Figure A.7: Similar to Figure A.5, but using SIMLR as clustering method. SHARP is based on random projection algorithm, which will lead to different clustering results from different runs. Thus, for each dataset, we run 50 times and report the mean of clustering accuracies and the standard errors.

A.8 Computational performance of FEAST

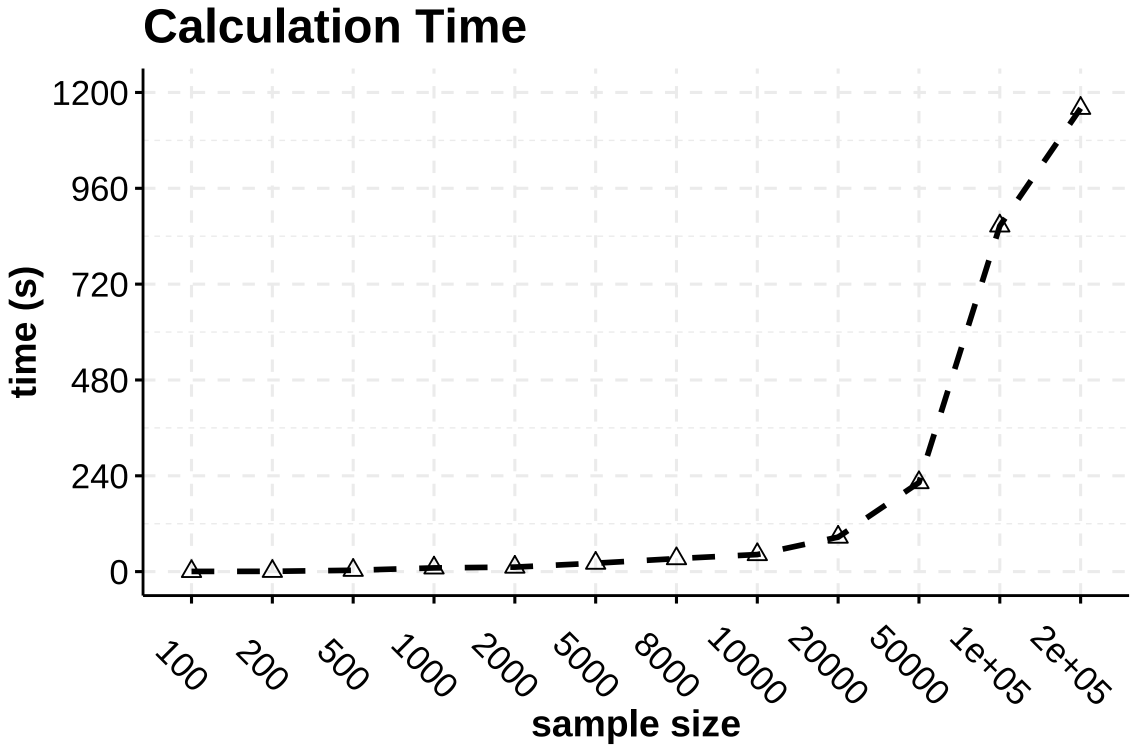


Figure A.8: Computational performance for FEAST. Figure shows the total running time for FEAST for different numbers of cells, without validation procedure. FEAST takes less than 1 minute for 10,000 cells, and less than 4 minutes for 50,000 cells. The running time is profiled on a Macbook Pro with 2.3GHz Intel Core i9 CPU.

Appendix B

Appendix for Chapter 3

B.1 Power analysis for Form II DE with respect to zero fractions

One unique phenomena in scRNA-seq is dropout event which will cause missing values (zero expressions) due to low amount of RNA amplification.Due to the high noise in the scRNA-seq data biologically and technically, it is compelling to investigate how the targeted powers associate to the dropout events. Here, we specifically stratify the Form II DE genes with respect to the zero fractions. Within a simulation of a certain total cell number, we find that stratified powers declines with the higher zero fractions (Figure S12G). Across simulations of different total cells, the overall stratified power curves move up as the increase of the sample sizes. A tendency of improved marginal powers with more cells is clear (Figure S12F). In each case of a certain number of total of cells (Figure S12A-E), the detailed counts of true DE genes and called DE genes for each zero fraction interval are listed.

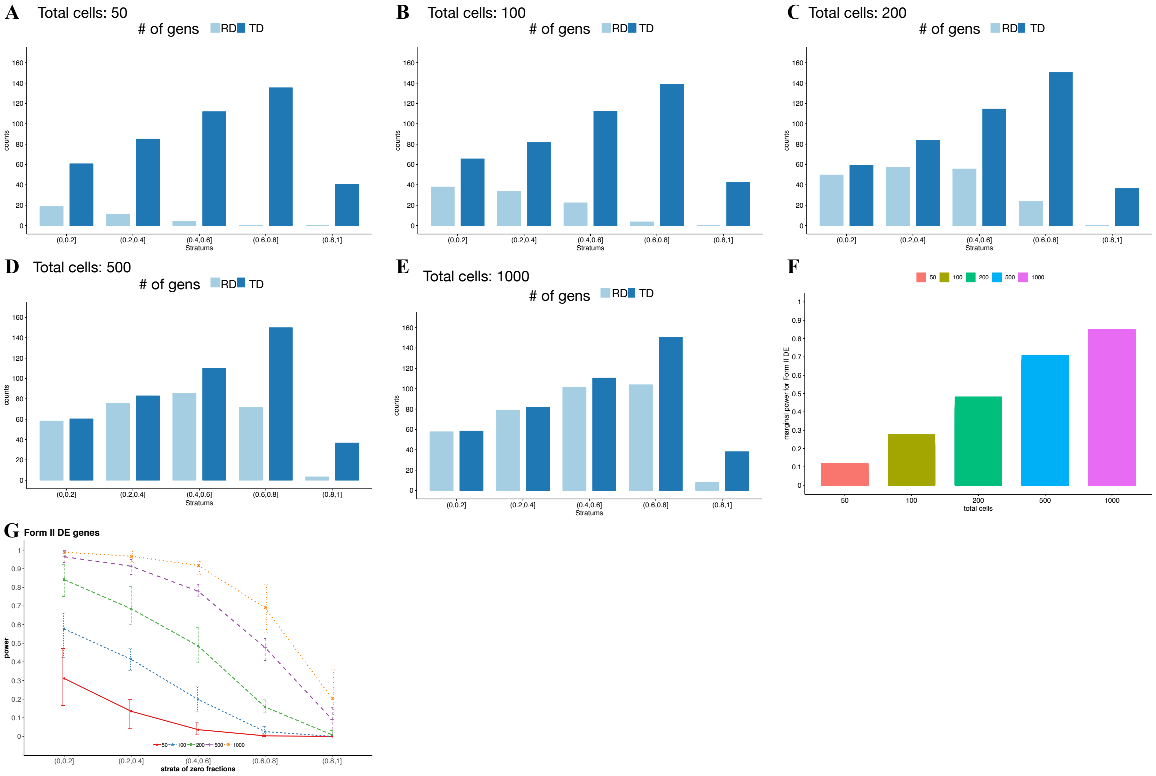


Figure B.1: The power evaluation for Form II DE. Using the template data, we simulate a series of data with different numbers of total cells (50, 100, 200, 500, and 1000). In each simulation, we count the numbers of the true (simulated) DE gens and the recovered DE genes in each stratum (A-E). The stratum is about the zero fractions which is related to dropout rate. We also calculate the marginal power for Form II DE (F). The simulation is repeated 50 times and illustrated in the (G).

B.2 Multiple cell types in Glioblastoma

We tested POWSC on a Glioblastoma (GBM) dataset (GSE57872) to demonstrate how the power changes for comparing different cell types in real case. This scRNA-seq dataset includes 5 individual tumors (MGH26, MGH28, MGH29, MGH30, and MGH31). Sample MGH31 is used as template for this simulation. Proportions for four cell types (cell type 1 to 4) are estimated by SC3, as 0.66, 0.15, 0.1, and 0.09. We first obtained cell-type-specific model parameters. Then, we simulated data under different total the cells from 1000 to 6000. Lastly, we performed DE analysis and calculated stratified targeted power for Form I and II DE. In Figure S13, the heatmap in each panel is a simulation case with specific total cells. The first row (A) is the power evaluation result for Form I DE, and the second row (B) is for Form II DE. In most pairwise comparisons, the powers are low even with many cells. This is because of the large within group variance (Figure S14), which makes the DE detection much challenging.

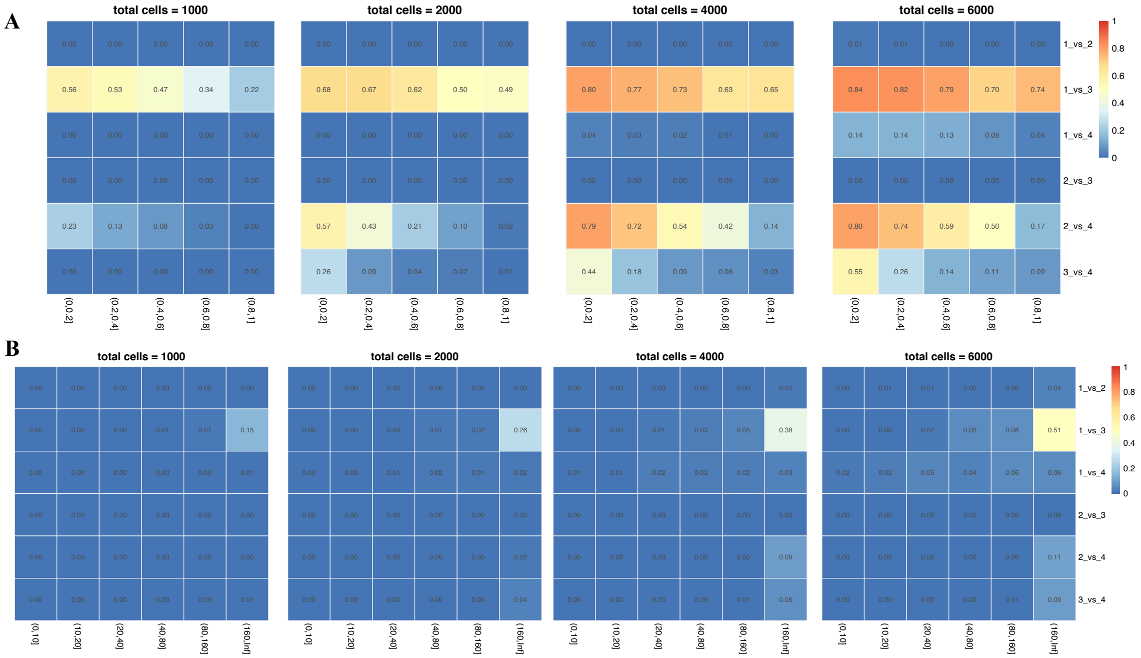


Figure B.2: Another test on scenario about the cross cell types comparisons under the same condition. It is performed on a real scRNA-seq dataset that contains 5 glioblastoma patients. We used MGH31 patient as the template for this case study.

B.3 Multiple cell types in another Glioblastoma (GSE84465)

We also tested POWSC on another GBM dataset from GEO with accession ID of GSE84465, to investigate how the powers improve in a different biological context. We found that the power (Figure S15A-B) for Form I and II DE have more similarity than difference compared with the case of GSE67835. However, the results are distinct from GSE57872 even if they are from the same biological system. It is indicated power analysis is case sensitive, and many unique factors can affect the power evaluations. For instance, the within group variations as well as between group variations can play essential roles for power. Specifically, if the within group variances are high, it will more difficult to detect DE genes.

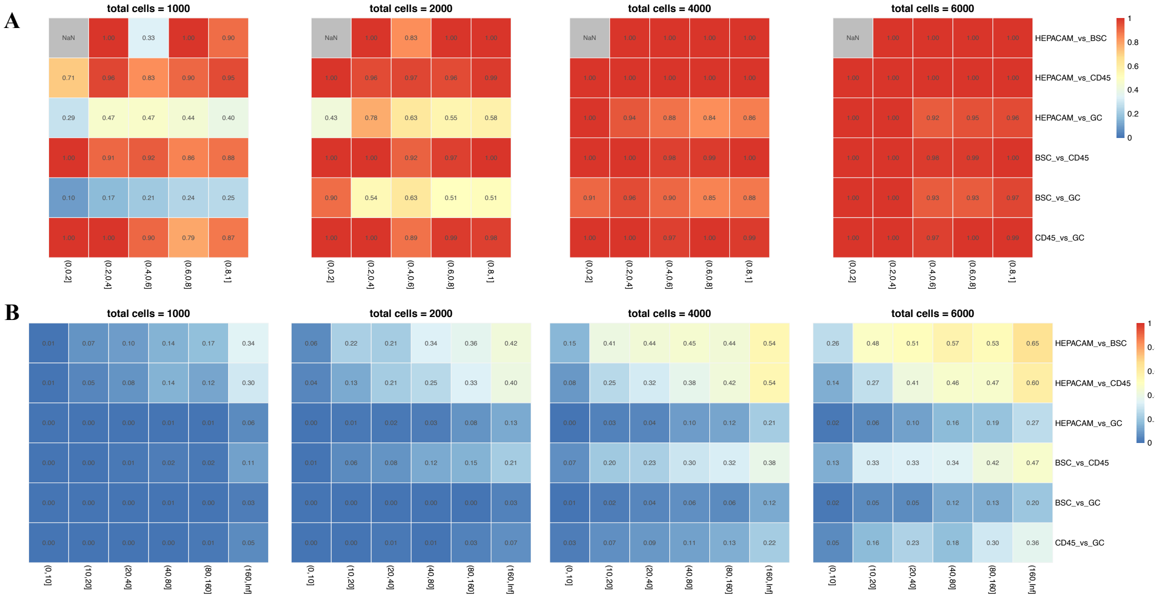


Figure B.3: One more test on the second scenario about the multiple cell types under the same condition. We used a real scRNA-seq dataset that contains 4 glioblastoma patients, and we utilize one patient with ID BT_S1 as template for this case study.

B.4 Test on 10X platform

10X community platform has become an popular tools for researchers to conduct clustering analysis; however, it is not known for DE analysis. Here, we test the ability of detecting DE genes by using POWSC. POWSC performs the simulation process based on a real 10X dataset about Peripheral blood mononuclear cells (PBMCs) https://support.10xgenomics.com/single- cell-gene-expression/datasets/2.1.0/pbmc4k. This dataset includes 4,340 cells with average sequencing reads 3 K. We found that the powers estimated from both forms of DE are not stable as from deeper sequencing depth cases 0.5 Million. We increased the sample size from 100 to 2000, and drew the distributions of the recovered (RD) and true DE (TD) genes for each case. Figure S17A-E are for Form I DE and Figure S18A-E are for Form II DE. The power evaluations are summarized in Figure S17F and Figure S18F. The marginal power Figure S19 shows the power for Form II DE has a slow growth from 0.02 to 0.46, but the power for Form I DE saturated after including 1000 cells.

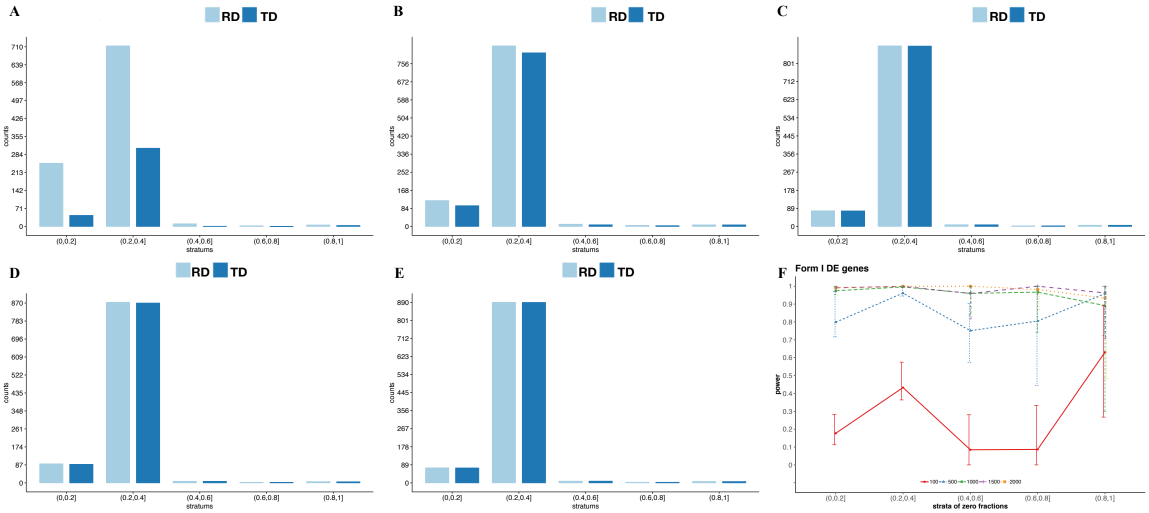


Figure B.4: Form II power evaluation for 10x data: A-E show the count distributions of the simulated DE genes and discovered DE genes for the Form I. They are averaged by the 50 runs. F shows the stratified targeted powers for Form I DE.

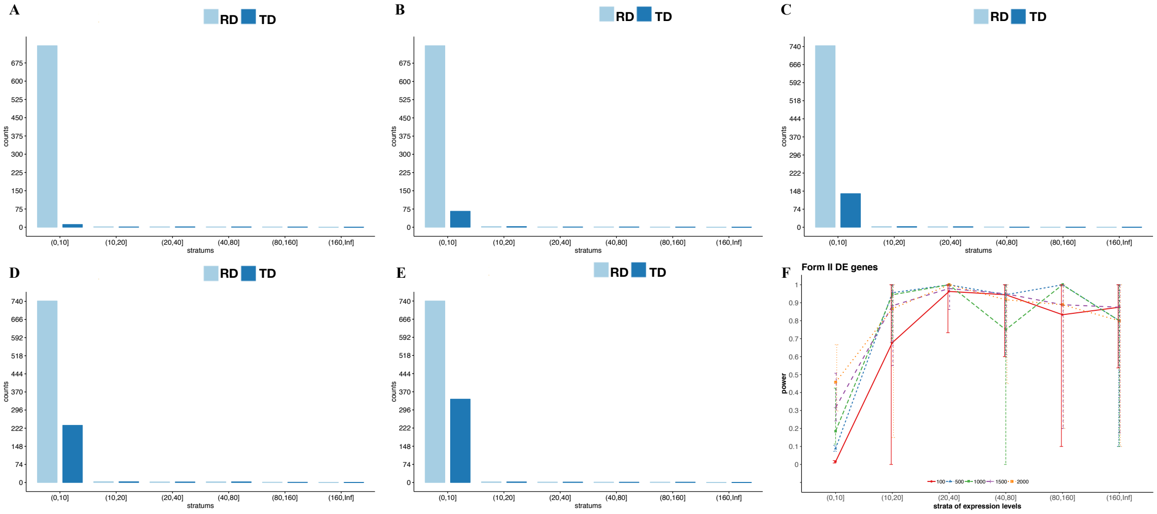


Figure B.5: Form II power evaluation for 10x data: A-E show the count distributions of the simulated DE genes and discovered DE genes for the Form II, averaged from 50 runs. F shows the stratified targeted powers for Form II DE. It shows that most of the DE genes are located in the interval of (0, 10] because of low sequencing depth. Thus, successfully detecting the DE genes at this interval will contribute to the marginal power evaluation.

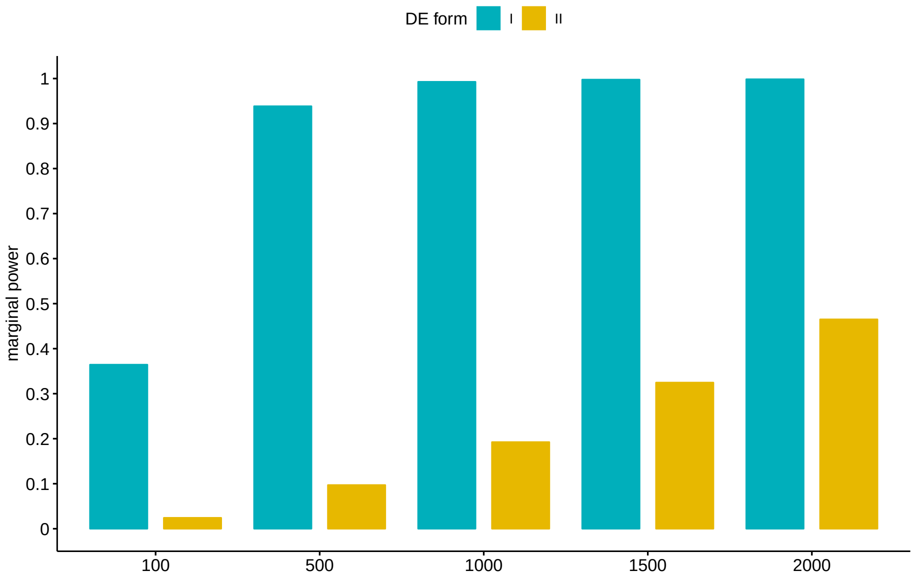


Figure B.6: The marginal power for Form I and II DE, it shows that with the increase of the total cell numbers the marginal power for Form I DE will become saturated, however the marginal power for Form II DE increases slowly from 0.02 to 0.46 mainly because of the loss of the detections at the stratum (0, 10].

Appendix C

Appendix for Chapter 4

Label	Cancer types	# of tumor samples	# of normal samples	# total samples
BLCA	Bladder Urothelial Carcinoma	408	19	427
BRCA	Breast invasive carcinoma	1,100	112	1,212
COAD	Colon adenocarcinoma	287	41	328
HNSC	Head and Neck squamous cell carcinoma	522	44	566
KICH	Kidney Chromophobe	66	25	91
KIRC	Kidney renal clear cell carcinoma	534	72	606
KIRP	Kidney renal papillary cell carcinoma	291	32	323
LIHC	Liver hepatocellular carcinoma	373	50	423
LUAD	Lung adenocarcinoma	517	59	576
LUSC	Lung squamous cell carcinoma	501	51	552
PRAD	Prostate adenocarcinoma	498	52	550
STAD	Stomach adenocarcinoma	415	35	450
THCA	Thyroid carcinoma	509	59	568
UCEC	Uterine Corpus Endometrial Carcinoma	177	24	201

Table C.1: Summary of the 14 Cancer types used in this study

Category	Sub-category or Annotations		# of Gene Sets	Average Size
C2	CGP: chemical and genetic perturbations		3433	110.52
	CP: Canonical pathways	BIOCARTA	217	20.88
		KEGG	186	69.22
		REACTOME	674	55.79
GO	BP: GO biological process		4436	114.11
	CC: GO cellular component		580	151.35
	MF: GO molecular function		901	106.11
Hallmark	Hallmark Gene Sets		50	146.48

Table C.2: Summary of pathways and gene sets used in this study

Pathway	P-value
DELYS_THYROID_CANCER_DN	7.65E-66
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_LOBULAR_NORMAL_DN	1.38E-65
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_DUCTAL_NORMAL_DN	1.25E-63
VECCHI_GASTRIC_CANCER_EARLY_DN	7.03E-63
LIU_PROSTATE_CANCER_DN	7.24E-62
SCHAEFFER_PROSTATE_DEVELOPMENT_AND_CANCER_BOX5_UP	6.53E-61
TOMLINS_PROSTATE_CANCER_DN	7.23E-61
SABATES_COLORECTAL_ADENOMA_DN	1.05E-60
YAO_HOXA10_TARGETS_VIA_PROGESTERONE_UP	1.38E-60
YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_0	1.69E-60

Table C.3: Top pathways in BRCA differentiated between tumor and normal in the iES of iPath.

Pathway	P-value	Ranking	Percentile
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_LOBULAR_NORMAL_DN	0.0000	2	0.04
MIKKELSEN_ES_ICP_WITH_H3K4ME3_AND_H3K27ME3	0.0000	325	6.87
FRASOR_TAMOXIFEN_RESPONSE_UP	0.0018	3339	70.61
MEDINA_SMARCA4_TARGETS	0.0018	3886	82.17
NIELSEN_LEIOMYOSARCOMA_DN	0.0018	148	3.13
NADERI_BREAST_CANCER_PROGNOSIS_DN	0.0019	285	6.03
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_DUCTAL_NORMAL_DN	0.0021	3	0.06
MATZUK_CENTRAL_FOR_FEMALE_FERTILITY	0.0026	107	2.26
REACTOME_INTRINSIC_PATHWAY	0.0027	1200	25.38
SCHAEFFER_PROSTATE_DEVELOPMENT_48HR_UP	0.0045	538	11.38

Table C.4: Top pathways in BRCA differentiated between tumor and normal in the ES of GSEA

C.1 The comparison between iPath and Human Pathology Atlas (HPA)on KIRC cancer type

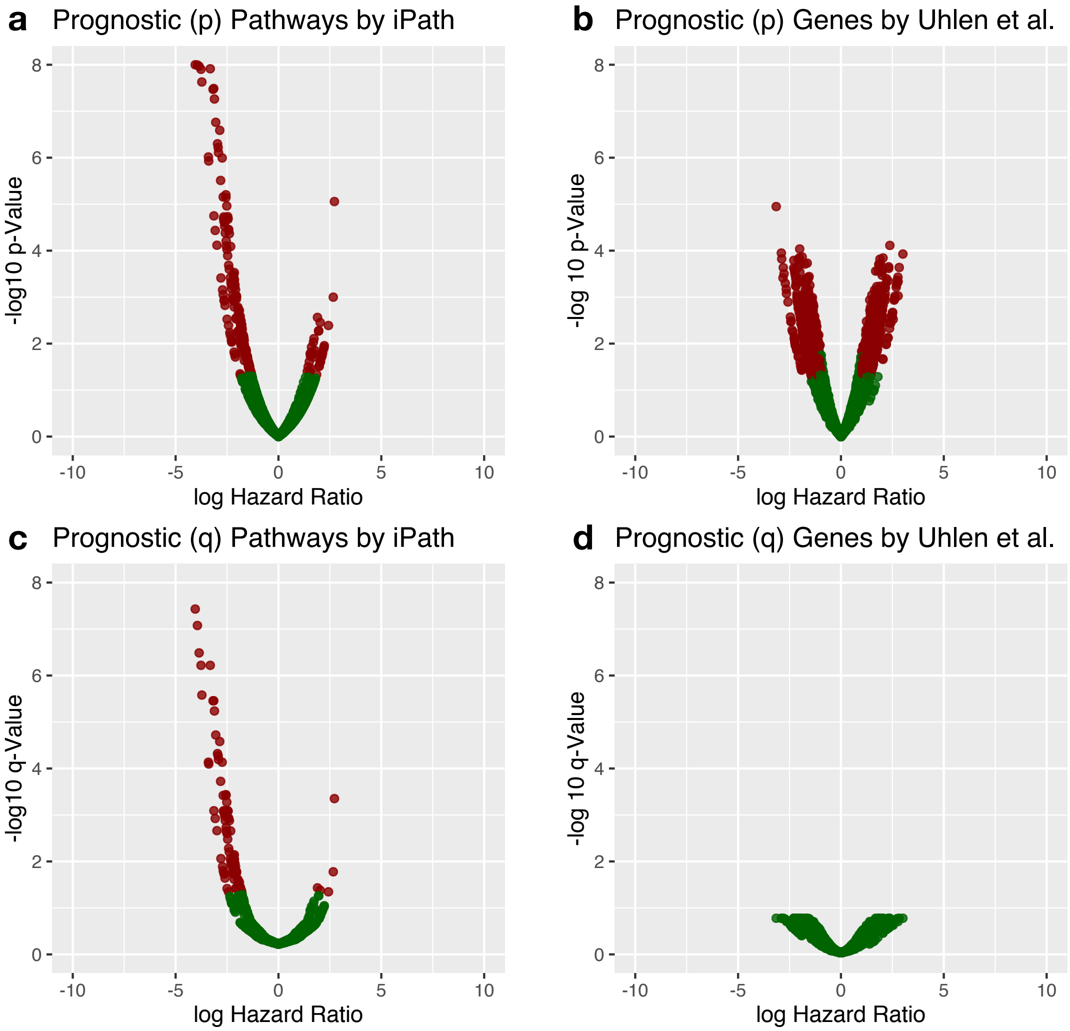


Figure C.1: The comparison between iPath and Human Pathology Atlas (HPA)on KIRC cancer type. We used p-value and q-value equal to 0.05 as the threshold to determine whether a pathway or gene is prognostic. a and b illustrate the prognostics biomarkers from both approaches when using p-value = 0.05. c and d are biomarkers from both approaches when using q-value = 0.05.

C.2 Selected prognostic C2 and GO pathways in BRCA

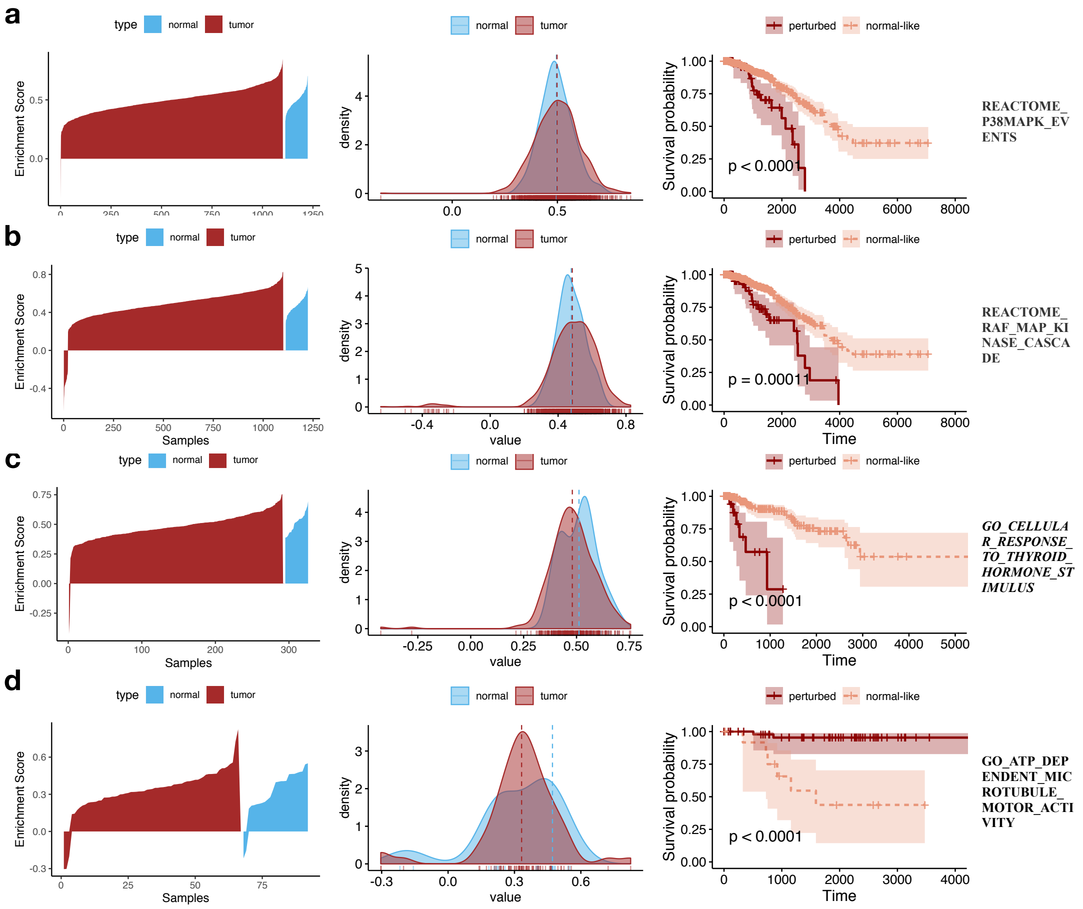


Figure C.2: Two prognostic biomarker pathways from the C2 category of MSigDB in BRCA, and two significant GO terms in kidney cancer type.

C.3 Test iPath on negative-control gene sets

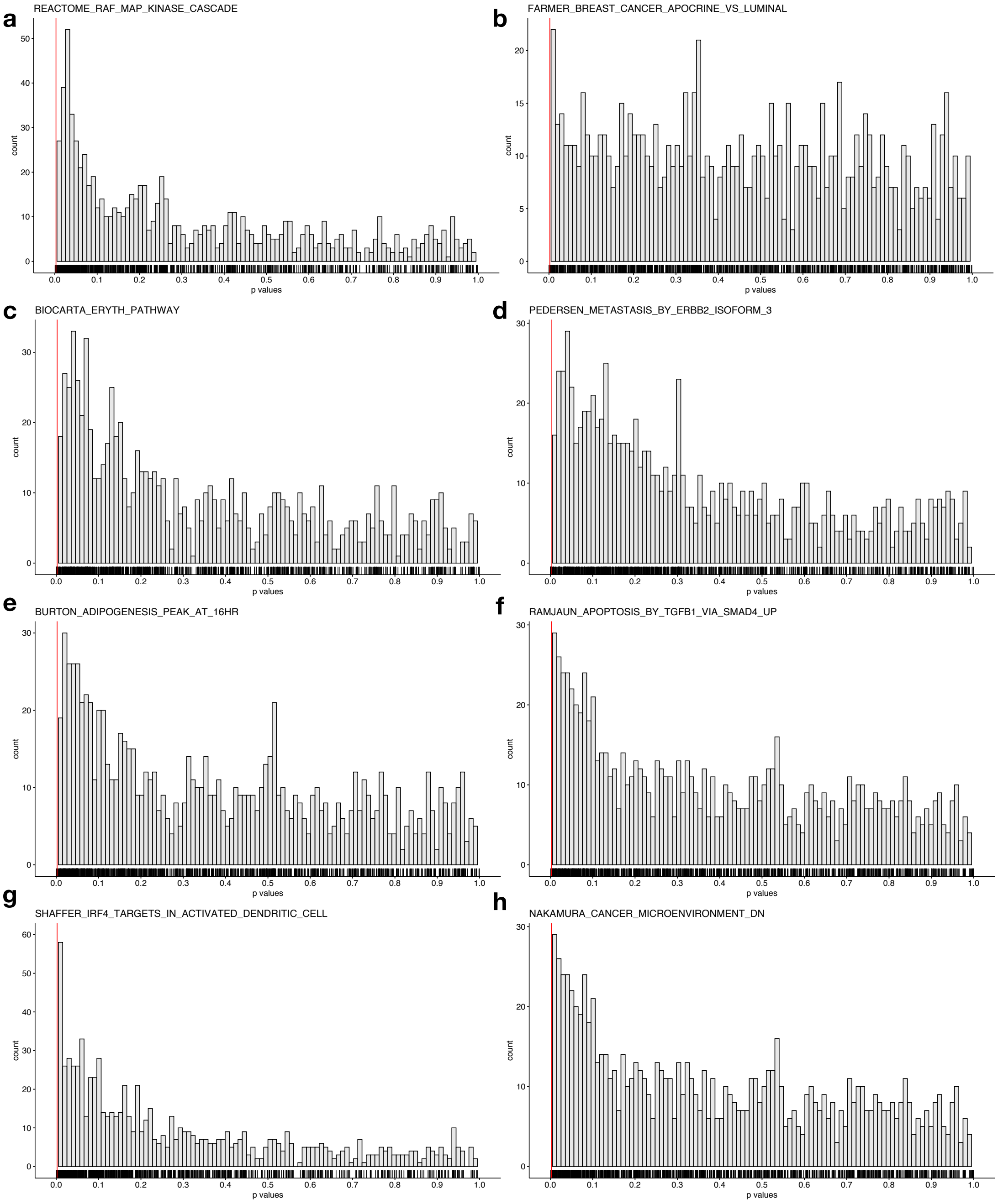


Figure C.3: Test iPath on randomly generated gene sets. For each of the biomarkers from BRCA (a-d) KIRP (e-f), and KIRC (g-h), we randomly simulate 1000 gene sets with the same sizes, and followed by iPath pipeline. The reported p values are shown in the histograms. The red vertical lines are the p values from the original gene sets which demonstrate the most significant level.

Bibliography

[1] Method of the Year 2013. *Nature Methods*, 11(1):1–1, January 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2801. URL `https://www.nature.com/articles/nmeth.2801`. Number: 1 Publisher: Nature Publishing Group.

[2] Hamed Al-Hussaini, Deepa Subramanyam, Michael Reedijk, and Srikala S. Sridhar. Notch signaling pathway as a therapeutic target in breast cancer. *Molecular Cancer Therapeutics*, 10(1):9–15, January 2011. ISSN 1538-8514. doi: 10.1158/1535-7163.MCT-10-0677.

[3] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[4] Tallulah S. Andrews and Martin Hemberg. False signals induced by single-cell imputation. *F1000Research*, 7, March 2019. ISSN 2046-1402. doi: 10.12688/f1000research.16613.2. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6415334/`.

[5] Richa Arya and Kristin White. Cell death in development: Signaling pathways and core mechanisms. *Seminars in Cell & Developmental Biology*, 39:12–19, March 2015. ISSN 1096-3634. doi: 10.1016/j.semcdb.2015.02.001.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill,

L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556.

[7] Eric J. Askeland, Vincent A. Chehval, Ryan W. Askeland, Placede G. Fosso, Zaina Sangale, Nafei Xu, Saradha Rajamani, Steven Stone, and James A. Brown. Cell cycle progression score predicts metastatic progression of clear cell renal cell carcinoma after resection. *Cancer Biomarkers: Section A of Disease Markers*, 15(6):861–867, 2015. ISSN 1875-8592. doi: 10.3233/CBM-150530.

[8] Elham Azizi, Ambrose J. Carr, George Plitas, Andrew E. Cornish, Catherine Konopacki, Sandhya Prabhakaran, Juozas Nainys, Kenmin Wu, Vaidotas Kiseliovas, Manu Setty, Kristy Choi, Rachel M. Fromme, Phuong Dao, Peter T. McKenney, Ruby C. Wasti, Krishna Kadaveru, Linas Mazutis, Alexander Y. Rudensky, and Dana Pe'er. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, 174(5):1293–1308.e36, August 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.05.060. URL `https://www.cell.com/cell/abstract/S0092-8674(18)30723-2`. Publisher: Elsevier.

[9] Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, Ron M Stewart, Michael Newton, and Christina Kendziorski. Scnorm: robust normalization of single-cell rna-seq data. *Nature methods*, 14(6):584, 2017.

[10] David A. Barbie, Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M. Chan, Martin L. Sos, Kathrin Michel, Craig Mermel, Serena J. Silver, Barbara A. Weir, Jan H. Reiling, Qing Sheng, Piyush B. Gupta, Raymond C. Wadlow, Hanh Le, Sebastian Hoersch,

Ben S. Wittner, Sridhar Ramaswamy, David M. Livingston, David M. Sabatini, Matthew Meyerson, Roman K. Thomas, Eric S. Lander, Jill P. Mesirov, David E. Root, D. Gary Gilliland, Tyler Jacks, and William C. Hahn. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112, November 2009. ISSN 1476-4687. doi: 10.1038/nature08460.

[11] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.

[12] Charles H Bennett, Péter Gács, Ming Li, Paul MB Vitányi, and Wojciech H Zurek. Information distance. *IEEE Transactions on information theory*, 44(4): 1407–1423, 1998.

[13] Inma M. Berenjeno, Roberto Piñeiro, Sandra D. Castillo, Wayne Pearce, Nicholas McGranahan, Sally M. Dewhurst, Valerie Meniel, Nicolai J. Birkbak, Evelyn Lau, Laurent Sansregret, Daniele Morelli, Nnennaya Kanu, Shankar Srinivas, Mariona Graupera, Victoria E. R. Parker, Karen G. Montgomery, Larissa S. Moniz, Cheryl L. Scudamore, Wayne A. Phillips, Robert K. Semple, Alan Clarke, Charles Swanton, and Bart Vanhaesebroeck. Oncogenic PIK3CA induces centrosome amplification and tolerance to genome doubling. *Nature Communications*, 8(1):1773, November 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-02002-4.

[14] Shuhui Bian, Yu Hou, Xin Zhou, Xianlong Li, Jun Yong, Yicheng Wang, Wendong Wang, Jia Yan, Boqiang Hu, Hongshan Guo, Jilian Wang, Shuai Gao, Yunuo Mao, Ji Dong, Ping Zhu, Dianrong Xiu, Liying Yan, Lu Wen, Jie Qiao, Fuchou Tang, and Wei Fu. Single-cell multiomics sequencing and

analyses of human colorectal cancer. *Science*, 362(6418):1060–1063, November 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aao3791. URL https://science.sciencemag.org/content/362/6418/1060. Publisher: American Association for the Advancement of Science Section: Report.

[15] Tom Bodenheimer, Mahantesh Halappanavar, Stuart Jefferys, Ryan Gibson, Siyao Liu, Peter J. Mucha, Natalie Stanley, Joel S. Parker, and Sara R. Selitsky. FastPG: Fast clustering of millions of single cells. *bioRxiv*, page 2020.06.19.159749, June 2020. doi: 10.1101/2020.06.19.159749. URL https://www.biorxiv.org/content/10.1101/2020.06.19.159749v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[16] Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, July 2015. ISSN 1476-4687. doi: 10.1038/nature14590. URL https://www.nature.com/articles/nature14590. Number: 7561 Publisher: Nature Publishing Group.

[17] Maria E. Cabanillas, Mabel Ryder, and Camilo Jimenez. Targeted Therapy for Advanced Thyroid Cancer: Kinase Inhibitors and Beyond. *Endocrine Reviews*, 40(6):1573–1604, December 2019. ISSN 1945-7189. doi: 10.1210/er.2019-00007.

[18] Laura Cantini, Laurence Calzone, Loredana Martignetti, Mattias Rydenfelt, Nils Blüthgen, Emmanuel Barillot, and Andrei Zinovyev. Classification of gene signatures for their information value and functional redundancy. *NPJ systems biology and applications*, 4:2, 2018. ISSN 2056-7189. doi: 10.1038/s41540-017-0038-8.

[19] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim,

Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[20] Scott L. Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W. Laird, Robert C. Onofrio, Wendy Winckler, Barbara A. Weir, Rameen Beroukhim, David Pellman, Douglas A. Levine, Eric S. Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, May 2012. ISSN 1546-1696. doi: 10.1038/nbt.2203.

[21] Gayathri Chandrasekaran, Péter Tátrai, and Fanni Gergely. Hitting the brakes: targeting microtubule motors in cancer. *British Journal of Cancer*, 113(5):693–698, September 2015. ISSN 1532-1827. doi: 10.1038/bjc.2015.264.

[22] Liang Chen, Lushun Yuan, Kaiyu Qian, Guofeng Qian, Yuan Zhu, Chin-Lee Wu, Han C. Dan, Yu Xiao, and Xinghuan Wang. Identification of Biomarkers Associated With Pathological Stage and Prognosis of Clear Cell Renal Cell Carcinoma by Co-expression Network Analysis. *Frontiers in Physiology*, 9:399, 2018. ISSN 1664-042X. doi: 10.3389/fphys.2018.00399.

[23] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):173, 2016.

[24] Stephen J. Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C. Marioni, Oliver Stegle, and Wolf Reik. scNMT-seq enables joint profiling of chromatin accessibility DNA methy-

lation and transcription in single cells. *Nature Communications*, 9(1):781, February 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03149-4. URL https://www.nature.com/articles/s41467-018-03149-4. Number: 1 Publisher: Nature Publishing Group.

[25] Michael B Cole, Davide Risso, Allon Wagner, David DeTomaso, John Ngai, Elizabeth Purdom, Sandrine Dudoit, and Nir Yosef. Performance assessment and selection of normalization procedures for single-cell rna-seq. *Cell systems*, 8(4):315–328, 2019.

[26] Adeline Crinier, Pierre Milpied, Bertrand Escalière, Christelle Piperoglou, Justine Galluso, Anaïs Balsamo, Lionel Spinelli, Inaki Cervera-Marzal, Mikaël Ebbo, Mathilde Girard-Madoux, Sébastien Jaeger, Emilie Bollon, Sami Hamed, Jean Hardwigsen, Sophie Ugolini, Frédéric Vély, Emilie Narni-Mancinelli, and Eric Vivier. High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. *Immunity*, 49 (5):971–986.e5, November 2018. ISSN 1074-7613. doi: 10.1016/j.immuni.2018.09.009. URL https://www.cell.com/immunity/abstract/S1074-7613(18)30424-2. Publisher: Elsevier.

[27] Carmen Criscitiello, Angela Esposito, Sabino De Placido, and Giuseppe Curigliano. Targeting fibroblast growth factor receptor pathway in breast cancer. *Current Opinion in Oncology*, 27(6):452–456, November 2015. ISSN 1531-703X. doi: 10.1097/CCO.0000000000000224.

[28] Helena L Crowell, Charlotte Soneson, Pierre-Luc Germain, Daniela Calini, Ludovic Collin, Catarina Raposo, Dheeraj Malhotra, and Mark D Robinson. On the discovery of population-specific state transitions from multi-sample multi-condition single-cell rna sequencing data. *BioRxiv*, page 713412, 2019.

[29] Hien Dang, Yotsawat Pomyen, Sean P. Martin, Dana A. Dominguez, Sun Young Yim, Ju-Seog Lee, Anuradha Budhu, Ashesh P. Shah, Adam S. Bodzin, and Xin Wei Wang. NELFE-Dependent MYC Signature Identifies a Unique Cancer Subtype in Hepatocellular Carcinoma. *Scientific Reports*, 9(1):3369, March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-39727-9.

[30] Spyros Darmanis, Steven A. Sloan, Derek Croote, Marco Mignardi, Sophia Chernikova, Peyman Samghababi, Ye Zhang, Norma Neff, Mark Kowarsky, Christine Caneda, Gordon Li, Steven D. Chang, Ian David Connolly, Yingmei Li, Ben A. Barres, Melanie Hayden Gephart, and Stephen R. Quake. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Reports*, 21(5):1399–1410, October 2017. ISSN 2211-1247. doi: 10.1016/j.celrep.2017.10.030. URL `https://www.cell.com/cell-reports/abstract/S2211-1247(17)31462-6`. Publisher: Elsevier.

[31] Jurrian K de Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, and Frank CP Holstege. Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic acids research*, 47(16):e95–e95, 2019.

[32] Patrick W. B. Derksen, Xiaoling Liu, Francis Saridin, Hanneke van der Gulden, John Zevenhoven, Bastiaan Evers, Judy R. van Beijnum, Arjan W. Griffioen, Jacqueline Vink, Paul Krimpenfort, Johannes L. Peterse, Robert D. Cardiff, Anton Berns, and Jos Jonkers. Somatic inactivation of E-cadherin and p53 in mice leads to metastatic lobular mammary carcinoma through induction of anoikis resistance and angiogenesis. *Cancer Cell*, 10(5):437–449, November 2006. ISSN 1535-6108. doi: 10.1016/j.ccr.2006.09.013.

[33] Kevin Dobbin and Richard Simon. Sample size determination in microarray

experiments for class comparison and prognostic classification. *Biostatistics*, 6 (1):27–38, 2005.

[34] S. M. Donnelly, E. Paplomata, B. M. Peake, E. Sanabria, Z. Chen, and R. Nahta. P38 MAPK contributes to resistance and invasiveness of HER2- overexpressing breast cancer. *Current Medicinal Chemistry*, 21(4):501–510, 2014. ISSN 1875-533X. doi: 10.2174/0929867320666131119155023.

[35] Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16):6388–6393, April 2013. ISSN 1091-6490. doi: 10.1073/pnas.1219651110.

[36] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

[37] Xiaoying Fan, Ji Dong, Suijuan Zhong, Yuan Wei, Qian Wu, Liying Yan, Jun Yong, Le Sun, Xiaoye Wang, Yangyu Zhao, et al. Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell rna-seq analysis. *Cell research*, 28(7):730–745, 2018.

[38] Jingya Fang, Cong Pian, Mingmin Xu, Lingpeng Kong, Zutan Li, Jinwen Ji, Yuanyuan Chen, and Liangyun Zhang. Revealing Prognosis-Related Pathways at the Individual Level by a Comprehensive Analysis of Different Cancer Transcription Data. *Genes*, 11(11), October 2020. ISSN 2073-4425. doi: 10.3390/genes11111281.

[39] Zhide Fang and Xiangqin Cui. Design and validation issues in RNA-seq experiments. *Briefings in bioinformatics*, 12(3):280–287, 2011.

[40] Pierre Farmer, Herve Bonnefoi, Veronique Becette, Michele Tubiana-Hulin, Pierre Fumoleau, Denis Larsimont, Gaetan Macgrogan, Jonas Bergh, David Cameron, Darlene Goldstein, Stephan Duss, Anne-Laure Nicoulaz, Cathrin Brisken, Maryse Fiche, Mauro Delorenzi, and Richard Iggo. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*, 24(29): 4660–4671, July 2005. ISSN 0950-9232. doi: 10.1038/sj.onc.1208561.

[41] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology*, 16(1):278, 2015.

[42] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383): 553–569, 1983.

[43] Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, 7, December 2018. ISSN 2046-1402. doi: 10.12688/f1000research.15809.2. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6124389/`.

[44] Milena Gasco, Shukri Shami, and Tim Crook. The p53 pathway in breast cancer. *Breast cancer research: BCR*, 4(2):70–76, 2002. ISSN 1465-5411. doi: 10.1186/bcr426.

[45] Dominic Grün. Revealing dynamics of gene expression variability in cell state space. *Nature methods*, 17(1):45–49, 2020.

[46] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637, 2014.

[47] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568): 251, 2015.

[48] Yanqin Gu, Linfeng Lu, Lingfeng Wu, Hao Chen, Wei Zhu, and Yi He. Identification of prognostic genes in kidney renal clear cell carcinoma by RNA-seq data analysis. *Molecular Medicine Reports*, 15(4):1661–1667, April 2017. ISSN 1791-3004. doi: 10.3892/mmr.2017.6194.

[49] Gunes Gundem and Nuria Lopez-Bigas. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. *Genome Medicine*, 4(3):28, March 2012. ISSN 1756-994X. doi: 10.1186/gm327.

[50] Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11):e1004575, 2015.

[51] David A. Gutman, Jake Cobb, Dhananjaya Somanna, Yuna Park, Fusheng Wang, Tahsin Kurc, Joel H. Saltz, Daniel J. Brat, and Lee A. D. Cooper. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *Journal of the American Medical Informatics Association: JAMIA*, 20(6):1091–1098, December 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2012-001469.

[52] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):1–15, 2019.

[53] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011. ISSN 1097-4172. doi: 10.1016/j.cell.2011.02.013.

[54] Steven N Hart, Terry M Therneau, Yuji Zhang, Gregory A Poland, and Jean-Pierre Kocher. Calculating sample size estimates for RNA sequencing data. *Journal of computational biology*, 20(12):970–978, 2013.

[55] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *Advances in neural information processing systems*, 18:507–514, 2005.

[56] Zhongshi He, Min Sun, Yuan Ke, Rongjie Lin, Youde Xiao, Shuliang Zhou, Hong Zhao, Yan Wang, Fuxiang Zhou, and Yunfeng Zhou. Identifying biomarkers of papillary renal cell carcinoma associated with pathological stage by weighted gene co-expression network analysis. *Oncotarget*, 8(17):27904–27914, April 2017. ISSN 1949-2553. doi: 10.18632/oncotarget.15842.

[57] Stephanie C. Hicks, Ruoxi Liu, Yuwei Ni, Elizabeth Purdom, and Davide Risso. mbkmeans: Fast clustering for single cell data using mini-batch k-means. *PLOS Computational Biology*, 17(1):e1008625, January 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008625. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008625`. Publisher: Public Library of Science.

[58] Yu-Jui Ho, Naishitha Anaparthy, David Molik, Grinu Mathew, Toby Aicher, Ami Patel, James Hicks, and Molly Gale Hammell. Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome research*, 28(9):1353–1363, 2018.

[59] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C Hicks. A systematic

evaluation of single-cell rna-sequencing imputation methods. *Genome biology*, 21(1):1–30, 2020.

[60] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research*, 26(3):304, 2016.

[61] Jian Hu, Xiangjie Li, Gang Hu, Yafei Lyu, Katalin Susztak, and Mingyao Li. Iterative transfer learning with neural network for clustering and cell type classification in single-cell rna-seq analysis. *Nature Machine Intelligence*, 2(10): 607–618, 2020.

[62] Ming-Wen Hu, Dong Won Kim, Sheng Liu, Donald J Zack, Seth Blackshaw, and Jiang Qian. Panoview: An iterative clustering method for single-cell rna sequencing data. *PLoS computational biology*, 15(8):e1007040, 2019.

[63] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[64] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, 14:7, January 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-7.

[65] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research*, 21 (7):1160–1167, 2011.

[66] Diego Adhemar Jaitin, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van

Oudenaarden, and Ido Amit. Dissecting immune circuits by linking crispr-pooled screens with single-cell rna-seq. *Cell*, 167(7):1883–1896, 2016.

[67] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

[68] Yunlu Jia, Jichun Zhou, Xiao Luo, Miao Chen, Yongxia Chen, Ji Wang, Hanchu Xiong, Xiaogang Ying, Wenxian Hu, Wenhe Zhao, Wuguo Deng, and Linbo Wang. KLF4 overcomes tamoxifen resistance by suppressing MAPK signaling pathway and predicts good prognosis in breast cancer. *Cellular Signalling*, 42: 165–175, January 2018. ISSN 1873-3913. doi: 10.1016/j.cellsig.2017.09.025.

[69] Aashi Jindal, Prashant Gupta, Debarka Sengupta, et al. Discovery of rare cells from voluminous single cell expression data. *Nature communications*, 9(1):1–9, 2018.

[70] Soobok Joe and Hojung Nam. Prognostic factor analysis for breast cancer using gene expression profiles. *BMC medical informatics and decision making*, 16 Suppl 1:56, July 2016. ISSN 1472-6947. doi: 10.1186/s12911-016-0292-5.

[71] J. Johnson, B. Thijssen, U. McDermott, M. Garnett, L. F. A. Wessels, and R. Bernards. Targeting the RB-E2F pathway in breast cancer. *Oncogene*, 35 (37):4829–4835, September 2016. ISSN 1476-5594. doi: 10.1038/onc.2016.32.

[72] Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F. McMichael, Matthew A. Wyczalkowski, Mark D. M. Leiserson, Christopher A. Miller, John S. Welch, Matthew J. Walter, Michael C. Wendl, Timothy J. Ley, Richard K. Wilson, Benjamin J. Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, October 2013. ISSN 1476-4687. doi: 10.1038/nature12634.

[73] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7):740, 2014.

[74] Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002375.

[75] Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, and Pengyi Yang. Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 20(6):2316–2326, 2019.

[76] Taj D. King, Mark J. Suto, and Yonghe Li. The Wnt/β-catenin signaling pathway: a potential therapeutic target in the treatment of triple negative breast cancer. *Journal of Cellular Biochemistry*, 113(1):13–18, January 2012. ISSN 1097-4644. doi: 10.1002/jcb.23350.

[77] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods*, 14(5):483, 2017.

[78] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362, 2018.

[79] Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5):273–282, May 2019. ISSN 1471-0064. doi: 10.1038/s41576-018-0088-9. URL `https://www.nature.com/articles/s41576-018-0088-9`. Number: 5 Publisher: Nature Publishing Group.

[80] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner.

Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[81] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[82] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome biology*, 17(1):222, 2016.

[83] Eilon Krashin, Agnieszka Piekiełko-Witkowska, Martin Ellis, and Osnat Ashur-Fabian. Thyroid Hormones and Cancer: A Comprehensive Review of Preclinical and Clinical Studies. *Frontiers in Endocrinology*, 10:59, 2019. ISSN 1664-2392. doi: 10.3389/fendo.2019.00059.

[84] Monika Krzak, Yordan Raykov, Alexis Boukouvalas, Luisa Cutillo, and Claudia Angelini. Benchmark and parameter sensitivity analysis of single-cell rna sequencing clustering methods. *Frontiers in genetics*, 10:1253, 2019.

[85] Blue B. Lake, Rizi Ai, Gwendolyn E. Kaeser, Neeraj S. Salathia, Yun C. Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293):1586–1590, June 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaf1204. URL `https://science.sciencemag.org/content/352/6293/1586`. Publisher: American Association for the Advancement of Science Section: Reports.

[86] Mark D. M. Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, February 2015. ISSN 1546-1718. doi: 10.1038/ng.3168.

[87] Kevin G. Leong, Kyle Niessen, Iva Kulic, Afshin Raouf, Connie Eaves, Ingrid Pollet, and Aly Karsan. Jagged1-mediated Notch activation induces epithelial-to-mesenchymal transition through Slug-induced repression of E-cadherin. *The Journal of Experimental Medicine*, 204(12):2935–2948, November 2007. ISSN 1540-9538. doi: 10.1084/jem.20071082.

[88] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234 (5323):34–35, 1971.

[89] Chenwei Li, Baolin Liu, Boxi Kang, Zedao Liu, Yedan Liu, Changya Chen, Xianwen Ren, and Zemin Zhang. Scibet as a portable and fast single cell type identifier. *Nature communications*, 11(1):1–8, 2020.

[90] Chung-I Li, Pei-Fang Su, Yan Guo, and Yu Shyr. Sample size calculation for differential expression analysis of RNA-seq data under poisson distribution. *International journal of computational biology and drug design*, 6(4), 2013.

[91] Hongjie Li, Felix Horns, Bing Wu, Qijing Xie, Jiefu Li, Tongchao Li, David J. Luginbuhl, Stephen R. Quake, and Liqun Luo. Classifying Drosophila Olfactory Projection Neuron Subtypes by Single-Cell RNA Sequencing. *Cell*, 171(5):1206–1220.e22, November 2017. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.

2017.10.019. URL https://www.cell.com/cell/abstract/S0092-8674(17)31241-2. Publisher: Elsevier.

[92] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, September 2011. ISSN 1932-6157, 1941-7330. doi: 10.1214/11-AOAS466. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-5/issue-3/Measuring-reproducibility-of-high-throughput/10.1214/11-AOAS466.full. Publisher: Institute of Mathematical Statistics.

[93] Shun Li, Ying Song, Christine Quach, Hongrui Guo, Gyu-Beom Jang, Hadi Maazi, Shihui Zhao, Nathaniel A. Sands, Qingsong Liu, Gino K. In, David Peng, Weiming Yuan, Keigo Machida, Min Yu, Omid Akbari, Ashley Hagiya, Yongfei Yang, Vasu Punj, Liling Tang, and Chengyu Liang. Transcriptional regulation of autophagy-lysosomal function in BRAF-driven melanoma progression and chemoresistance. *Nature Communications*, 10(1):1693, April 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09634-8.

[94] Wei Vivian Li and Jingyi Jessica Li. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics*, 35(14):i41–i50, 2019.

[95] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, 27(12):1739–1740, June 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr260.

[96] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database (MSigDB)

hallmark gene set collection. *Cell Systems*, 1(6):417–425, December 2015. ISSN 2405-4712. doi: 10.1016/j.cels.2015.12.004.

[97] Wei-Jiun Lin, Huey-Miin Hsueh, and James J Chen. Power and sample size estimation in microarray studies. *Bmc Bioinformatics*, 11(1):48, 2010.

[98] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15 (12):1053–1058, 2018.

[99] Aaron TL Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):1–14, 2016.

[100] Chongyuan Luo, Christopher L. Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, Brian Bui, Terrence J. Sejnowski, Timothy T. Harkins, Eran A. Mukamel, M. Margarita Behrens, and Joseph R. Ecker. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604, August 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan3351. URL `https://science.sciencemag.org/content/357/6351/600`. Publisher: American Association for the Advancement of Science Section: Report.

[101] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. ISSN 1533-7928. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

[102] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki,

Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[103] Iñigo Martincorena and Peter J. Campbell. Somatic mutation in cancer and normal cells. *Science (New York, N.Y.)*, 349(6255):1483–1489, September 2015. ISSN 1095-9203. doi: 10.1126/science.aab4082.

[104] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature Methods*, 18(1):9–14, 2021.

[105] Todd M. Morgan, Rohit Mehra, Placede Tiemeny, J. Stuart Wolf, Shulin Wu, Zaina Sangale, Michael Brawer, Steven Stone, Chin-Lee Wu, and Adam S. Feldman. A Multigene Signature Based on Cell Cycle Proliferation Improves Prediction of Mortality Within 5 Yr of Radical Nephrectomy for Renal Cell Carcinoma. *European Urology*, 73(5):763–769, May 2018. ISSN 1873-7560. doi: 10.1016/j.eururo.2017.12.002.

[106] Gayathri Nagaraj and Cynthia Ma. Revisiting the estrogen receptor pathway and its role in endocrine therapy for postmenopausal women with estrogen receptor-positive metastatic breast cancer. *Breast Cancer Research and Treatment*, 150(2):231–242, April 2015. ISSN 1573-7217. doi: 10.1007/s10549-015-3316-4.

[107] A Narayanan. A note on parameter estimation in the multivariate beta distribution. *Computers & Mathematics with Applications*, 24(10):11–17, 1992.

[108] Karlynn E Neu, Qingming Tang, Patrick C Wilson, and Aly A Khan. Single-cell genomics: approaches and utility in immunology. *Trends in immunology*, 38(2):140–149, 2017.

[109] Quy H. Nguyen, Nicholas Pervolarakis, Kerrigan Blake, Dennis Ma, Ryan Tevia Davis, Nathan James, Anh T. Phung, Elizabeth Willey, Raj Kumar, Eric

Jabart, Ian Driver, Jason Rock, Andrei Goga, Seema A. Khan, Devon A. Lawson, Zena Werb, and Kai Kessenbrock. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nature Communications*, 9(1):2028, May 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04334-1. URL `https://www.nature.com/articles/s41467-018-04334-1`. Number: 1 Publisher: Nature Publishing Group.

[110] Vasilis Ntranos, Lynn Yi, Páll Melsted, and Lior Pachter. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nature methods*, 16(2):163, 2019.

[111] Sumanta Kumar Pal, Robert A. Figlin, and Karen Reckamp. Targeted therapies for non-small cell lung cancer: an evolving landscape. *Molecular Cancer Therapeutics*, 9(7):1931–1944, July 2010. ISSN 1538-8514. doi: 10.1158/1535-7163.MCT-10-0239.

[112] Junya Peng, Bao-Fa Sun, Chuan-Yuan Chen, Jia-Yi Zhou, Yu-Sheng Chen, Hao Chen, Lulu Liu, Dan Huang, Jialin Jiang, Guan-Shen Cui, et al. Single-cell rna-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research*, 29(9):725–738, 2019.

[113] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

[114] Ren Qi, Anjun Ma, Qin Ma, and Quan Zou. Clustering and classification methods for single-cell rna-sequencing data. *Briefings in bioinformatics*, 21(4): 1196–1208, 2020.

[115] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E. Mason, Nicholas D. Socci, and Doron Betel. Com-

prehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95, 2013. ISSN 1474-760X. doi: 10.1186/gb-2013-14-9-r95.

[116] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The Human Cell Atlas. *eLife*, 6:e27041, December 2017. ISSN 2050-084X. doi: 10.7554/eLife.27041. URL `https://doi.org/10.7554/eLife.27041`. Publisher: eLife Sciences Publications, Ltd.

[117] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, Daniele Merico, and Gary D. Bader. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, 14(2):482–517, February 2019. ISSN 1750-2799. doi: 10.1038/s41596-018-0103-9.

[118] Jeffrey S. Ross, Kai Wang, Christine E. Sheehan, Ann B. Boguniewicz, Geoff Otto, Sean R. Downing, James Sun, Jie He, John A. Curran, Siraj Ali, Roman Yelensky, Doron Lipson, Gary Palmer, Vincent A. Miller, and Philip J. Stephens. Relapsed classic E-cadherin (CDH1)-mutated invasive lobular breast cancer shows a high frequency of HER2 (ERBB2) gene mutations. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 19(10):2668–2676, May 2013. ISSN 1557-3265. doi: 10.1158/1078-0432.CCR-13-0295.

[119] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[120] Anna-Leena Salmela and Marko J. Kallio. Mitosis as an anti-cancer drug target. *Chromosoma*, 122(5):431–449, October 2013. ISSN 1432-0886. doi: 10.1007/s00412-013-0419-8.

[121] Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K. Chatila, Augustin Luna, Konnor C. La, Sofia Dimitriadoy, David L. Liu, Havish S. Kantheti, Sadegh Saghafinia, Debyani Chakravarty, Foysal Daian, Qingsong Gao, Matthew H. Bailey, Wen-Wei Liang, Steven M. Foltz, Ilya Shmulevich, Li Ding, Zachary Heins, Angelica Ochoa, Benjamin Gross, Jianjiong Gao, Hongxin Zhang, Ritika Kundra, Cyriac Kandoth, Istemi Bahceci, Leonard Dervishi, Ugur Dogrusoz, Wanding Zhou, Hui Shen, Peter W. Laird, Gregory P. Way, Casey S. Greene, Han Liang, Yonghong Xiao, Chen Wang, Antonio Iavarone, Alice H. Berger, Trever G. Bivona, Alexander J. Lazar, Gary D. Hammer, Thomas Giordano, Lawrence N. Kwong, Grant McArthur, Chenfei Huang, Aaron D. Tward, Mitchell J. Frederick, Frank McCormick, Matthew Meyerson, Cancer Genome Atlas Research Network, Eliezer M. Van Allen, Andrew D.

Cherniack, Giovanni Ciriello, Chris Sander, and Nikolaus Schultz. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, 173(2):321–337.e10, April 2018. ISSN 1097-4172. doi: 10.1016/j.cell.2018.03.035.

[122] Hirak Sarkar, Avi Srivastava, and Rob Patro. Minnow: a principled framework for rapid simulation of dscRNA-seq data at the read level. *Bioinformatics*, 35 (14):i136–i144, 2019.

[123] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495, 2015.

[124] T. Schmelzle and M. N. Hall. TOR, a central controller of cell growth. *Cell*, 103(2):253–262, October 2000. ISSN 0092-8674. doi: 10.1016/s0092-8674(00) 00117-3.

[125] Michael Schubert, Bertram Klinger, Martina Klünemann, Anja Sieber, Florian Uhlitz, Sascha Sauer, Mathew J. Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications*, 9(1):20, January 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02391-6.

[126] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[127] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236, 2013.

[128] Kuanwei Sheng, Wenjian Cao, Yichi Niu, Qing Deng, and Chenghang Zong. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature Methods*, 14(3):267–270, March 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4145. URL `https://www.nature.com/articles/nmeth.4145`. Number: 3 Publisher: Nature Publishing Group.

[129] Robert R Sokal and F James Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40, 1962.

[130] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, 14:91, March 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-91.

[131] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods*, 15(4):255, 2018.

[132] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4380. URL `https://www.nature.com/articles/nmeth.4380`. Number: 9 Publisher: Nature Publishing Group.

[133] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

[134] Kenong Su, Zhijin Wu, and Hao Wu. Simulation, power evaluation and sample size recommendation for single-cell rna-seq. *Bioinformatics*, 36(19):4860–4868, 2020.

[135] Kenong Su, Tianwei Yu, and Hao Wu. Accurate feature selection improves single-cell rna-seq cell clustering. *Briefings in Bioinformatics*, 2021.

[136] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102.

[137] Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell rna-seq analysis. *Genome biology*, 20(1):1–21, 2019.

[138] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, pages 1–4, 2020.

[139] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, May 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1315. URL `https://www.nature.com/articles/nmeth.1315`. Number: 5 Publisher: Nature Publishing Group.

[140] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in RNA-seq: a matter of depth. *Genome research*, 21(12):2213–2223, 2011.

[141] Sarah Teichmann and Mirjana Efremova. Method of the year 2019: single-cell multimodal omics. *Nat. Methods*, 17(1), 2020.

[142] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single cell RNA-Seq based on a multinomial model. *bioRxiv*, page 574574, 2019.

[143] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21 (1):12, January 2020. ISSN 1474-760X. doi: 10.1186/s13059-019-1850-9. URL `https://doi.org/10.1186/s13059-019-1850-9`.

[144] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[145] Mathias Uhlen, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu, and Fredrik Ponten. A pathology atlas of the human cancer transcriptome. *Science (New York, N.Y.)*, 357(6352), August 2017. ISSN 1095-9203. doi: 10.1126/science.aan2507.

[146] Marc J. van de Vijver, Yudong D. He, Laura J. van't Veer, Hongyue Dai, Augustinus A. M. Hart, Dorien W. Voskuil, George J. Schreiber, Johannes L. Peterse, Chris Roberts, Matthew J. Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony van der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T. Rutgers, Stephen H. Friend, and René

Bernards. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347(25):1999–2009, December 2002. ISSN 1533-4406. doi: 10.1056/NEJMoa021967.

[147] Edwin Vans, Ashwini Patil, and Alok Sharma. Feats: Feature selection based clustering of single-cell rna-seq data. *bioRxiv*, 2020.

[148] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics (Oxford, England)*, 26(12):i237–245, June 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq182.

[149] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, 33(21):3486–3488, 2017.

[150] Marie-Claire Wagle, Daniel Kirouac, Christiaan Klijn, Bonnie Liu, Shilpi Mahajan, Melissa Junttila, John Moffat, Mark Merchant, Ling Huw, Matthew Wongchenko, Kwame Okrah, Shrividhya Srinivasan, Zineb Mounir, Teiko Sumiyoshi, Peter M. Haverty, Robert L. Yauch, Yibing Yan, Omar Kabbarah, Garret Hampton, Lukas Amler, Saroja Ramanujan, Mark R. Lackner, and Shih-Min A. Huang. A transcriptional MAPK Pathway Activity Score (MPAS) is a clinically relevant biomarker in multiple cancer types. *NPJ precision oncology*, 2(1):7, 2018. ISSN 2397-768X. doi: 10.1038/s41698-018-0051-4.

[151] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145, 2016.

[152] Shibiao Wan, Junil Kim, and Kyoung Jae Won. Sharp: hyperfast and accurate

processing of single-cell rna-seq data via ensemble random projection. *Genome research*, 30(2):205–213, 2020.

[153] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature methods*, 14(4):414–416, 2017.

[154] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences*, 115(28):E6437–E6446, 2018.

[155] Zhe Wang, Junming Hu, W Evan Johnson, and Joshua D Campbell. scruff: an r/bioconductor package for preprocessing single-cell rna-sequencing data. *BMC bioinformatics*, 20(1):1–9, 2019.

[156] Rebekka Wegmann, Marilisa Neri, Sven Schuierer, Bilada Bilican, Huyen Hartkopf, Florian Nigsch, Felipa Mapa, Annick Waldt, Rachel Cuttat, Max R. Salick, Joe Raymond, Ajamete Kaykas, Guglielmo Roma, and Caroline Gubser Keller. CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. *Genome Biology*, 20(1): 142, July 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1739-7. URL `https://doi.org/10.1186/s13059-019-1739-7`.

[157] Caimiao Wei, Jiangning Li, and Roger E Bumgarner. Sample size for detecting differentially expressed genes in microarray experiments. *BMC genomics*, 5(1): 87, 2004.

[158] Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, June

2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.05.006. URL https://www.sciencedirect.com/science/article/pii/S0092867419305045.

[159] Agnieszka K. Witkiewicz and Erik S. Knudsen. Retinoblastoma tumor suppressor pathway in breast cancer: prognosis, precision medicine, and therapeutic interventions. *Breast cancer research: BCR*, 16(3):207, May 2014. ISSN 1465-542X. doi: 10.1186/bcr3652.

[160] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

[161] Hao Wu, Chi Wang, and Zhijin Wu. PROPER: comprehensive power evaluation for differential expression using rna-seq. *Bioinformatics*, 31(2):233–241, 2014.

[162] Zhijin Wu and Hao Wu. Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. *Genome Biology*, 21(1):123, May 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02027-x. URL https://doi.org/10.1186/s13059-020-02027-x.

[163] Zhijin Wu, Yi Zhang, Michael L Stitzel, and Hao Wu. Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics*, 34(19):3340–3348, 2018.

[164] Kaikun Xie, Yu Huang, Feng Zeng, Zehua Liu, and Ting Chen. scAIDE: clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types. *NAR Genomics and Bioinformatics*, 2(lqaa082), December 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa082. URL https://doi.org/10.1093/nargab/lqaa082.

[165] Gang Xin, Ryan Zander, David M. Schauder, Yao Chen, Jason S. Weinstein, William R. Drobyski, Vera Tarakanova, Joseph Craft, and Weiguo Cui. Single-cell RNA sequencing unveils an IL-10-producing helper subset that sustains

humoral immunity during persistent infection. *Nature Communications*, 9(1): 5037, November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07492-4. URL `https://www.nature.com/articles/s41467-018-07492-4`. Number: 1 Publisher: Nature Publishing Group.

[166] Qinghua Xu, Jinying Chen, Shujuan Ni, Cong Tan, Midie Xu, Lei Dong, Lin Yuan, Qifeng Wang, and Xiang Du. Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 29(6):546–556, June 2016. ISSN 1530-0285. doi: 10.1038/modpathol.2016.60.

[167] Lu Yang, Jiancheng Liu, Qiang Lu, Arthur D Riggs, and Xiwei Wu. Saic: an iterative clustering approach for analysis of single cell rna-seq data. *BMC genomics*, 18(6):9–17, 2017.

[168] Yuchen Yang, Ruth Huh, Houston W Culpepper, Yuan Lin, Michael I Love, and Yun Li. SAFE-clustering: Single-cell Aggregated (from Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics*, 35(8):1269–1277, April 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty793. URL `https://doi.org/10.1093/bioinformatics/bty793`.

[169] Xiaomao Yin, Jianfeng Wang, and Jin Zhang. Identification of biomarkers of chromophobe renal cell carcinoma by weighted gene co-expression network analysis. *Cancer Cell International*, 18:206, 2018. ISSN 1475-2867. doi: 10.1186/s12935-018-0703-z.

[170] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[171] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg,

Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

[172] Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell RNA-Seq. *Cell systems*, 9(4):383–392, 2019.

[173] Lihua Zhang and Shihua Zhang. Comparison of Computational Methods for Imputing Single-Cell RNA-Sequencing Data. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(2):376–389, April 2020. ISSN 1557-9964. doi: 10.1109/TCBB.2018.2848633.

[174] Qiming Zhang, Yao He, Nan Luo, Shashank J. Patel, Yanjie Han, Ranran Gao, Madhura Modak, Sebastian Carotta, Christian Haslinger, David Kind, Gregory W. Peet, Guojie Zhong, Shuangjia Lu, Weihua Zhu, Yilei Mao, Mengmeng Xiao, Michael Bergmann, Xueda Hu, Sid P. Kerkar, Anne B. Vogt, Stefan Pflanz, Kang Liu, Jirun Peng, Xianwen Ren, and Zemin Zhang. Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell*, 179(4): 829–845.e20, October 2019. ISSN 1097-4172. doi: 10.1016/j.cell.2019.10.003.

[175] Wei Zhang and Hui Tu Liu. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Research*, 12(1):9–18, March 2002. ISSN 1001-0602. doi: 10.1038/sj.cr.7290105.

[176] Yiqun Zhang, Lixing Yang, Melanie Kucherlapati, Fengju Chen, Angela Hadjipanayis, Angeliki Pantazi, Christopher A. Bristow, Eunjung A. Lee, Harshad S. Mahadeshwar, Jiabin Tang, Jianhua Zhang, Sahil Seth, Semin Lee, Xiaojia Ren, Xingzhi Song, Huandong Sun, Jonathan Seidman, Lovelace J. Luquette, Ruibin Xi, Lynda Chin, Alexei Protopopov, Wei Li, Peter J. Park, Raju Kucherlapati, and Chad J. Creighton. A Pan-Cancer Compendium of Genes Deregulated by

Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Reports*, 24(2):515–527, July 2018. ISSN 2211-1247. doi: 10.1016/j.celrep.2018.06.025.

[177] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One*, 9(1):e78644, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0078644.

[178] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.

[179] Lingxue Zhu, Jing Lei, Lambertus Klei, Bernie Devlin, and Kathryn Roeder. Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471, 2019.

[180] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell RNA sequencing methods. *Molecular cell*, 65(4):631–643, 2017.

[181] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[182] Shuguang Zuo, Xinhong Zhang, and Liping Wang. A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. *Scientific Reports*, 9(1):2615, February 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-39273-4.