



THE UNIVERSITY OF
SYDNEY

COMPUTATIONAL METHODS FOR SINGLE CELL OMICS DATA ANALYSIS

LIJIA YU

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science
The University of Sydney

December 2024

*This thesis is dedicated to my loving parents
Ying Qian and Zhongde Yu for their
unconditional love and support.*

STATEMENT OF ORIGINALITY

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature:

Lijia Yu, December 2024

ABSTRACT

The rapid advancement of single cell technologies always brings new challenges in the development of computational methods to effectively analyse increasingly large-scale and high dimensional single cell datasets. In this work, we explore and address several issues in the analysis of single cell data, with a focus on better understanding of cell types.

To evaluate clustering methods on their ability to automatically estimate the number of cell types, we performed a systematic benchmarking study to evaluate 14 methods in four categories of different techniques for estimating the number of cell types in single cell RNA data. We uncover the performance differences between these methods and summarise these multi-faceted evaluations into a recommendation.

To improve the clustering of multimodal single cell omics data, we present SnapCCESS, an ensemble deep learning framework for generating multi-view integrated embeddings for consensus clustering of multimodal single cell omics data. We show that by using the snapshot ensemble learning technique, snapCCESS can achieve high performance in cell clustering and improve computational speed.

To jointly examine the effect of cell types or other potentially related factors of differential abundance signatures, we present iDAS, an ANOVA-based framework for separating gene signatures that are common to all individuals from those that are specific to a particular cell type or cell state. In the case study, we use iDAS to identify the cell type-specific gene signatures that respond to immunotherapy.

Overall, this thesis addresses several questions in the analysis of single cell data, focusing on the estimation and development of methods for cell type

identification using single cell RNA-seq and multi-omics data, and the use of cell type information for a better understanding of complex gene signatures.

PUBLICATIONS

Most of the work presented in this thesis has appeared in publication. These are listed below:

- **Yu, L.**, Cao, Y., Yang, J. and Yang, P., (2022). Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biology*, 23(1).
- **Yu, L.**, Liu, C., Yang, J. Y. H., and Yang, P. (2023). Ensemble deep learning of embeddings for clustering multimodal single-cell omics data. *Bioinformatics*, 39(6):btad382

Besides the studies mentioned above that I led during my candidacy, I have also collaborated with others and contributed to additional projects. Below, I list the publicly available ones.

- Cao, Y., **Yu, L.**, Torkel, M., Kim, S., Lin, Y., Yang, P., Speed, T. P., Ghazanfar, S., and Yang, J. Y. H. (2023). The current landscape and emerging challenges of benchmarking single-cell methods. *bioRxiv*.
- Kim, D, Chen, C., **Yu, L.**, Yang, J. Y. H., and Yang, Pengyi. (2024). CLUEY enables knowledge-guided clustering and cell type detection from single-cell omics data. *bioRxiv*.

AUTHORSHIP ATTRIBUTION STATEMENT

Chapter 2 of this thesis is published as Yu et al. (2022). I collected the data, conducted the experiment and interpreted the results and wrote the majority of the manuscript under the supervision of A/Prof. Pengyi Yang and Prof. Jean Yang, with input from Dr. Yue Cao.

Chapter 3 of this thesis is published as Yu et al. (2023). I collected the data, wrote the deep learning package, conducted the experiment, interpreted the results with input from Dr. Chunlei Liu. I wrote the majority of manuscript under the supervision of A/Prof. Pengyi Yang, Dr. Chunlei Liu and Prof. Jean Yang.

I acknowledge that the publisher's policy grants permission for the use of published material within this thesis.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Signature:

Lijia Yu, December 2024

As primary supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Signature:

Pengyi Yang, December 2024

As secondary supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Signature:

Jean Yang, December 2024

ACKNOWLEDGMENTS

Looking back on the four years of my PhD journey, I have been incredibly lucky to meet, study and work with so many amazing people. This journey would not have been possible without the help of many people to whom I would like to express my gratitude.

I would like to start by thanking my supervisors, Pengyi and Jean, for their unwavering support, guidance, and mentorship throughout my PhD journey. Pengyi gave me two opportunities to enrol and supported me throughout my studies. He guided me into the fields of single-cell and deep learning, sharing his knowledge and helping me grow. I am deeply thankful to Jean for everything she has done for me—sharing her experience, knowledge, ideas, and time. Without her, I would not be where I am today for so many reasons. From Jean, I have learned not only from her expertise but also how to manage and navigate the difficult tasks in my life.

I would also like to thank my friends and colleagues at the Sydney Precision Data Science Centre and the Children’s Medical Research Institute, for their feedback and engagement, particularly: Adam, Alex, Andy, Cabiria, Carissa, Chuhan, Chunlei, Daniel, Dario, Di, Ellis, Elijah, Farhan, Garth, Hani, Harry, Helen, Hao, Jackson, John, Katie, Manoj, Miranda, Mohammad, Qiuzhuang, Rajan, Sam, Shila, Tom, Xiangnan, Yingxin, Yue and Yunwei. Special thanks to Daniel, Alex and Yue for proofreading this thesis.

In addition, I gratefully acknowledge the funding received towards my PhD from Pengyi and the University of Sydney, which has made my PhD journey financially possible.

Last but not least, I would like to express my greatest gratitude to my parents for their love and support that encouraged me throughout my entire life.

CONTENTS

1	INTRODUCTION	1
1.1	Single-cell sequencing technologies	2
1.2	Computational analysis on single-cell omics data	4
1.2.1	Initial data analysis (IDA)	5
1.2.2	Downstream analysis	7
1.3	Challenges of single-cell analysis	9
1.3.1	Cell type clustering	9
1.3.2	Interpret gene signatures	11
1.3.3	Challenges addressed in this thesis	12
1.4	Thesis outline and contributions	14
2	A BENCHMARK STUDY OF ESTIMATING THE NUMBER OF CELL TYPES FROM SINGLE-CELL RNA-SEQUENCING DATA	17
2.1	Introduction	18
2.2	Methods	21
2.2.1	Sampling of Tabula Muris and Tabula Sapiens data	21
2.2.2	Proposed clustering stability-based approach for number of cell type estimation	23
2.2.3	Clustering methods that estimate the number of cell types	23
2.2.4	Cell clustering evaluation metrics	24
2.2.5	Assessment of the run time and peak memory usage	26
2.2.6	Performance summary criteria	26
2.3	Results	28
2.3.1	Benchmark framework for clustering algorithms on esti- mating the number of cell types	28
2.3.2	Overall performance of clustering algorithms on the num- ber of cell type estimation	30

2.3.3	Impact of number of cells on number of cell type estimation and clustering	32
2.3.4	Bias analysis of clustering algorithms on estimating the number of cell types	34
2.3.5	Impact of imbalance of cell ratios on number of cell type estimation and clustering	36
2.3.6	Cross-comparison on datasets with high cell numbers using Tabula Muris and Tabula Sapiens data	38
2.3.7	Running time and peak memory usage	40
2.4	Discussion	41
2.5	Conclusions	44
3	ENSEMBLE DEEP LEARNING OF EMBEDDINGS FOR CLUSTERING MULTIMODAL SINGLE-CELL OMICS DATA	46
3.1	Introduction	47
3.2	Methods	50
3.2.1	SnapCCESS framework for generating embeddings of multimodal single-cell data	50
3.2.2	Clustering algorithms for ensemble clustering of embeddings	52
3.2.3	Evaluation settings and datasets	53
3.2.4	Performance evaluation criteria	57
3.3	Results	58
3.3.1	SnapCCESS is an effective and efficient ensemble deep learning method for clustering multimodal single-cell omics data	58
3.3.2	Diagnostic analysis of SnapCCESS reveals its ability to improve major and minor cell type identification	62
3.3.3	Integrated embedding of multimodality generally leads to more precise cell clustering	63
3.3.4	SnapCCESS improves various clustering algorithms	64

3.3.5	SnapCCESS performs competitively to the state-of-the-art embedding generating methods for multimodal single-cell clustering	65
3.4	Discussion and conclusion	67
4	INTERPRETABLE DIFFERENTIAL ABUNDANCE SIGNATURE	70
4.1	Background and current approach	70
4.2	Data	72
4.2.1	Single-cell and bulk RNA sequencing	72
4.2.2	Spatially resolved transcriptomics data	73
4.3	iDAS Framework	74
4.3.1	ANOVA-based test to classify genes	75
4.3.2	Differential expression analysis for signature identification and interpretation	80
4.4	Application on single-cell data	81
4.4.1	Identification of cell states specific marker (main effect F_1)	82
4.4.2	Identification of responding and non-responding signatures (main effect F_2)	84
4.4.3	Identification of cell state specific responding signature (interaction effect $F_1 \times F_2$)	85
4.5	External validation on bulk data	86
4.6	Extensibility of iDAS	88
4.7	Discussion and conclusion	91
5	CONCLUSION	94
A	APPENDIX FOR CHAPTER 2	100
B	APPENDIX FOR CHAPTER 3	109
C	APPENDIX FOR CHAPTER 4	114
C.1	Supplementary figures	114
C.2	Supplementary tables	121
	BIBLIOGRAPHY	130

INTRODUCTION

The cell is the fundamental unit of a living organism. Single-cell research is about studying the behaviour and characteristics of cells, including the type of cells, how cells communicate with each other, how cells changes under different conditions and many more. With the rapid technological development in this field, we now can measure the molecular profile of single cells. The explosion of single-cell data enables scientists to systematically study cell function, development and differentiation, facilitating the understanding of disease and the discovery of new personalised treatment strategies, paving the way for precision medicine.

The typical single-cell dataset now ranges from thousands to hundreds of thousands of cells with different omic modalities. The complexity of these extremely high-dimensional datasets poses both challenges in data science and opportunities in addressing biomedical questions. Examples of these challenges and opportunities include discovering complex differential patterns in gene expression (Lähnemann et al., 2020), integrating single-cell data across multiple modalities (Argelaguet et al., 2021), and identifying rare cell types (Kiselev et al., 2019). Taking on these opportunities and addressing these issues will help unlock the full potential of single-cell technology. Therefore, novel computational techniques must be developed.

This thesis examines some of these challenges and opportunities associated with "cell types." First, we evaluate the estimation of the number of clusters to better understand the number of cell types. Next, we use multimodal single-cell omics data to better identify different cell types. Finally, we leverage cell

type information to construct better signatures. My specific contribution to the single-cell research field is as follows.

1. Benchmarked clustering methods on their ability to estimate the number of cell types, addressing the question of effectively identifying major and minor cell types.
2. Developed a multi-omics ensemble deep learning clustering method, it improves cell type annotation by incorporating multi-omics information compare to single modality (scRNAseq), and enhances the speed of integrating multimodal single-cell omics data under an ensemble deep learning model.
3. Developed an interpretable algorithm to identify differential abundance signatures linked with cell type and experimental factors, addressing the challenge of understanding how cell populations respond to perturbations in single-cell studies by the gene signatures.

In this chapter, we will begin with an introduction to the single-cell technologies (Section 1.1). Next, we review the typical steps in single-cell RNA data analysis (Section 1.2). We will then discuss some of the key computational challenges in method development (Section 1.3). Lastly, we will outline the content of this thesis in Section 1.4.

1.1 SINGLE-CELL SEQUENCING TECHNOLOGIES

Since the first single-cell RNA sequencing technology was published in 2009, molecular biology research has been dramatically transformed by the analysis of transcriptome profiles at the level of individual cells. Today, as shown in Figure 1.1, Smart-seq2, Drop-seq and 10X Genomics have emerged as the three dominant single-cell RNA sequencing technologies, each with more than 3000 citations. The most important step in single-cell sequencing is the process of isolating individual cells, using technologies such as microfluidics for cell isolation and fluorescence-activated or magnetic-activated cell sorting (FACS or MACS)

(Hu et al., 2016). Using single-cell isolation technology, a variety of single-cell sequencing protocols have been established to address different research needs and technological challenges. Notable methods such as Smart-seq2 (Picelli et al., 2013), SCR-seq (Soumillon et al., 2014) and MARS-seq (Jaitin et al., 2014) are often coupled with FACS, while CEL-seq2 Hashimshony et al. (2016), Drop-seq (Macosko et al., 2015), inDrops (Zilionis et al., 2016), 10x Genomics (Zheng et al., 2017) and Seq-Well (Aicher et al., 2019) use advanced microfluidic techniques.

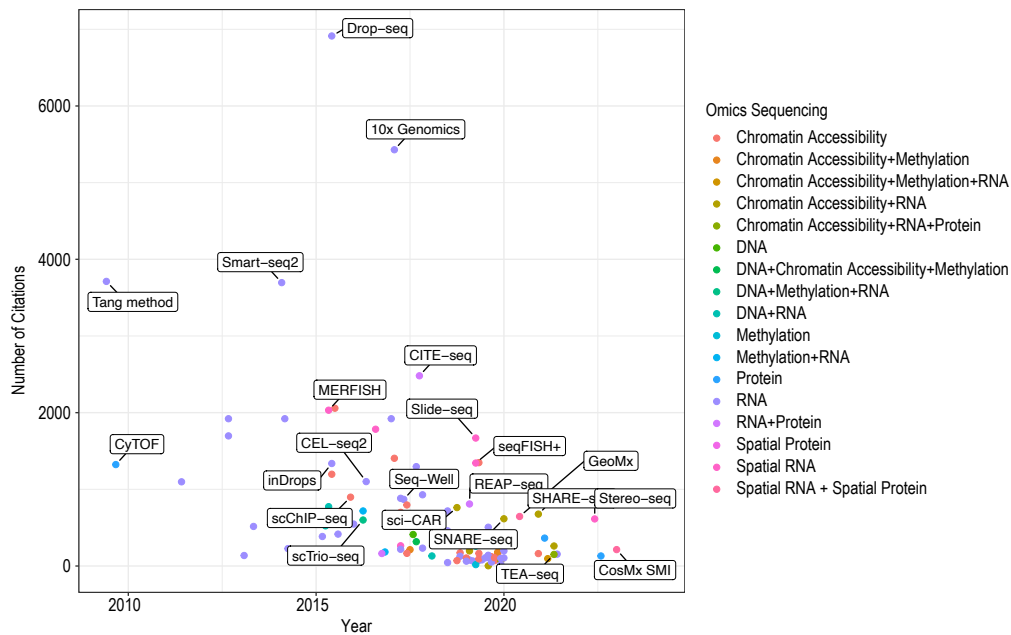


Figure 1.1: Scatter plot of single-cell omics technologies with the x-axis representing the technologies' paper publication year, and the y-axis showing the total number of citations of the original publication. The points are colored by the omics sequencing type.

In recent years, the types of single-cell sequencing technologies have expanded beyond transcriptome profiling. Scientists can apply scBS-seq (Smallwood et al., 2014) to analyse DNA methylation patterns, scATAC-seq (Buenrostro et al., 2015) and scHi-C (Nagano et al., 2013) to study chromatin accessibility and 3D chromatin interactions, scChIP-seq (Rotem et al., 2015) to learn about histone modification in single cells and CyTOF (Bandura et al., 2009) to analyse protein abundance.

There has also been a surge in advanced single-cell multimodal sequencing technologies. Examples of joint profiling of two modalities include CITE-seq

(Stoeckius et al., 2017) and REAP-seq (Peterson et al., 2017), which jointly profile RNA expression and protein abundance, and sci-CAR (Cao et al., 2018), SNARE-seq (Chen et al., 2019a) and SHARE-seq (Ma et al., 2020), which jointly analyse RNA and chromatin accessibility. There are also technologies that allow scientists to profile three modalities together at the single-cell level, such as TEA-seq (Swanson et al., 2021), scTrio-seq (Hou et al., 2016), and scNMT-seq (Clark et al., 2018). This remarkable diversity of methods underlines the dynamic and rapidly evolving landscape of single-cell multimodal sequencing, a field recognised by Nature Methods as the "Method of 2019" (Nature, 2020).

In addition, a growing field in single-cell sequencing technology is spatial transcriptomics. It can capture the location information on slides with the transcriptomics of each location, bringing single-cell analysis into the new era of spatially resolved dimensions. The most popular methods within this domain are categorized by their underlying technologies (Chen et al., 2023): Sequencing-Based Methods such as 10X Genomics Visium, Slide-seq (Rodriques et al., 2019), and Stereo-seq (Chen et al., 2022). Probe-Based Methods like GeoMx. And Imaging-Based Methods, including MERFISH (Chen et al., 2015), seqFISH+ (Eng et al., 2019), and CosMx SMI.

1.2 COMPUTATIONAL ANALYSIS ON SINGLE-CELL OMICS DATA

This chapter provides an overview of the typical scRNA-seq analysis workflow and the general tools involved at each stage. In its simplest form, the workflow consists of the following steps:

Initial data analysis (IDA): In this stage, the sequencing data is prepared to generate a cell-by-gene expression matrix, followed by quality control measures to refine the dataset for further analysis. The next steps may include imputation, normalisation, batch correction, data integration, and dimension reduction to ensure that the data is optimised for downstream analysis.

Downstream analysis: This stage involves the analysis of single-cell via cell annotation and gene identification. Key analysis includes clustering, cell annotation, cell trajectory analysis and differential expression analysis.

1.2.1 *Initial data analysis (IDA)*

1.2.1.1 *Quality assessment and droplet detection*

Due to technical issues in single-cell RNA sequencing (scRNAseq) such as the low-quality cells and existence of doublets, scRNA-seq expression data often contains low-expressed cells and genes. To obtain high quality data for downstream analysis, apoptotic or lysed cells with high levels of mitochondrial gene expression are often removed during quality assessment. A simple way to filter out low-expressed cells and genes is to apply cutoff-based filtering. Commonly used packages include Scater (McCarthy et al., 2017) and Scrublet (Wolock et al., 2019). In addition to removing low quality cells, we can also address missing values in gene expression data using imputation methods such as scImpute (Li and Li, 2018), SAVER (Huang et al., 2018), DrImpute (Gong et al., 2018) and scVI (Lopez et al., 2018).

1.2.1.2 *Normalisation*

Normalisation is used to remove technical variation. There are two different ways to normalise single-cell RNA sequencing data. One way is to normalise the UMI (Unique Molecular Identifier) count to the library size, which is typically applied for 10X datasets. A widely used tool for this approach is Scrn (Lun et al., 2016), which uses a pooling-based strategy to estimate size factors for each cell. Another popular method is sctransform (Hafemeister and Satija, 2019), which uses a regularized negative binomial regression to model gene expression as a function of library size and other covariates, such as the percentage of mitochondrial genes. The other way is to normalise the read count to FPKM or TPM, typically for Smart-seq2 datasets.

1.2.1.3 *Batch correction and data integration*

To increase the amount of single-cell data available for analysis, more and more scientists are collecting data from multiple batches or combining publicly available data from different groups and experimental conditions for analysis. To better integrate multiple datasets, we need to perform batch correction, which aims to remove unwanted variation in the data to minimise potential confounding in downstream analyses. To date, several tools and algorithms have been developed to address batch effects, including Harmony (Korsunsky et al., 2019), scMerge (Lin et al., 2019), LIGER (Liu et al., 2020), Scanorama (Hie et al., 2019) and ComBat.

In addition to the tools mentioned, other widely used integration methods include Seurat’s canonical correlation analysis (CCA) approach (Butler et al., 2018) and its anchor-based method (Stuart et al., 2019), which have been extensively applied in single-cell data integration. Recent benchmarking studies (Tran et al., 2020; Luecken et al., 2021) provide a comprehensive comparison of batch correction and data integration methods, and highlight their strengths and limitations. Choosing the most appropriate method depends on the dataset and specific analysis goals. For example, Harmony performs well on datasets with shared cell types and across different technologies, while scMerge is effective at recovering differentially expressed genes in batch-corrected data.

1.2.1.4 *Dimension reduction*

Dimension reduction is the process of reducing the number of features, either by feature selection or feature compression. Feature selection is usually done by identifying highly variable genes (HVGs), methods include BASiCS (Vallejos et al., 2015), Brennecke (Brennecke et al., 2013), and scLVM (Buettner et al., 2015). Feature compression is usually done by compressing the original features into a low dimensional space, techniques include PCA, LDA, t-SNE (van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018), as well as deep learning methods such as autoencoder and variational autoencoder.

1.2.2 *Downstream analysis*

1.2.2.1 *Clustering*

Clustering is a fundamental task in the analysis of scRNA-seq data and can assist in identifying the number of cell-types and -states. This is an essential preliminary step before further downstream analyses such as differential expression and cell trajectory analysis. Over one hundred tools have been released for clustering cells which attempt to address a range of technical and biological challenges (Zappia et al., 2018; Zappia and Theis, 2021). The current most popular approach utilizes graph-based clustering (e.g., Louvain clustering, Leiden clustering) and has been implemented in many single-cell pipeline packages, such as Seurat (Butler et al., 2018) and Scanpy (Wolf et al., 2018). In addition to traditional methods, deep learning techniques have also been developed for clustering in single-cell omics data. These approaches leverage neural network architectures to enhance clustering performance by learning representations that are more conducive to grouping similar cells. Notable examples of deep learning-based tools include scCCESS (Geddes et al., 2019), scAIDE (Xie et al., 2020), and scDeepCluster (Tian et al., 2019), which offer new solutions for tackling the complexities of single-cell data analysis.

1.2.2.2 *Cell type annotation*

After clustering similar cells into groups, each cluster is annotated to known cell types using marker genes. While manual annotation is the standard approach, this can also be done using classification methods with previously annotated references, such as the Human Cell Atlas (Regev et al., 2017). An advantage of classification is that it uses existing knowledge to avoid time-consuming manual annotation and interpretation of clusters. Commonly used supervised classification methods for cell type annotation include scmap (Kiselev et al., 2018), CaSTLe (Lieberman et al., 2018), CellTypist (Domínguez Conde et al., 2022) and SingleR (Aran et al., 2019). However, these methods typically classify cells into a single level of cell types, which cannot help researchers un-

derstand the hierarchical structures between cell types. To address this, several multi-scale classification methods have been developed to address the issue of the hierarchical structure of cell types, such as scPred (Alquicira-Hernandez et al., 2019), CHETAH (de Kanter et al., 2019) and scClassify (Lin et al., 2020). Additionally, the anchor-based method in Seurat (Stuart et al., 2019) can also be used to transfer cell type labels from reference datasets to query data.

1.2.2.3 *Trajectory inference*

Besides traditional methods that classify cells into distinct groups, there are other analysis strategies have been developed for certain biological scenarios. For example, in developmental studies, methods to arrange cells along a continuous trajectory are more informative than classifying them into distinct groups. Methods in this area include Monocle (Van den Berge et al., 2020), TSCAN (Ji and Ji, 2016), SLICER (Welch et al., 2016), and Slingshot (Street et al., 2018).

1.2.2.4 *Differential expression analysis*

Once cells have been organised by clustering or trajectory analysis, the next step is to decipher the biological significance of these groups. For datasets with clustered cells, this typically involves identifying genes that show differential expression (DE) across groups, or marker genes that are unique to a single cluster. A number of methods have been proposed to test for DE, including Limma (Ritchie et al., 2015), edgeR (Robinson et al., 2009) and DEseq2 (Love et al., 2014). Originally designed for bulk sample analysis, these methods can also be applied to single cells by aggregating expressions to create "pseudobulk" samples that can then be analysed using bulk analysis methods. In addition, specialised tools such as DESingle (Miao et al., 2018), MAST (Finak et al., 2015), scDD (Korthauer et al., 2016) and SCDE (Kharchenko et al., 2014) have been developed specifically for single-cell analysis. In addition, approaches such as Cepo (Kim et al., 2021) have introduced the concept of differential stable expression to single-cell research, further enriching the toolkit available for interpreting grouped or ordered cell data.

1.3 CHALLENGES OF CLUSTERING SINGLE-CELL OMICS DATA AND INTERPRET GENE SIGNATURES

Both single-cell clustering and gene signature analysis (e.g. DE analysis) have been widely used in disease-related studies to help scientists better understand cell types within diseases. For example, Mathys et al. (2019) used cell clustering to discover new neuronal subtypes that play distinct roles in neuronal circuitry and Alzheimer’s disease. In oncology, Nofech-Mozes et al. (2023) used clustering to identify distinct subpopulations of cancer, immune and stromal cells to better understand the tumour microenvironment. In addition, single-cell RNA sequencing and differential expression (DE) analysis are helping scientists discover gene signatures that differ in drug sensitivity from different cancer subpopulations (Aissa et al., 2021). In this section, we provide a brief introduction to the two techniques and the challenges addressed in this thesis.

1.3.1 Cell type clustering

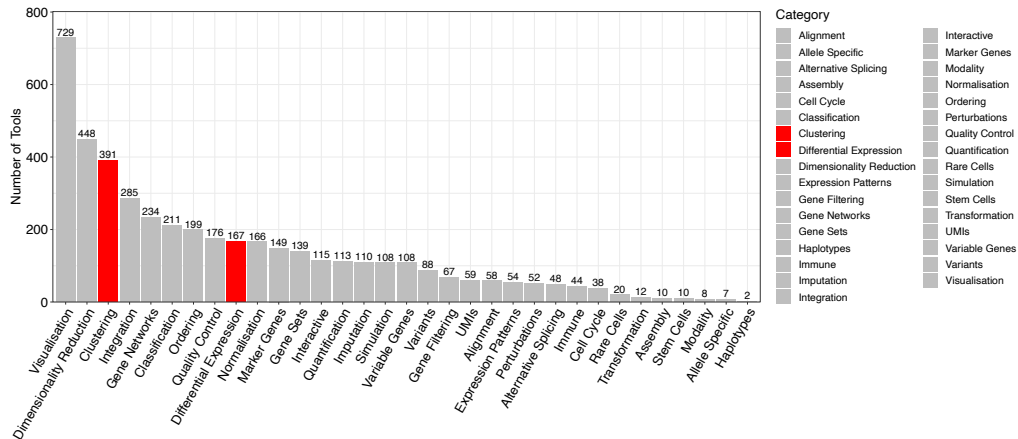


Figure 1.2: Barplot of single-cell omics analysis tools included on the scRNA-tools website. The x-axis represents the categories of each tool, and the y-axis shows the total number of tools in each category. Bars related to clustering and differential expression are highlighted in red.

To date (May, 2024), 391 single-cell RNA-seq clustering methods have been published (see Figure 1.2) and can be categorised according to several factors (Zappia et al., 2018; Zappia and Theis, 2021). These include the ability to handle

multi-omics data (such as scRNA-seq, CITE-seq, and spatial data), automatic estimation of the number of clusters, and the capability to partition data into multiple resolutions or scales. Furthermore, the methodologies that underpin the algorithms can vary and present diverse approaches for achieving the same goal: including deep learning-based methods, hierarchical clustering, centroid-based clustering, and graph community detection. The following paragraphs provide a brief overview of cell clustering steps in single-cell omics data and commonly used methods.

The first step in single-cell clustering analysis is quality control and normalisation of the data. This can help to filter out doublets and remove technical confounders. Next, highly variable genes are selected for downstream analysis. If there are still a large number of features in the data, dimension reduction techniques can be applied to compress the data and further minimise the feature space. Subsequently, any clustering method and graph community detection approach can be applied to the data to assign cells to clusters, such as Seurat(Butler et al., 2018), CIDR (Lin et al., 2017), GiniClust (Jiang et al., 2016), SC₃ (Kiselev et al., 2017).

Below is a brief introduction to each of the above clustering methods.

- **Seurat** is a community detection based method, it calculates the cell-cell similarity by gene expression to build a cell-cell graph and then applies community detection technologies such as Louvain or Leiden to find the communities (clusters) from the graph.
- **CIDR** is specifically designed to deal with dropout and to impute the gene expression matrix by "implicit imputation", it uses a hierarchical method to cluster cells.
- **GiniClust** uses the Gini index to identify genes that are either up- or down-regulated in rare cells and uses DBSCAN algorithm (Ester et al., 1996) to cluster the cells.

- **SC₃** uses different distance calculation methods to construct multiple distance matrices and, after multi-step processing, produces a consensus matrix, and finally applies hierarchical clustering to the consensus data.

In addition to single-cell RNA sequencing, multimodal single-cell sequencing technologies can now provide multiple layers of cellular information. These advances can dramatically improve cell type clustering through additional informatics, such as the CITE-seq technique, which provides cell surface protein data to better annotate cells into specific cell types. Multi-omics data clustering relies critically on the integration of different omics layers using strategies such as early, intermediate and late data integration. Early integration, where multiple datasets are concatenated at the beginning, is less favoured due to increased dimensionality and variable distribution issues. Intermediate integration is characterised by its efficiency. Some methods include CiteFuse for similarity-based integration, intNMF (Chalise and Fridley, 2017) and MOFA+ (Argelaguet et al., 2020) for joint dimensionality reduction, and Bi-CCA (Dou et al., 2022) for statistical modelling. Deep learning approaches are also commonly used, such as totalVI (Gayoso et al., 2021), multiVI (Ashuach et al., 2023) and GLUE (Cao and Gao, 2022), which transform data sets into low dimension embeddings for clustering. Late data integration involves analysing each omics layer separately before integrating the results to reach a consensus on cell clustering, using techniques such as the Monti consensus clustering algorithm (Monti et al., 2003) to determine optimal cluster numbers.

1.3.2 *Interpret gene signatures*

Over the past 5 years, there has been a remarkable increase in the availability of single-cell multi-sample and -condition data. This development has shifted the focus from the mere clustering of cell types to the in-depth investigation of cell-level responses to biological perturbations. Such studies allow the detailed characterisation of different cell states. A variety of cell state identification methods have been developed to help researchers better understand cell states in

response to biological perturbations. These methods can be roughly divided into two categories, differential abundance identification (using tools such as MiloR (Dann et al., 2021), DASEq (Zhao et al., 2021) and MELD (Burkhardt et al., 2021)) and phenotype-associated cell subpopulation identification (using approaches such as CNA (Reshef et al., 2021), DEGAS (Johnson et al., 2022), Augur (Skinnider et al., 2020), and Scissor (Sun et al., 2021)). In the first category, differential abundance analysis is used to uncover cell subpopulations whose presence correlates significantly with experimental perturbations. In contrast, the second category focuses primarily on determining whether each cell is in a perturbed or control state.

Once cell states have been identified, a key question is whether there are specific gene signatures associated with particular phenotype-related cell populations. Differential expression (DE) analysis is often used to answer this question. There are 167 published DE methods available for such analysis, as listed on the scRNA-tools website (see Figure 1.2). Among all methods, a clustering-free method called MiloDE (Missarova et al., 2023) has been developed to identify DE genes in the context of DA analysis. The authors suggest that miloDE can improve the sensitivity and specificity of DE testing compared to traditional fixed clustering methods. However, DE methods cannot identify signatures associated with multiple factors. While particular phenotype-related cell populations may include multiple cell types and other experimental factors.

1.3.3 *Challenges addressed in this thesis*

Building on our brief exploration of methods for cell type clustering, differential abundance and differential expression analysis, we now turn to the specific challenges of single-cell omics data analysis for understanding cell types. When dealing with large multi-omics data, there are still several challenges presents in single-cell clustering, such as the need for scalable computational methods to process growing single-cell datasets, and the need for the effective integration of multi-omics data to provide comprehensive insights. In addition, how

to better capture dynamic cell states and the identification of gene signatures with multiple experimental factors are also critical questions. Below we describe three specific challenges that we address in this work.

1.3.3.1 *Estimating the number of cell types*

Most clustering methods require the user to pre-specify parameters, such as the estimated number of cell types (k) or the resolution (r) for clustering. This requirement raises the question of how to automatically determine these parameters and thus reduce user-induced bias. This challenge is part of a larger question in the single-cell field: how to efficiently estimate the number of cell types and accurately divide the cells into clusters. It should also be noted that not all clustering methods have the ability to automatically estimate the number of clusters.

To help users choose the most efficient methods, in this thesis, we conducted a benchmarking study to evaluate single-cell clustering algorithms, specifically focusing on their ability to estimate the number of cell types. We also developed a stability-based ensemble clustering method, and explored the ability of ensemble methods to estimate the number of cell types accurately. This benchmarking aims to encourage the creation of new and more accurate clustering methods for single-cell omics data analysis.

1.3.3.2 *Multimodal single-cell omics data clustering*

Compared to classification, clustering approaches are easily adaptable to various omics data types, allowing for the integrative analysis of multiple layers of information within individual cells. To perform multimodal single-cell omics clustering, integration methods are generally using "intermediate integration". The multimodality integrated embeddings produced by intermediate integration methods can be employed for cell clustering using any clustering algorithms that accept embeddings as input. Consequently, the effectiveness and quality of the multimodality-integrated embeddings will significantly influence the clustering results in multimodal single-cell omics data.

In this thesis, we introduce SnapCCESS, a deep learning-based ensemble framework that creates multi-view, multimodality-integrated embeddings to enhance clustering of multimodal single-cell omics data and improve computational speed. By generating snapshots of embeddings using variational autoencoders, SnapCCESS can be coupled with various clustering algorithms to produce consensus clustering of cells. Our results show that SnapCCESS not only achieves high-performance cell clustering but also addressing the computational efficiency problems associated with conventional ensemble deep learning frameworks.

1.3.3.3 *Interpret differential abundance signature analysis*

Current methods for identifying perturbation states at the single-cell level vary significantly. For example, approaches based on abundance analysis and those relying on bulk-level information often identify single cells associated with different drug phenotypes, leading to complex interpretations of the sub-population gene signatures derived from these methods. This variation highlights the need for a method that offers a more interpretable way to characterize the gene signatures associated with different experimental or patient factors.

In this thesis, we used ANOVA-like hypothesis tests to help interpret gene signatures associated with two factors and created a framework called iDAS. We applied this framework on a melanoma single-cell dataset, classifying genes associated with cell type, immunotherapy responding phenotypes, and their interactions. By interpreting gene signatures linked to interaction effect, we identified some genes as tissue-specific markers of responding or non-responding to immunotherapy. We further validate the findings using external bulk datasets.

1.4 THESIS OUTLINE AND CONTRIBUTIONS

This thesis addresses three challenges in single-cell omics data analysis, incorporating methods, analyses, and results that the author has published (or ongoing project) and contribute to the field of single-cell data science. Chapter 2

focuses on estimating the number of cell types. It proposes a clustering stability-based ensemble clustering method for this purpose and systematically benchmarks the current methods in their ability to estimate cell type numbers. This work has been published in *Genome Biology*. Chapter 3 presents an ensemble deep learning-based multimodal single-cell clustering method, demonstrating the efficacy of a variational autoencoder model to integrate multimodal single-cell data, including CITE-seq, SHARE-seq, SNARE-seq, and TEA-seq data, and has been published in *Bioinformatics*. Chapter 4 addresses the challenge of interpreting complex differential abundance signatures by developing a framework to classify signatures into groups related to multiple factors. This framework identifies genes associated with multiple experimental factors. We demonstrate its application using a melanoma single-cell data and validate some of the results with two external bulk datasets.

Together, these chapters tackle multiple inter-connected challenges in single-cell data analysis, moving from cell clustering to gene signature classification. Clustering is a fundamental step that lays the groundwork for identifying distinct cell types and states, enabling researchers to map cellular heterogeneity and structure. However, clustering alone is not sufficient to fully characterise cellular responses to biological perturbations or treatment conditions. Building on this foundation, later chapters of the thesis shift the focus to gene signature classification, specifically the identification of differential abundance and phenotype-associated signatures. This transition highlights how clustering insights are essential for providing the structural context needed for the functional analyses of cellular behaviour.

By integrating clustering-based insights with gene signature classification, this thesis provides some tools for characterizing both the structural and functional aspects of cellular behavior in complex datasets. For example, improving the ability to estimate cell type numbers and integrating multimodal data for clustering enables better characterization of cellular heterogeneity, which is critical for understanding tissue composition in health and disease. Additionally, the framework for differential abundance signatures identifies gene markers that

most reflect treatment responses or disease progression while avoiding interference from other experimental factors, contributing to translational research efforts.

In summary, this thesis explores strategies to estimate the number of cell types, applies ensemble deep learning for clustering multimodal omics data, and method for finding interpretable signatures related to experimental or patient factors. These approaches address specific technical challenges in single-cell data analysis. The publicly available frameworks and methods developed in this thesis will benefit the current research community and inspire future methodologies.

A BENCHMARK STUDY OF ESTIMATING THE NUMBER OF CELL TYPES FROM SINGLE-CELL RNA-SEQUENCING DATA

Clustering plays an important role in the analysis of scRNA-seq data. It enables the identification and categorization of single cells into similar groups, such as cell types, for a better understanding of cellular diversity and its function in health and disease. Traditionally, the community has relied on clustering algorithms to group cells and estimate the number of cell types (k). So clustering is often a necessary step for annotating cell types and guiding downstream analyses such as visualization or differential expression studies. Although a large number of algorithms have been designed and evaluated, the task of accurately estimating the number of cell types present in scRNA-seq datasets has not been as thoroughly explored.

Estimating the number of cell types (k) in scRNA-seq datasets has a lot of challenges. scRNA-seq data is highly noisy due to technical variation introduced during cell capture, library preparation, and sequencing. This noise can make it hard to clearly define the boundaries between cell types. Additionally, cell populations in biological tissues are often hierarchical and continuous rather than discrete. For example, transitional or intermediate cell states can blur the lines between clusters, which makes it difficult to define clear-cut clusters that represent distinct cell types. Another challenge is that these datasets often contain rare or underrepresented cell types, which are easily overlooked by clustering algorithms optimized for larger populations.

Several studies have evaluated the performance of various clustering algorithms on clustering the cells into known types. While many of these clustering algo-

rithms also offer to estimate the number of cell types in a dataset, this task has mostly been ignored in their performance assessment. Given the potential impact of this step in cell clustering and identifying new cell populations, this gap highlights the need for systematic evaluation and comparison of existing methods under a common framework to understand their strengths, limitations and applicability to various scRNA-seq data characteristics.

To address this gap and catalyse future advancements, in this chapter, we describe the establishment of a systematic benchmark framework designed specifically for evaluating the performance of clustering algorithms in estimating the number of cell types from scRNA-seq data. We tested 13 single-cell clustering algorithms, spanning four major categories, including proposing a novel stability-based approach, called "scCCESS" (Geddes et al., 2019), against a set of performance metrics. The contents of this chapter are published in *Genome Biology* in 2022 (Yu et al., 2022). To the best of our knowledge, this is the first systematic benchmark study on single-cell clustering methods for estimating the number of cell types. The outcomes of this extensive evaluation were distilled into a comprehensive set of recommendations for users, aiding in the selection of the most suitable clustering tools for their specific needs. Moreover, the development and availability of our stability-based approach as an R package represents a significant contribution to the toolkit available to researchers, offering a new method for tackling the challenge of cell type estimation in scRNA-seq data analysis. As the first author, my contribution towards this project involved performing the entire benchmarking analysis and extending the original scCCESS script with function to estimate the number of cell types and build the package with input from A/Prof. Pengyi Yang, Prof. Jean Yang and Dr. Yue Cao.

2.1 INTRODUCTION

Much attention has been given to clustering cells into cell type groups. Estimating the number of cell types in a given scRNA-seq dataset is a closely related

task and critical for scRNA-seq data analysis but has so far received less attention. This is highlighted in the literature, where only a small proportion (Zhang et al., 2023b) of the large number of clustering algorithms (Duò et al., 2020; Freytag et al., 2018) designed for cell clustering is capable of estimating the number of cell types. Estimating the number of cell types can be considered as finding the optimal number of clusters for a given scRNA-seq data with the assumption that each cluster corresponds to a unique cell type in the dataset (Andrews and Hemberg, 2018). Under this assumption, current clustering methods that estimate the number of cell types can be loosely classified into the following categories: (i) intra- and inter-cluster similarity, (ii) modularity in community detection, (iii) eigenvector-based metrics, and (iv) stability metrics. Given the lack of systematic evaluation of clustering algorithms on their performance on estimating the number of cell types, in this study, we set out to systematically assess the estimation of the number of cell types for a collection of clustering algorithms from each category summarised below.

Intra- and inter-cluster similarity is one of the most widely applied approaches for estimating the optimal number of clusters in a given dataset (Kothari and Pitts, 1999; Sugar and James, 2003). This involves calculating indices that measure the closeness of items in each cluster and separations among clusters. In scRNA-seq data analysis, methods in this category include scLCA (Cheng et al., 2019) which uses Silhouette index (Rousseeuw, 1987), CIDR (Lin et al., 2017) which uses Calinski-Harabasz (CH) index (Caliński and Harabasz, 1974), and SHARP (Wan et al., 2020) that relies on both indices (Silhouette and CH) and hierarchical heights of the clustering to determine the number of clusters. RaceID (Grün et al., 2015) uses the Gap statistic (Tibshirani et al., 2001), follows the idea of intra- and inter-cluster similarity but introduces a statistical test to compare within-cluster dispersion. Similarly, SINCERA (Guo et al., 2015) uses a minimum distance approach to obtain “non-singleton” cell clusters. The second category of community detection-based techniques mostly relies on the Louvain algorithm (Blondel et al., 2008) and Leiden algorithm (Traag et al., 2019) to optimise community structure to find the best possible grouping. This strategy is implemented by a number of scRNAseq clustering methods including AC-

TIONet(Mohammadi et al., 2020), Monocle3 (Cao et al., 2019), and Seurat(Hao et al., 2021). The third category involves using eigenvector-based techniques, where the methods typically apply eigengap heuristic to estimate the number of cell types (von Luxburg, 2007). Examples such as SIMLR (Wang et al., 2017) partition the data into a specific number of clusters that maximise the eigengap. In Spectrum (John et al., 2019), the authors extended the idea of eigengap heuristic and built a multimodality gap heuristic algorithm in which can be applied to Gaussian or non-Gaussian structures. Similarly, SC3 (Kiselev et al., 2017) partitions the data by examining the eigenvalue based on the Tracy-Widom test (Tracy and Widom, 1994; Patterson et al., 2006). Finally, clustering stability is another commonly employed metric for determining the number of clusters in the computation and data science literature (Lange et al., 2004; Lord et al., 2017). The intuition behind these approaches is that clustering output, generated under the optimal number of clusters, would lead to more stable or reproducible clusters compared to those generated under suboptimal number of clusters. An example in this fourth category is densityCut (Ding et al., 2016), which estimates the number of cell types from a given dataset by modelling the density of cell distributions for generating a hierarchical cluster tree and subsequently selecting clusters that are most stable in the hierarchical cluster tree. In this study, we propose an alternative stability-based approach by taking advantage of scCCESS, a random sampling-based ensemble deep clustering model, previously proposed for scRNA-seq data clustering (Geddes et al., 2019) for estimating the number of cell types. Our key assumption is that clustering from using the optimal number of clusters would be the most robust to small perturbations in the data, such as those introduced by random resampling, compared to those generated under the suboptimal number of clusters.

Together with the current state-of-the-art methods and our proposed stability-based approach, we present a systematic and quantitative analysis of single-cell clustering algorithms, focusing on their performance on estimating the number of cell types. Specifically, we benchmark the above fourteen clustering methods from each of the four categories (Figure 1a) across a large number of datasets sampled from the Tabula Muris project (The Tabula Muris Consortium,

2018) representing different data characteristics in various settings. We evaluate the accuracy on determining the number of cell types, performance of cell clustering, and computing time and peak memory usage of each method on each of all datasets. We further cross-compared the performance of clustering algorithms on datasets with a large number of cells using both Tabula Muris and Tabula Sapiens data (The Tabula Sapiens Consortium, 2022). We summarise these findings into a multi-aspect recommendation to the users, and highlight potential areas requiring future research.

2.2 METHODS

2.2.1 *Sampling of Tabula Muris and Tabula Sapiens data*

To create datasets with varying but well-defined characteristics, we took advantage of the large number of cells and cell types profiled by the Tabula Muris project (The Tabula Muris Consortium, 2018) by subsampling from this dataset different numbers of cells and types and creating different imbalance ratios (described below). Specifically, Tabula Muris dataset contains 53,760 cells (FACS sorted and sequenced using Smart-Seq2 protocol) from 81 cell types of 20 organs of 7 mice. Cell types that include no more than 300 cells were excluded from this study. Overall, 38 cell types with 39,712 cells and 23,433 genes were included in the selection pool after filtering.

Four settings were considered in this benchmark study. The first is to benchmark the performance of clustering algorithms on estimating the number of cell types in datasets that contain different numbers of cell types. In particular, we held the number of cells in each cell type at a constant of 200, while varying the number of cell types by randomly sampling from all 38 candidate cell types in Tabula Muris dataset 200 cells each from 5 to 20 cell types (by step of 1). We repeated this sampling procedure 10 times where 10 datasets were sampled for each number of combination for the purpose of estimating variability, resulting in a total of 160 datasets (i.e. $(20 - 4) \times 10$).

In the second setting, we benchmarked the impact of different numbers of cells in each cell type on the number of cell type estimation. In this setting, we set the number of cell types in a dataset to be 5, 10, 15, and 20 by sampling from the 38 cell types in the Tabula Muris dataset and varied the number of cells in each cell type from 50 to 250 (increment by 50). Similar to above, we repeated the sampling 10 times for each number of cell type and number of cell combination, resulting in a total of 200 datasets (i.e. $4 \times 5 \times 10$).

In the third setting, we tested the impact of datasets with different imbalance ratios by creating major and minor cell types that contain different numbers of cells. First, we sampled from Tabula Muris dataset 10 or 20 cell types and divided them into two equal groups for creating major and minor cell types (i.e. 5 or 10 cell types each for 10 and 20, respectively). We then created major and minor cell types by setting the number of cells in the major cell types as 200 and varying the number of cells in the minor cell types as 100, 50, and 20, leading to imbalance ratios of 2:1, 4:1, and 10:1 for the major and minor cell types. Again, we repeated the sampling 10 times, resulting in 60 datasets (i.e. $2 \times 3 \times 10$).

To assess the performance of clustering algorithms on datasets with large number of cells while also validating if the benchmark results obtained from using Tabula Muris datasets are consistent compared to datasets sampled from another source, in the last setting, we sampled from Tabula Sapiens data (The Tabula Sapiens Consortium, 2022) to create datasets with either five or ten cell types and each with either 500 or 1000 cells in each cell type. These lead to datasets with 2500 to 10,000 cells. The same sampling procedures were repeated on the Tabula Muris data to create datasets with the matching sizes so the performance of each clustering algorithm on the datasets with same sizes can be compared across different data sources. As above, we repeated the sampling 10 times, resulting in 80 datasets with large numbers of cells (i.e. $2 \times 2 \times 2 \times 10$).

Finally, all subsampled datasets are un-normalised raw count matrices. For methods that require normalised and/or log-transformed count matrix, we con-

verted the raw count matrix to log-normalised count matrix using the ‘scater’ (McCarthy et al., 2017) package.

2.2.2 *Proposed clustering stability-based approach for number of cell type estimation*

Building on single-cell Consensus Clusters of Encoded Subspaces (scCCESS), an ensemble clustering algorithm we developed previously (Geddes et al., 2019), here we propose a clustering stability-based approach for the number of cell type estimation. scCCESS implements an autoencoder-based cluster ensemble for single cell clustering. It first generates multiple random projections from the original input scRNA-seq dataset and trains a collection of autoencoders, unsupervised deep learning neural networks, each on a random projection. It then encodes the data to multiple low dimensional data from which multiple clustering outputs are generated for creating the ensemble. We hypothesised that the number of cell types in a dataset is best estimated when multiple clustering output, each from a random project and dimension reduction, are highly concordant with each other (i.e. high stability). To this end, we take advantage of the multiple clustering output generated from the collection of autoencoders by quantifying stability of these clustering results across a range of k values using either k -means or SIMLR (Wang et al., 2017) clustering algorithms and use the k with the highest overall stability score (median of all pairwise concordance scores measured using normalised mutual information [NMI]) as the estimation of the number of cell types in an input scRNA-seq dataset.

2.2.3 *Clustering methods that estimate the number of cell types*

We examined eleven established single-cell clustering approaches as well as two clustering stability-based methods that we proposed above. The methods were chosen from the scRNA-tools database (Zappia et al., 2018), representing a wide range of popular clustering algorithms used for cell clustering and the number of cell type estimation from scRNA-seq data (Figure 2.1a). Table 1 summarises

the details of each approach, including the version of the code utilised in this benchmark analysis and its publication.

Table 1: scRNA-seq clustering methods for number of cell type estimation evaluated in this study.

Methods	Platform	Clustering type	Category	Ref.	Version
Monocle3	R	Leiden clustering	Community detection	Cao et al. (2019)	0.2.3.0
scLCA	R	Spectral clustering	Intra- and inter-cluster similarity	Cheng et al. (2019)	0.0.0.9
scCCESS-SIMLR	R	Ensemble of SIMLR	Stability metric	Geddes et al. (2019)	0.0.1
ACTIONet	R/C++	Leiden clustering	Community detection	Mohammadi et al. (2020)	2.0.18
Seurat	R	Louvain clustering	Community detection	Hao et al. (2021)	4.0.1
scCCESS-Kmeans	R	Ensemble of K-means	Stability metric	Geddes et al. (2019)	0.0.1
CIDR	R	Hierarchical clustering	Intra- and inter-cluster similarity	Lin et al. (2017)	0.1.5
SC3	R	Hierarchical clustering	Eigenvector-based metrics	Kiselev et al. (2017)	1.18.0
SIMLR	R	Spectral clustering	Eigenvector-based metrics	Wang et al. (2017)	1.18.0
RaceID	R/C++	K-means	Intra- and inter-cluster similarity	Grün et al. (2015)	0.2.3
SINCERA	R	Hierarchical clustering	Intra- and inter-cluster similarity	Guo et al. (2015)	0.99.0
Spectrum	R	Spectral clustering	Eigenvector-based metrics	John et al. (2019)	1.1
densityCut	R	Hierarchical clustering	Stability metric	Ding et al. (2016)	0.0.1
SHARP	R	Meta-clustering	Intra- and inter-cluster similarity	Wan et al. (2020)	1.1.0

All parallel computing methods, including RaceID, CIDR, ACTIONet, scCCESS, SC3, and SHARP, were benchmarked by using eight cores. For methods such as scCCESS, SIMLR, scLCA, and Spectrum that allow explicit specification of the range of k values to be evaluated, we tested these from 2 to the upper bound of 25 for estimating the number of cell types. For scCCESS, the ensemble size was set as 20. When using SIMLR, we set the principal component as 15 and used SIMLR_Large_Scale function to avoid the data size limitation problem on SIMLR function. All other parameters of each method were set as the default values. To benchmark the elapsed time and peak memory usage, we evaluate all processing steps of each method, including gene and cell filtering if they are part of clustering steps implemented in a clustering package.

2.2.4 Cell clustering evaluation metrics

To benchmark the cell clustering results from the four settings, four evaluation measures were employed to quantify the concordance of clustering results on each scRNAseq dataset with respect to their predefined cell-type annotations (Kim et al., 2019). These included adjusted Rand index (ARI), normalised mutual information (NMI), Fowlkes–Mallows index (FM), and Jaccard index (Jaccard).

Let S be a set of N cells, then a clustering U on S is a way of partitioning S into non-overlap subset $\{U_1, U_2, \dots, U_R\}$, where $\cup_{i=1}^R U_i = S$ and $U_i \cap U_j = \emptyset$ for $i \neq j$. Here, we define $U = \{U_1, U_2, \dots, U_R\}$ as the gold standard cell type labels, $V = \{V_1, V_2, \dots, V_C\}$ is a partition generated by a clustering. Pair counting based measures can be used for counting pairs of items on which the partition U and V agree or disagree. Specifically, the $\binom{N}{2}$ item pairs in S can be classified into one of the four types: (i) N_{11} : the number of pairs that are in the same partition in both U and V ; (ii) N_{00} : the number of pairs that are in different partitions in both U and V ; (iii) N_{01} : the number of pairs that are in the same partition in U but in different partitions in V ; (iv) N_{10} : the number of pairs that are in different partitions in U but in the same partition in V . Following this, ARI, NMI can be defined as follows (Wagner and Wagner, 2007; Vinh et al., 2010):

$$ARI(U, V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}$$

$$NMI(U, V) = \frac{I(U; V)}{H(U) + H(V)}$$

$$FM(U, V) = \sqrt{\left(\frac{N_{11}}{N_{11} + N_{01}}\right) \left(\frac{N_{11}}{N_{11} + N_{10}}\right)}$$

$$Jaccard(U, V) = \frac{N_{11}}{N_{11} + N_{10} + N_{01}}$$

where $I(U; V)$ is the mutual information between U and V , defined as:

$$I(U; V) = \sum_{i=1}^R \sum_{j=1}^C \frac{|U_i \cap V_j|}{N} \log_2 \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

and $H(\cdot)$ is the entropy of partitions, in which $H(U)$ and $H(V)$ are calculated:

$$H(U) = - \sum_{i=1}^R \frac{|U_i|}{N} \log_2 \frac{|U_i|}{N}$$

$$H(V) = - \sum_{j=1}^C \frac{|V_j|}{N} \log_2 \frac{|V_j|}{N}$$

2.2.5 Assessment of the run time and peak memory usage

All benchmark tasks were allocated on a research server with dual Intel(R) Xeon(R) CPU E5-2637 v4 @ 3.50GHz processor (16 cores and 64 Gb total memory). The elapsed run time was evaluated by the R function `system.time`; timings for each method include all pre-processing steps. The usage of peak memory was monitored by R function `gc`. The same seed was set for all steps involving stochasticity (i.e. dimension reduction and clustering) in each evaluating task.

2.2.6 Performance summary criteria

Figure 8 summarises the performance of the evaluated methods across four criteria categories, including (i) deviation of the estimated number of cell types from the ground truth, which assesses the ability to estimate the number of cell types under the four settings; (ii) variability of the number of cell type estimation, which evaluate the variability of the estimated number of cell types across the four settings; (iii) cell clustering concordance with respect to the predefined cell type labels in the four settings, and (iv) the average speed and memory required for clustering across the four settings. For each metric, the performance of each method is considered as “good”, “intermediate”, or “poor”. Here, we list the criteria used to categorise the methods for each evaluation metric.

1. Deviation of the estimated number of cell types compared to the ground truth, defined as $\frac{(\# \text{predicted_cell_type} - \# \text{true_cell_type})}{\# \text{true_cell_types}}$

- Good: the deviation from the true number of cell types is $\leq 20\%$
- Intermediate: the deviation from the true number of cell types is $20\% \leq$
50%

- Poor: the deviation from the true number of cell types is $> 50\%$
2. Variability of the number of cell type estimation:
- Good: the standard deviation is ≤ 2 cell types
 - Intermediate: the standard deviation is $2 \leq 5$ cell types
 - Poor: the standard deviation is > 5 cell types
3. Clustering concordance based on 4 concordance metrics (range from 0 to 1), evaluating the clustering concordance from the predefined cell type labels:
- Good: the average value of metrics score is > 0.7
 - Intermediate: the average value of metrics score is $0.5 \leq 0.7$
 - Poor: the average value of metrics score is ≤ 0.5
4. Speed, summarising of running time of each method:
- Good: the average running time for clustering a single dataset is ≤ 120 s
 - Intermediate: the average running time for clustering a single dataset is $120 \text{ s} \leq 360 \text{ s}$
 - Poor: the average running time for clustering a single dataset is > 360 s
5. Memory and summary of peak memory usage of each method:
- Good: the average peak memory usage for clustering a single dataset is $\leq 4\text{Gb}$
 - Intermediate: the average peak memory usage for clustering a single dataset is $4\text{Gb} \leq 8\text{Gb}$
 - Poor: the average peak memory usage for clustering a single dataset is $> 8\text{Gb}$

2.3 RESULTS

2.3.1 Benchmark framework for clustering algorithms on estimating the number of cell types

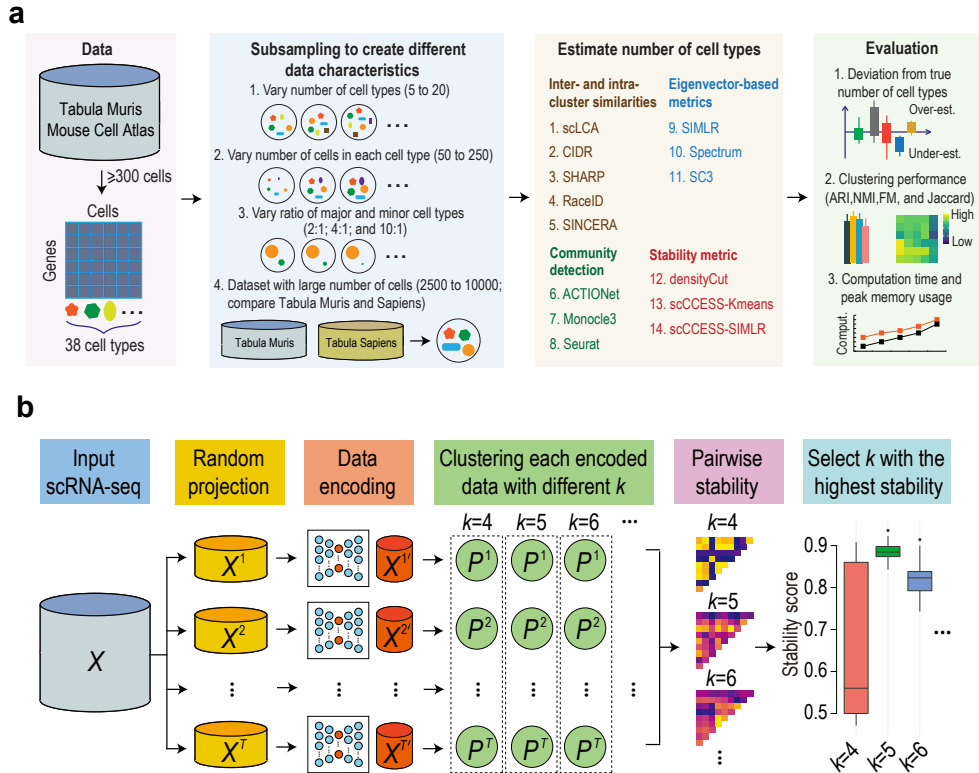


Figure 2.1: Schematic summaries of **(a)** benchmark workflow and **(b)** clustering stability measure. **(a)** Summary of the benchmark workflow. A panel of fourteen scRNA-seq clustering methods that perform the estimation of the number of cell types were evaluated under four main settings for creating different data characteristics via sampling from the Tabula Muris and Tabula Sapiens data. Evaluation includes deviation from the true number of cell types, clustering concordance with predefined cell type labels, and computational time and memory usage of each method. **(b)** Illustration of clustering stability, implemented as part of the single-cell Consensus Clusters of Encoded Subspaces (scCEESS) (Geddes et al., 2019), for estimating the number of cell types in a given scRNA-seq dataset

To evaluate the performance of clustering methods on estimating the number of cell types from data with various characteristics (Figure 2.1a), we subsampled from the Tabula Muris dataset (The Tabula Muris Consortium, 2018) to create three main settings including (i) varying the number of true cell types from

5 to 20 (increment by 1) while fixing the number of cells in each cell type as 200; (ii) varying the number of cells from 50 to 250 (increment by 50) while fixing the number of cell types at 5, 10, 15, and 20; and (iii) varying the ratio of cells between major and minor cell types (i.e. 2:1, 4:1, and 10:1) while fixing the number of cell types at 10 and 20 (see the “Methods” for details). In particular, the number of cells is kept the same among all cell types in setting 1 and 2, whereas in setting 3, the number of cells is different between major and minor cell types. In addition, we also subsampled from both the Tabula Muris and the Tabula Sapiens (The Tabula Sapiens Consortium, 2022) datasets to create a fourth setting in which datasets are with a large number of cells (2500 to 10,000). This last setting allows us to evaluate the performance of clustering methods on datasets with high cell numbers while also assessing if the findings are comparable across different species and data sources.

Among the fourteen methods compared in this benchmark study, we include twelve published clustering methods and two proposed stability-based methods. The two proposed methods use clustering stability extracted from scCCESS (Geddes et al., 2019), an ensemble clustering algorithm, for the number of cell type estimation (Figure 2.1b). Specifically, scCCESS samples multiple random projections from the original input scRNA-seq dataset and encodes the random projections to a lower dimension via autoencoders. Next, it clusters each encoded dataset and creates consensus from these clustering. We take advantage of the multiple clustering output generated from the encoded datasets for evaluating clustering stability of a cluster numbers k by employing scCCESS across a range of k values (by default $k \in [2, K]$; where K is the maximum number of clusters). Intuitively, the number of cell types in the dataset is determined by the k value that leads to the most stable clustering output among all encoded datasets. Thus, we calculate the pairwise agreement score of all clustering output on encoded datasets and select the k that gives the highest average score. Since scCCESS can be used with any clustering algorithm that allows user-specified k values, we coupled scCCESS with a basic k -means clustering algorithm and SIMLR (Wang et al., 2017), a single-cell specific clustering algo-

rithm. We refer to them as scCCESS-Kmeans and scCCESS-SIMLR, respectively, thereafter.

2.3.2 *Overall performance of clustering algorithms on the number of cell type estimation*

We first compare each clustering method for correctly identifying the number of cell types by applying each method on 160 datasets that contain 5 to 20 cell types randomly sampled from the Tabula Muris dataset. The number of cells in each cell type was held constant at 200. Figure 2.2a shows for each method the deviations between the estimated number of cell types and the true number of cell types, with positive deviation representing over-estimation and negative deviation representing under-estimation. Across these datasets (i.e. sampled with 5 to 20 cell types), we found that Monocle3, scLCA, and scCCESS-SIMLR in general have a smaller median deviation compared to other methods, and, as expected, increasing the number of cell types in the dataset leads to higher under- and over-estimation. To summarise these results, we calculated the overall distribution of deviation across all datasets (Figure 2.2b). The summary confirmed the performance of the above methods while highlighting the high instability found in methods such as Specturm, SINCERA, and RaceID, and the bias in underestimation (e.g. SHARP, densityCut) and overestimation (e.g. SC3, ACTIONet, Seurat).

Since a clustering method may correctly estimate the number of cell types in a dataset but still generate poor clustering of the cells, we next assessed the concordance between the clustering output and the predefined cell type labels (obtained from the original publication of Tabula Muris) using four evaluation metrics including Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows index (FM), and Jaccard index (Jaccard). The average clustering concordance and standard deviations across datasets sampled with different numbers of cell types were shown in Figure 2.2c, and the detailed results are presented in Appendix for Chapter 2: Figure S1. The assessment re-

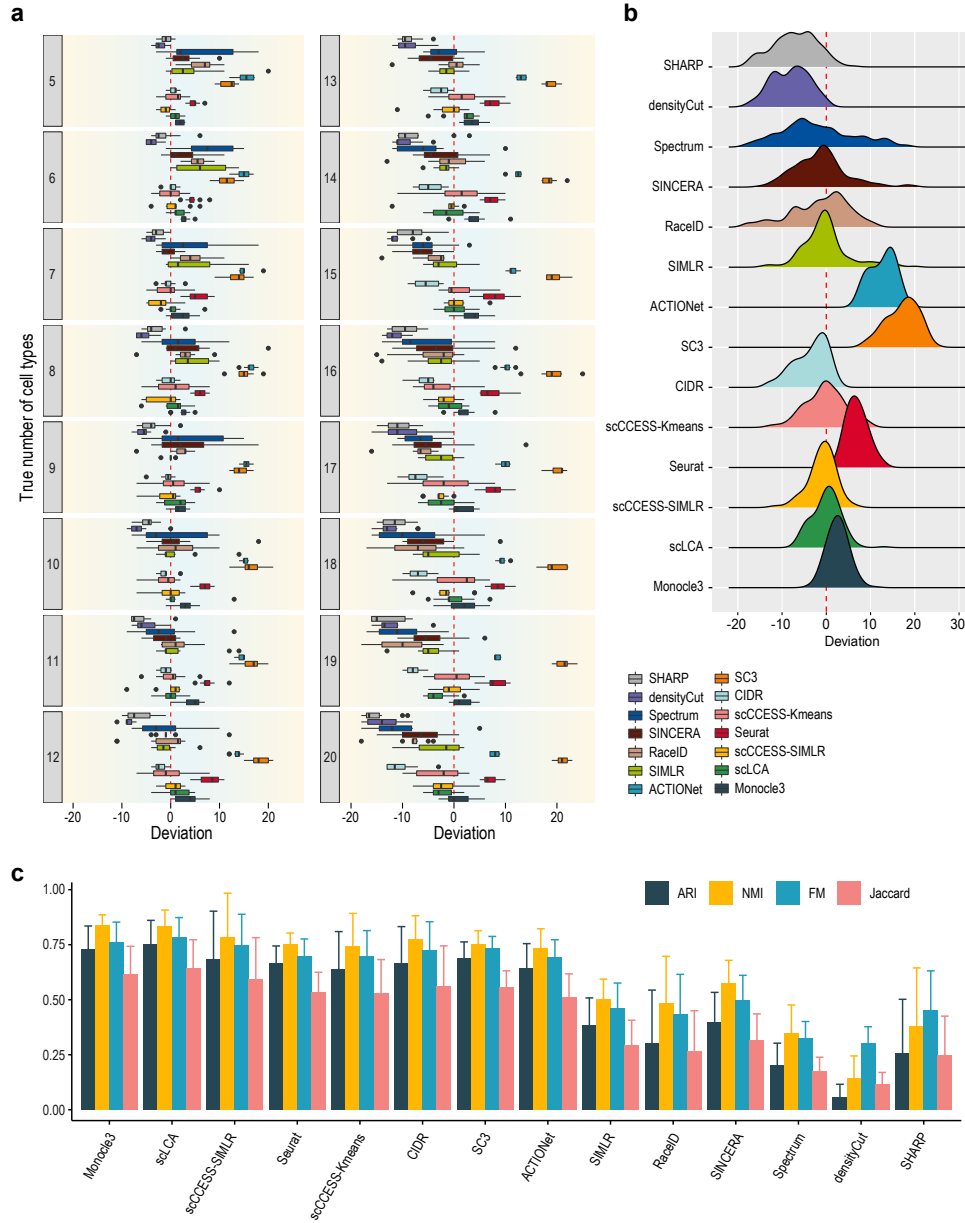


Figure 2.2: Overall performance and impact of cell type numbers on the number of cell type estimation. **(a)** Deviation of the estimated and the true number of cell types for each of the 14 clustering methods. Positive deviation represents over-estimation and negative deviation represents under-estimation. The true number of cell types ranges from 5 to 20 and each was repeated 10 times to capture the estimation variability. **(b)** Density plot summarising overall deviation across all numbers of cell types in (a). **(c)** Concordance of clustering output and predefined cell type labels as quantified by four concordance measures. Each bar represents the average performance across datasets with 5 to 20 cell types, and error bars represent the standard deviation

sults from the four evaluation metrics are highly correlated. Overall, Monocle3, scLCA, scCCESS-SIMLR, CIDR, Seurat, and scCCESS-Kmeans show higher cell type clustering concordance with predefined cell type labels (≥ 0.5), in agreement with their better performance in estimating the number of cell types compared to other methods. Notably, however, higher cell clustering concordance does not necessarily mean a more accurate number of cell type estimation. For example, SC3 has comparatively high cell clustering concordance compared to other best performing methods (e.g. Monocle3) but was significantly over-estimating the number of cell types (Figure 2.2a, b). These results highlight the importance of evaluating the number of cell types estimation accuracy of clustering algorithms independent of their performance on clustering cells.

2.3.3 *Impact of number of cells on number of cell type estimation and clustering*

Besides comparing the performance of clustering algorithms on datasets with a fixed number of cells (i.e. 200) in each cell type, we examine the impact of the number of cells on the number of cell type estimation and the clustering of cells. To this end, we varied the number of cells in each cell type from 50 to 250 (increments by 50 in each test) and assessed the accuracy of the estimated number of cell types when the true number of cell types were set as 5, 10, 15, and 20. We found that, in general, increasing the number of cells helps most clustering algorithms reduce the variability in the number of cell type estimation but have a limited impact on their estimation deviation (Figure 2.3a). However, SC3, ACTIONet, and Seurat are a few exceptions, showing a clear increase in variability in their number of cell type estimation on datasets with larger numbers of cells. A closer look at these results suggests that, interestingly, while most clustering algorithms show a similar level of deviation in the number of cell type estimation, SC3, ACTIONet, and Seurat, to a lesser degree, tend to over-estimate when the number of cells in each cell type increases (Figure 2.3a and Appendix for Chapter 2: Figure S2).

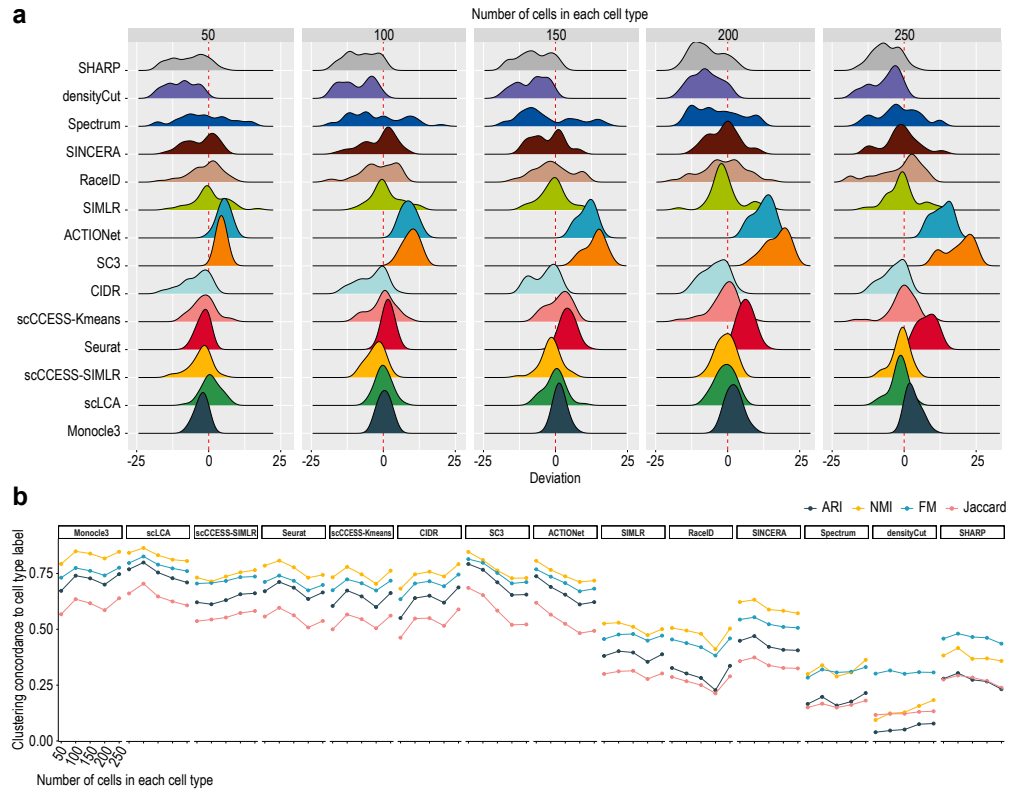


Figure 2.3: Impact of number of cells on the performance of number of cell type estimation. **(a)** Density plot summarising overall deviation in number of cell type estimation across all number of cell types (i.e. 5, 10, 15, and 20) with different number of cells in each cell type. **(b)** Concordance of clustering output and pre-defined cell type labels quantified by four concordance measures for datasets with different number of cells in each cell type (i.e. 50, 100, 150, 200, and 250). Each dot represents the mean value of concordance score

On the clustering of cells, we found that the performance in terms of clustering concordance to the pre-defined cell labels does not necessarily increase with an increasing number of cells in each cell type (Figure 2.3b; Appendix for Chapter 2: Figure S3-S4). In fact, the performance on cell clustering deteriorates for many clustering algorithms when the number of cells increases. This is particularly prominent for SC₃ and ACTIONet, probably due to these methods considerably over-estimate the number of cell types when the number of cells increases in the datasets (Figure 2.3a).

Together, these results suggest that methods such as SC₃, ACTIONet, and Seurat tend to over-estimate the number of cell types, especially when the number of cells is large in each cell type, and also having higher variability in their estimation, unveiling a limitation of these clustering methods when dealing with datasets with a relatively large number of cells.

2.3.4 *Bias analysis of clustering algorithms on estimating the number of cell types*

Given the tendency of consistent under- and over-estimation of the number of cell types we noticed in some clustering algorithms in the benchmarking results, we set out to analyse if there is a systematic bias in the number of cell type estimation for each method and how that is confounded by the number of cell types and the number of cells per cell type in the datasets. Notably, we found while most clustering algorithms tend to underestimate the number of cell types when the true number of cell types increases in the datasets (e.g. densityCut, SHARP, Spectrum, CIDR), SC₃ and Seurat, counter-intuitively, appear to over-estimate when applied to data with larger cell type numbers (Figure 2.4a).

The analysis of the coefficients (slopes) of the linear regression lines fitted to the number of cell types estimated by each method confirms these observations (Figure 2.4b). Monocle3, scCCESS-Kmeans, and scCCESS-SIMLR are the only methods that do not show bias with respect to the changing number of cell types (Figure 2.4b). In addition, while some methods display similar devi-

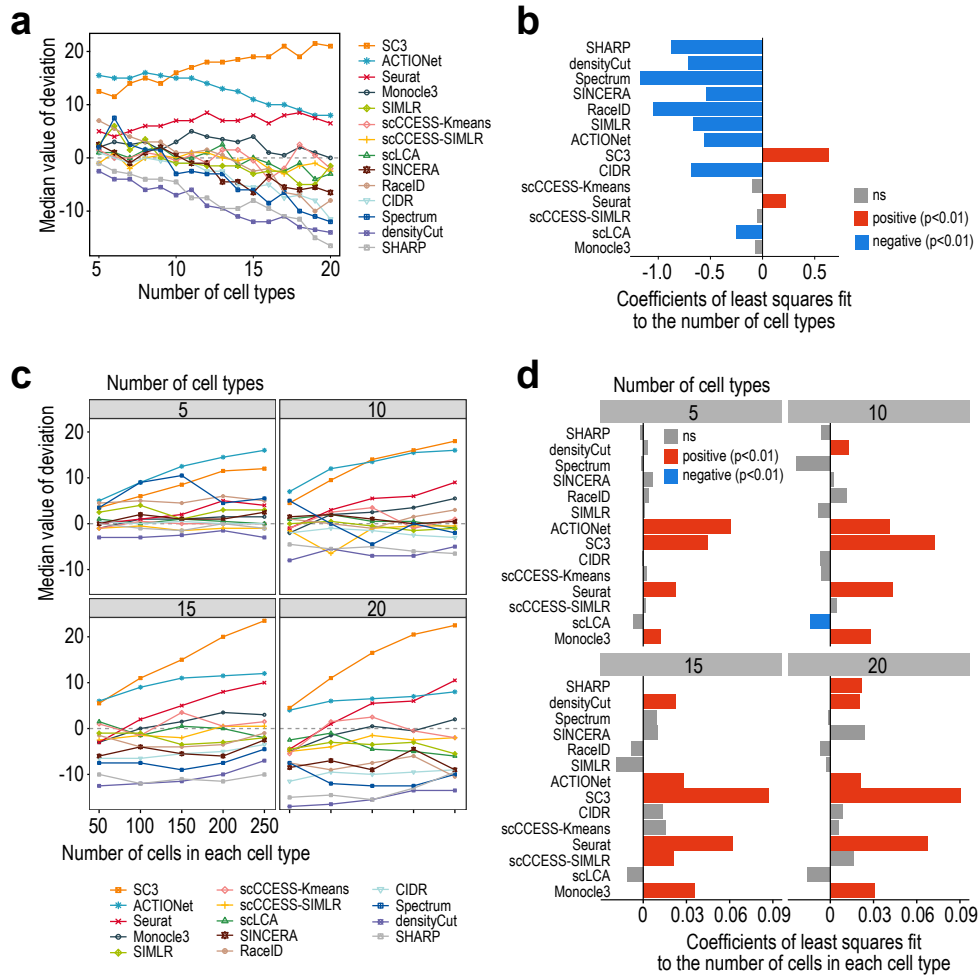


Figure 2.4: Bias analysis for each clustering algorithm on number of cell type estimation. **(a)** Median deviation values of each clustering algorithm on datasets with different number of cell types. **(b)** Coefficients of least squares fit to the median deviation value with respect to the number of cell types. **(c)** Median deviation values of each clustering algorithm on datasets with different number of cells (ranging from 50 to 250) in each cell type. Results are split by the number of cell types (i.e. 5, 10, 15, and 20). **(d)** Coefficients of least squares fit to the median deviation value with respect to the the number of cells in each cell type

ation irrespective of the changes in the number of cells in each cell type (e.g. RaceID, SINCERA, SIMLR), we found that, again, several methods, especially SC3, ACTIONet, and Seurat, increasingly over-estimate cell type number when the number of cells per cell type increases (Figure 2.4c, d). These results highlight the existence of systematic biases in some of the clustering algorithms in the number of cell type estimation, and the biases are compounded by both the number of cell types and the number of cells per cell type in the datasets. In particular, the number of cell types in a dataset could have an opposite effect on their estimation depending on which clustering algorithm was used, and methods such as SC3 and Seurat tend to significantly overestimate the cell type numbers when applied to datasets under two settings (i) with a large number of cell types, and (ii) a large number of cells in each cell type.

2.3.5 *Impact of imbalance of cell ratios on number of cell type estimation and clustering*

While the previous experiments simulated datasets with different numbers of cell types and also different numbers of cells in each cell type, the number of cells among all cell types was kept the same. As such, these settings facilitate the isolation and testing of the performance of clustering algorithms conditioned on these two key aspects (e.g. number of cell types; number of cells) on estimating the number of cell types, most scRNA-seq experiments generate data that capture cell types with different number of cells, sometimes with highly imbalanced ratios. To test the impact of imbalanced ratios of cells among different cell types, we set out subsampling from Tabula Muris data to create major and minor cell types (each with 5 or 10 cell types) with imbalanced cell ratios of 2:1, 4:1, and 10:1 (see the “Methods” section for details).

Figure 2.5a shows the performance on the number of cell type estimation in each imbalanced setting for each of the 14 clustering algorithms, and Figure 2.5b quantifies the concordance of cell clustering to cell type labels averaged across the results from 10 and 20 cell types. We found that the imbalanced ratio

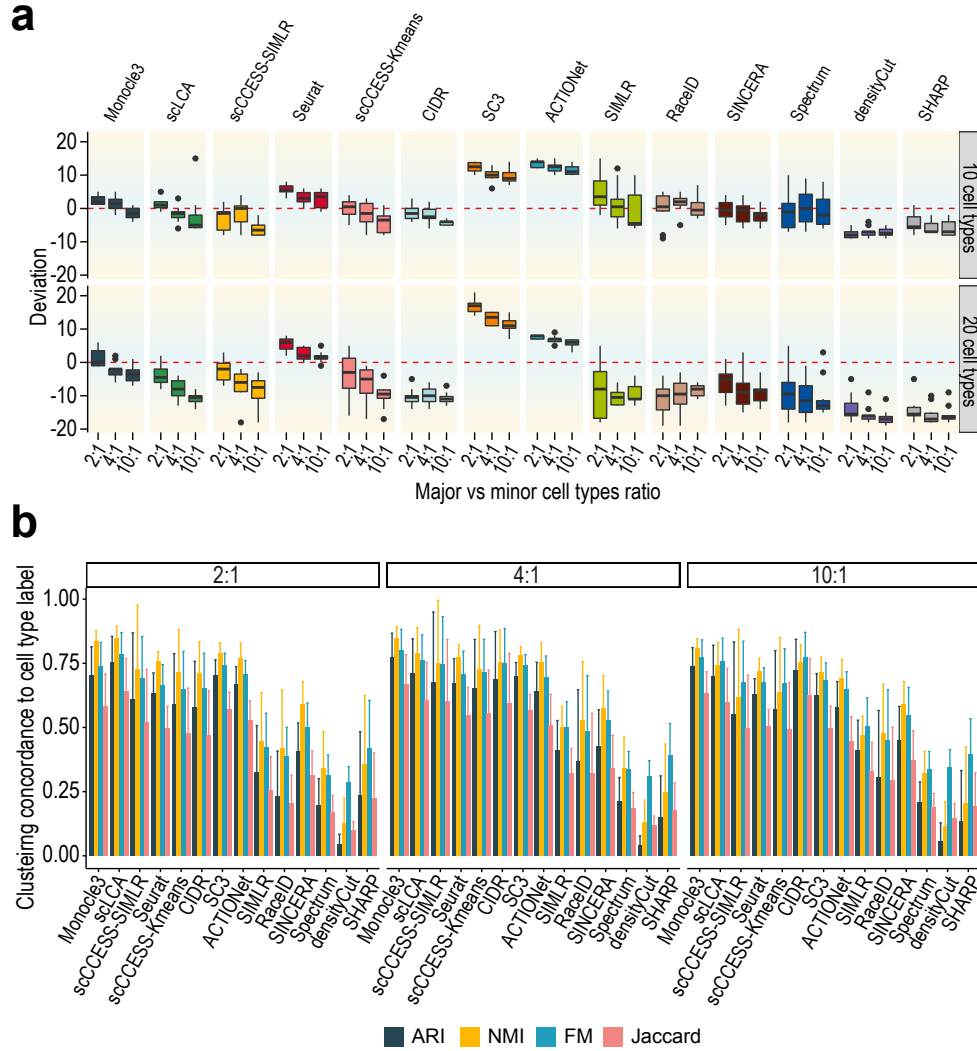


Figure 2.5: Impact of imbalance ratios of cells among cell types on the performance of the number of cell type estimation. **(a)** Deviation of the estimated and the true number of cell types for each clustering method on datasets with different imbalance ratios (i.e. 2:1, 4:1, and 10:1) and different number of cell types (i.e. 10 and 20). Each combination was repeated 10 times for estimating variability. **(b)** Concordance of clustering output and predefined cell type labels as quantified by four concordance measures for datasets with different imbalance ratios of 2:1, 4:1, and 10:1, and error bars represent the standard deviation

of cells in major and minor cell types in general leads to a reduction in the estimated number of cell types when the ratio increases (Figure 2.5a). This unintended reduction helps methods that tend to over-estimate the number of cell types (e.g. Seurat, SC3, ACTIONet) but results in under-estimation for others (e.g. scLCA, scCCESS-Kmeans, scCCESS-SIMLR). Again, the performance on cell clustering does not always match the accuracy of the number of cell type estimation (e.g. Seurat) (Figure 2.5b and Appendix for Chapter 2: Figure S5). These results suggest that the imbalance ratio of cells in major and minor cell types has an uneven impact on different clustering algorithms and the number of cell types (e.g. 10 and 20) in a dataset tends to be the key driver on their performance of the number of cell type estimation.

2.3.6 Cross-comparison on datasets with high cell numbers using *Tabula Muris* and *Tabula Sapiens* data

We further cross-compared the performance of different clustering algorithms on datasets with a large number of cells using both *Tabula Muris* and *Tabula Sapiens* data. Again, we observed clear overestimations of the number of cell types from SC3, ACTIONet, and Seurat (Figure 2.6a) on these datasets, confirming the results from the bias analysis using datasets with increasing numbers of cells (Figure 2.4c, d). When the overestimations are extreme, the performance of cell clustering from these methods do suffer (Figure 2.6b and Appendix for Chapter 2: Figure S6). Nevertheless, methods that do not show large over or underestimation of number of cell types do not necessary perform better than those that significantly overestimate. The most striking examples include densityCut, Spectrum, and RaceID. Finally, with few exceptions (e.g. SINCERA), the performance of different clustering algorithms on *Tabula Muris* and *Tabula Sapiens* datasets are highly consistent across the number of cell types and number of cells tested in our experiments (Figure 2.6 and Appendix for Chapter 2: Figure S6). These results confirm that the performance assessments are generalisable across datasets from different species and sources.

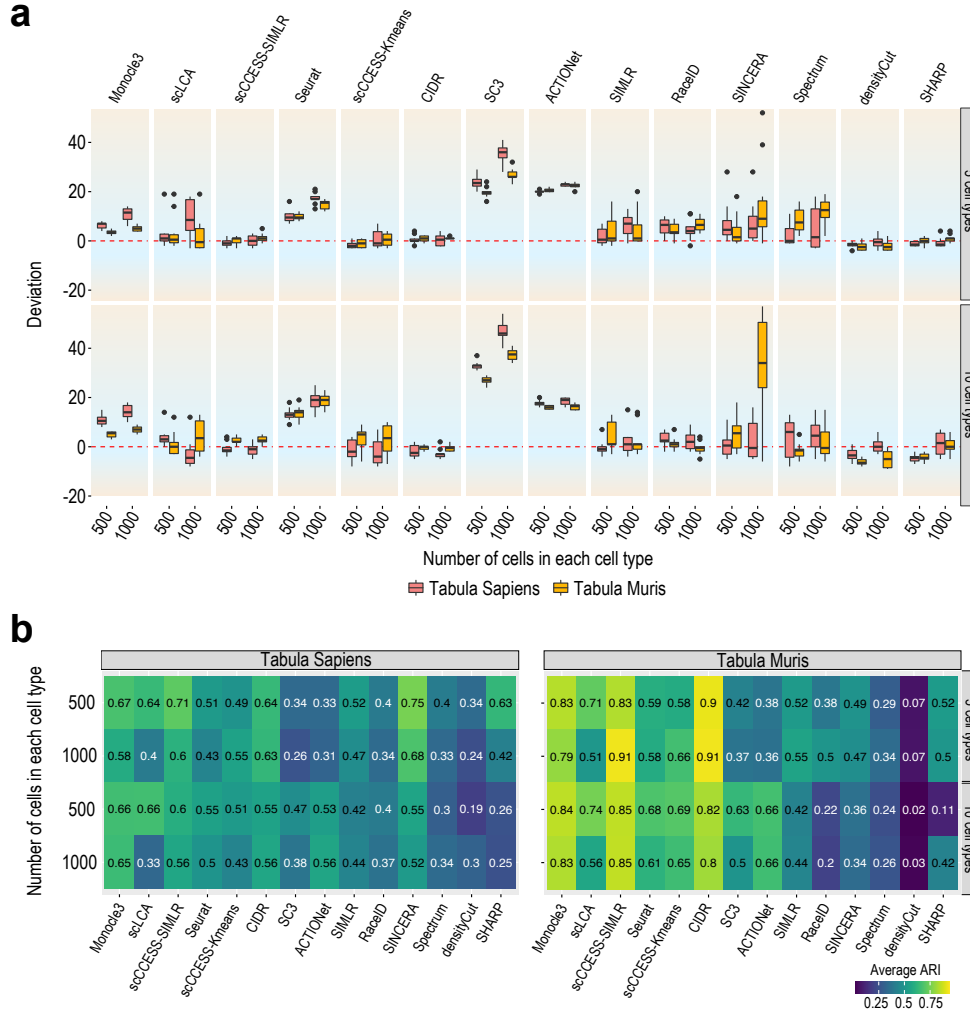


Figure 2.6: Performance of clustering methods on datasets with large numbers of cells sampled from Tabula Sapiens and Tabula Muris. **(a)** Deviation of the estimated and the true number of cell types for each clustering method on datasets with different number of cells (i.e. 2500 and 10000) depending on the number of cell types (i.e. 5 and 10) and the number of cells in each cell type. Each combination was repeated 10 times for estimating variability. **(b)** Concordance of clustering output and predefined cell type labels as quantified by ARI for each method

2.3.7 Running time and peak memory usage

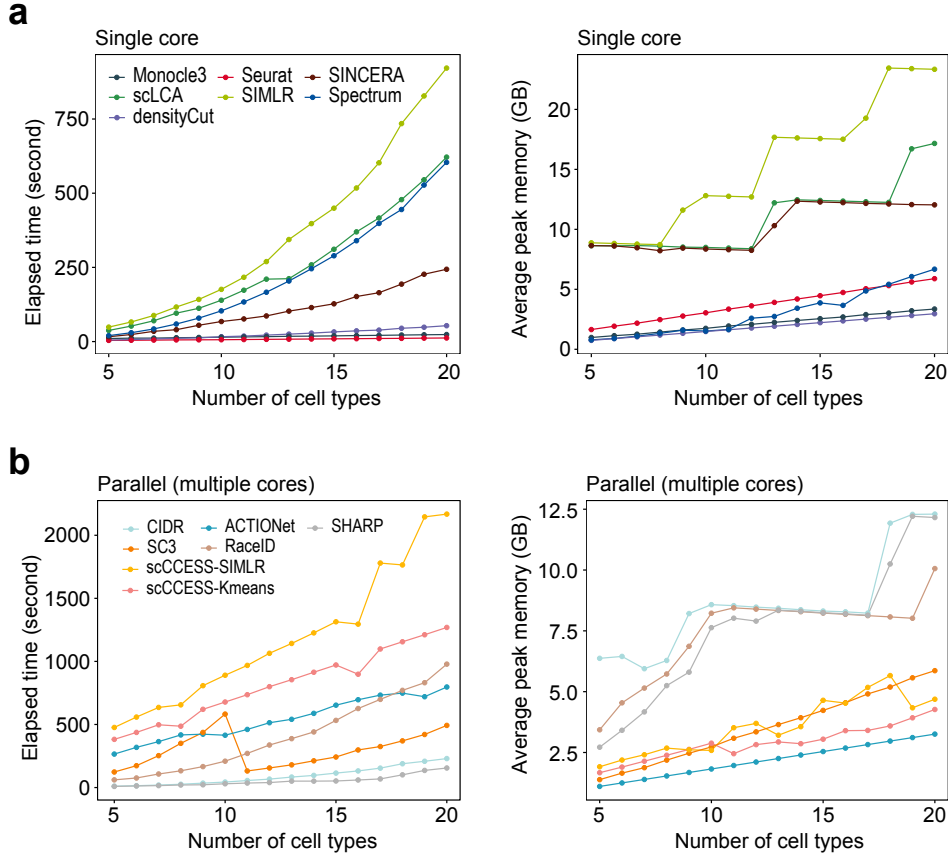


Figure 2.7: Benchmark of the elapsed time and peak memory usage across datasets with different number of cell types, ranging from 5 to 20. **(a)** The running time and peak memory usage of methods that uses only a single CPU core (i.e. densityCut, scLCA, SIMLR, Monocle3, Seurat, Spectrum, and SINCERA). **(b)** The running time and peak memory usage of methods that employ multiple cores for parallel computing (i.e. ACTIONet, RaceID, scCCESS-Kmeans, SHARP, CIDR, SC3 and scCCESS-SIMLR)

Lastly, we benchmarked the computational time and peak memory usage for all clustering methods in each of the four settings (i) varying the number of cell types (Figure 2.7), (ii) varying the number of cells in each cell type (Appendix for Chapter 2: Figure S7); (iii) varying the imbalanced ratio of cells among cell types (Appendix for Chapter 2: Figure S8), and (iv) datasets with large number of cells (Appendix for Chapter 2: Figure S9). Seven out of 14 methods (densityCut, scLCA, SIMLR, Monocle3, Seurat, Spectrum, and SINCERA) use only a single thread to perform clustering analysis, whereas the remainders

run on parallel computing mode that utilise multiple CPU cores, if available, by default. In general, clustering on large datasets needs more computing time and uses more memory in all settings (e.g. increasing number of cell types, increasing number of cells), and, interestingly, several single thread methods take shorter computing times than most parallel methods, and on average, more peak memory usage than parallel methods (Figure 2.7).

Specifically, on computational time, three single thread methods (Seurat, Monocle3, densityCut) and two parallel methods (CIDR, SHARP) greatly outperform others, where the clustering on any benchmark datasets were completed in less than 200 s. scLCA is one of the top-performing methods on the number of cell type estimation but uses significantly more time compared to other faster methods, revealing a trade-off between estimation accuracy and computational efficiency for this method. As expected, ensemble-based clustering methods (i.e. scCCESS-Kmeans and scCCESS-SIMLR) take a longer time to complete, since multiple clustering needs to be performed and then combined to get a consensus output. Nevertheless, when dealing with datasets with a large number of cells, RaceID appears to be significantly slower than all other methods (Appendix for Chapter 2: Figure S8).

On memory usage, SIMLR is the most memory-consuming method compared to others and may prohibit its application to large-scale datasets (Figure 2.7 and Appendix for Chapter 2: Figure S8). However, scCCESS-SIMLR reduces the feature dimension of the input data and therefore significantly reduces the peak memory usage. CIDR is reputable for its ultrafast clustering time, but uses the highest amount of peak memory among all parallel methods, highlighting a trade-off between memory usage and computational efficiency for this method.

2.4 DISCUSSION

Overall, we observed that methods based on community detection and clustering stability performed more favourably than methods from other categories across most of the evaluation criteria (Figure 2.8). Eigenvector-based meth-

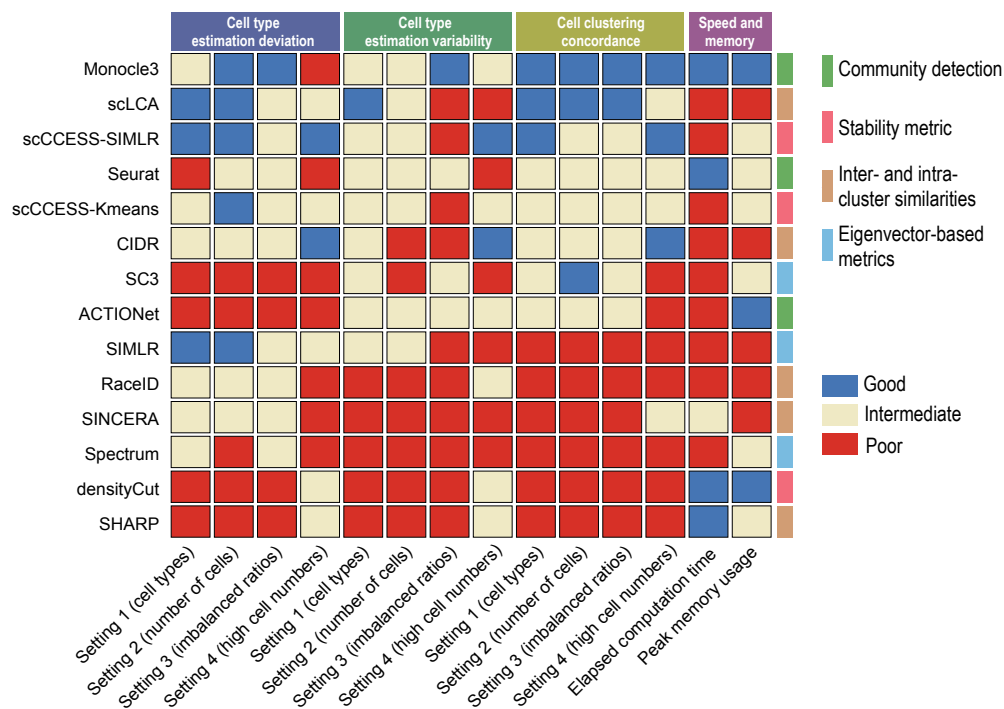


Figure 2.8: Summary of clustering method performance across all major evaluation criteria. Criteria and cutoff values for performance categories are described in detail in the Method section. Clustering methods are ranked by their overall performance across all criteria

ods, in comparison, performed unfavourably in general and methods based on inter- and intra-cluster similarities show a very broad range of performance from those that performed very well (i.e. scLCA) to those poorly (i.e. SHARP and RaceID). These findings are largely consistent with a recent study reporting that a stability-based clustering method performed the best while methods such as SC₃ overestimates the number of cell types in many cases and RaceID takes more computation time than alternatives such as Seurat (Peyvandipour et al., 2020). While it is hard to pinpoint the factors contributing to the performance difference among different categories of clustering algorithms, we suspect that methods based on community detection and clustering stability may share a similar implementation strategy whereas the implementation strategies for methods based on inter- and intra-cluster similarities are more diverse and hence may have contributed to the wider range of performance. Having said that, these results do indicate that, while there may be a general trend in performance for certain categories of methods, the specific implementation of each method also plays a significant role in determining its performance in each of these evaluated aspects. Furthermore, these results also demonstrate that, while there is an overall concordance in performance across most of the evaluation criteria, each criterion does shed a unique light on a specific aspect of each clustering algorithm. Most importantly, we found that better performance on cell clustering does not necessarily imply accuracy in estimating the number of cell types. This could happen when a method is able to correctly estimate the number of cell types even when its clustering of cells to their respective groups is less precise. These results highlight the importance of evaluating the number of cell types separately from clustering accuracy, as these two tasks address different aspects of cell type detection in scRNA-seq data. Clustering groups cells based on similarity, often relying on thresholds or assumptions that may not accurately capture the true number of cell types. On the other hand, estimating the number of cell types (k) is a quantitative assessment that provides cell type number information for downstream analysis. Therefore, separating the evaluation of k from the evaluation of clustering precision ensures a more accurate

and objective assessment of a clustering method's ability to detect and define cellular diversity.

Although the sampling framework we used in this study creates scRNA-seq data with discrete cell types from the Tabula Muris resource, cell type structures from many biological systems are hierarchical with subpopulations of cells residing in each major cell type (Lin et al., 2020). Related to this, the improving capability of scRNA-seq for profiling complex tissues and organs has led to datasets with increasing numbers of cell types and potentially further compounded by hierarchical cell type relationships. Thus, developing frameworks that are capable of benchmarking multi-resolution or multi-scale clustering on datasets with large numbers of cell types is a critical future research direction. Another recent expansion in single-cell omics field is the increasing availability of multimodal single-cell omics data produced by new enabling sequencing technologies such as single-cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq) (Cusanovich et al., 2015) and cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) (Stoeckius et al., 2017) among others. Therefore, integrative methods, such as multi-view clustering (Yang and Wang, 2018; Cao et al., 2020), that use several omics types generated from the same set of samples have the potential to reveal information, including hierarchical cell type structure, that are not accessible by examining only a single data type. Evaluating clustering algorithms for their performance on estimating the number of cell types from multimodal datasets is challenging and requires further methodological innovation.

2.5 CONCLUSIONS

Clustering is an essential technique for scRNA-seq data analysis. While a tremendous amount of work has been done for designing and evaluating algorithms for clustering cells into cell types, their performance on estimating the number of cell types from a scRNA-seq dataset is yet to be systematically assessed. In this study, we have benchmarked 14 scRNA-seq clustering methods on esti-

inating the number of cell types in datasets with various characteristics. We have also assessed other related aspects of performance including cell clustering and their scalability in terms of running time and memory usage. We compiled these multi-faceted evaluations into a recommendation (Figure 2.8), summarising the performance of each of the methods according to each evaluation criteria. We expect this work will foster future development of scRNA-seq data clustering methods by providing a reference to the performance on the estimation of number of cell types in such data.

ENSEMBLE DEEP LEARNING OF EMBEDDINGS FOR CLUSTERING MULTIMODAL SINGLE-CELL OMICS DATA

The availability of multimodal single-cell omics technologies - the 2019 Nature Method of the Year - allows multiple molecular profiles to be acquired from each individual cell, generating multidimensional data with the potential to significantly improve the molecular profiling of each cell. For example, the complementary information present in different omics layers could be used to improve cell type annotation using unsupervised clustering, and existing methods tailored for clustering scRNA-seq data (see Chapter 2) can be applied by treating each modality independently and identifying consensus clusters. However, it is unclear whether these strategies are most effective when using complementary information from multiple levels. With this in mind, emerging approaches to analysing such multimodal data integrate different layers (modalities) into a unified latent space. Cells can then be clustered in this space based on their embedded representations. The ability to generate integrated embeddings that capture information from multiple modalities will contribute to a better understanding of cellular landscapes.

Despite the potential of multimodal single-cell omics technologies, the quality of the data modality integration can have a substantial impact on the effectiveness and efficiency of clustering algorithms. Using an ensemble method to generate multiple views of integrated data modalities has been found to improve the performance of cell clustering (Geddes et al., 2019). Nonetheless, ensemble techniques are generally more time-consuming. This highlights the need for innovative methods that can not only integrate multimodal data more effectively but also do so in a way that is computationally accessible for large-scale single-cell studies.

In response to these challenges, we introduce SnapCCESS, an ensemble deep learning framework designed to generate multi-view integrated embeddings from multimodal single-cell omics data, facilitating consensus clustering of single cells. Drawing inspiration from our previous ensemble deep learning in scRNA-seq data analysis (Geddes et al., 2019), SnapCCESS extends this approach to multimodal data, integrating multiple molecular features into a cohesive latent space. A key innovation of SnapCCESS is its "snapshot" ensemble learning process, which markedly enhances computational efficiency without compromising the quality of clustering. The contents of this chapter are as published in *Bioinformatics* in 2023 (Yu et al., 2023). As the first author of the manuscript, my contribution towards this project involved leading the method development alongside A/Prof. Pengyi Yang, Dr. Chunlei Liu and Prof. Jean Yang. Additionally, I curated the benchmarking data, implemented all data analytics, and developed the R and Python package available on Github (<https://github.com/PYangLab/SnapCCESS>).

3.1 INTRODUCTION

The development of novel single-cell technologies, such as cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) (Stoeckius et al., 2017), shared single-cell profiling of RNA and chromatin (SHARE-seq) (Ma et al., 2020) and trimodal single-cell profiling by TEA-seq (Swanson et al., 2021), enables the profiling of gene expression, protein abundance, and/or chromatin accessibility in the same cell. The availability of multiple data modalities in individual cells promises more precise characterisation of cells such as clustering cells into distinctive cell types (Zhu et al., 2020). To analyse such multimodal single-cell omics data, however, requires effective computational methods that are capable of integrating data modalities for extracting the underlying biological signals. By "effective," we refer to methods that are computationally efficient, scalable, and runnable on typical datasets produced by multimodal sequencing technologies. By "capable," we mean methods that can process and integrate diverse types of single-cell data beyond scRNA-seq alone, which remains a lim-

itation of many existing tools. These standards ensure that the methods can handle the complexity and scale of multimodal data while providing meaningful biological insights.

While various methods exist for enabling the clustering of multimodal single-cell omics data, such as by simple feature concatenation or more sophisticated methods that integrate clustering output from each modality (Miao et al., 2021; Adossa et al., 2021) or construct similarity graphs for joint clustering (Wu et al., 2022), a popular approach is to integrate data modalities through learning an embedding that encodes multiple data modalities into a shared latent space, from which any clustering algorithm that accepts the embedding as input could be used for clustering cells (Lin et al., 2022; Liu et al., 2023). These embeddings capture shared features between modalities, summarizing common biological signals while reducing noise specific to each modality. Unlike simpler methods, such as feature concatenation, embeddings avoid overemphasizing noisy or redundant features and are better at uncovering patterns shared across modalities.

For example, Jvis-learn performs joint dimension reduction of data modalities for generating embeddings of multimodal single-cell omics data (Do and Canzar, 2021). It automatically determines the relative importance of each data modality that emphasises distinguishing characteristics while reducing noise. MOFA+ implements a Bayesian group factor analysis framework to infer a low-dimensional embedding that captures shared variation across multiple modalities (Argelaguet et al., 2020). Recently, deep learning-based methods such as totalVI (Gayoso et al., 2021) use a variational autoencoder (VAE) to learn an embedding for integrating RNA and ADT modalities such as in CITE-seq data, and MultiVI (Ashuach et al., 2023) uses a VAE to integrate RNA and ATAC modalities such as in SHARE-seq data. Deep learning models, such as VAEs, are effective for generating integrated embeddings because they learn latent spaces that capture meaningful biological variation while removing technical noise. This allows the embeddings to represent both shared and unique signals from different data modalities. As a result, they provide insights into cell

type-specific or condition-specific relationships. Compared to simpler methods, these embeddings can better balance contributions from each modality, reducing the risk of overrepresenting noisy or redundant features.

The multimodality integrated embeddings generated from these methods can be subsequently applied for cell clustering using any clustering algorithms that accept embeddings as input. Thus, the utility and quality of the multimodality integrated embeddings will have a large impact on clustering multimodal single-cell omics data.

Here we aim to improve the clustering of multimodal single-cell omics data by developing an ensemble deep learning-based framework that generates a multi-view of multimodality integrated embeddings. This is motivated by our previous work on autoencoder-based ensemble clustering of scRNA-seq data that demonstrates consensus derived from multiple embeddings each generated from perturbing the input data can lead to significantly better clustering results (Geddes et al., 2019). Such a single-cell consensus clusters of encoded subspaces (scCCESS) approach benefits from the multi-view of the input data (Cao et al., 2020) and is generic to clustering algorithms. Built on this concept, we propose SnapCCESS, an ensemble clustering framework that uses VAE and the snapshot ensemble learning technique (Huang et al., 2017) to learn multiple embeddings each encoding multiple data modalities, and subsequently generate consensus clusters for multimodal single-cell omics data by combining clusters from each embedding. The innovation in SnapCCESS includes (i) implementing an ensemble deep learning framework for creating a multi-view of latent spaces from which multimodality embeddings of multimodal single-cell omics data can be generated, and (ii) designing a snapshot ensemble learning approach to significantly improve the computational efficiency of the proposed framework.

By applying SnapCCESS to multimodal single-cell omics datasets generated by various biotechnologies and protocols, we show that SnapCCESS is effective and computationally efficient in learning multiple embeddings compared to conventional ensemble deep learning methods. We found that multimodal

data generally offer more information than single modality alone and SnapCCESS leverages such information to improve cell clustering. We also show that embeddings learned from SnapCCESS are generalisable and can be coupled with various clustering algorithms for improving consensus clustering of multimodal single-cell omics data. Lastly, we demonstrate the competitive performance of SnapCCESS with the other state-of-the-art methods for generating embeddings of data modalities for clustering cells. Together, our work showcases the effectiveness of a novel unsupervised ensemble deep learning framework for performing clustering analysis of multimodal single-cell omics data.

3.2 METHODS

3.2.1 *SnapCCESS framework for generating embeddings of multimodal single-cell data*

To integrate the high-dimensional feature space in each modality of multimodal single-cell omics data, SnapCCESS encodes features from multiple data modalities into a latent space using the VAE component of our recently published Matilda framework (Liu et al., 2023) to jointly learn to reconstruct each data modality (Figure 3.1a).

Specifically, SnapCCESS first learns different data modalities using modality-specific encoders and decoders (denoted using different colours in Figure 3.1a). The encoders in the VAE component include one learnable point-wise parameters layer and one fully connected layer to the input layer. Because surface protein modality has significantly fewer features than RNA and ATAC modalities, we empirically set the numbers of neurons for encoders of RNA, surface proteins, and ATAC modalities to be 185, 30, and 185, respectively. To learn a latent space that integrates the information across modalities, SnapCCESS concatenates the output from the encoder trained from each data modality to perform joint learning using a fully connected layer with 100 neurons. The

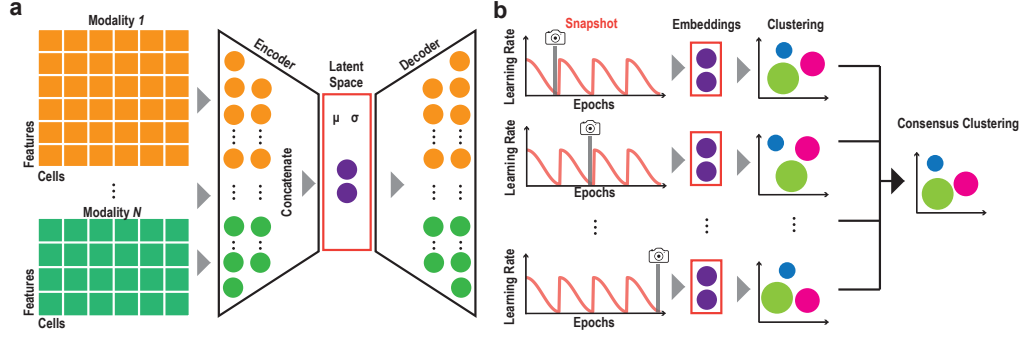


Figure 3.1: The proposed SnapCCESS framework of ensemble deep learning of embeddings for multimodal single-cell data clustering. **(a)** A variational autoencoder (VAE) is used to encode the high-dimensional features from multimodal data to a low-dimensional latent space. **(b)** The training process of SnapCCESS is based on the snapshot ensemble deep-learning model using learning rate annealing cycles where the model converges to and then escapes from multiple local minima, and multiple snapshots were taken at these minima for creating a multi-view of embeddings. The schematic illustrates using epoch of 10 for generating snapshots. The consensus clustering is derived from combining individual clustering results each from a snapshot embedding.

embeddings of input data were obtained from the latent space of VAE by minimising the loss function L defined as follows:

$$L = \sum_{m=1}^M \left\| \left(X_m - \hat{X}_m \right) \right\|^2 + \text{KL}[N(\mu_{x_m}, \sigma_{x_m}), N(0, 1)]$$

where x represents the original input, $\hat{x}$ represents the reconstructed data, i is the i^{th} modality, M is the number of modalities in a multimodal data. $N(\cdot)$ presents a normal distribution, which is learnable in VAE.

SnapCCESS employs the snapshot technique for ensemble learning (Huang et al., 2017). The key idea behind snapshot ensemble learning is to train multiple versions of a single model by using a cyclic learning rate scheduler. Since the ensemble is formed from a single training process, snapshot ensemble learning is significantly faster while maintaining a similar performance when compared to conventional ensemble methods that train individual models from multiple training processes (see [Huang et al., 2017] for more comparative analyses). In SnapCCESS, the learning rate of the training model was set up to be the shifted cosine function, which could help the training model converge to multiple lo-

cal minima and then get multiple lower-dimensional embeddings (Figure 3.1b). This is defined as follows:

$$S(t) = \frac{S_0}{2} \left(\cos \left(\frac{\pi \text{mod}(t-1, \text{ceil}(\frac{T}{E}))}{\text{ceil}(\frac{T}{E})} \right) + 1 \right)$$

where S_0 is the initial learning rate, t is the iteration number, T is the total number of training iterations, E is the number of learning rate cycles, $\text{mod}(\cdot)$ refers to the modulo operator, and $\text{ceil}(\cdot)$ refers to the smallest integer greater than or equal to a given value.

3.2.2 Clustering algorithms for ensemble clustering of embeddings

To create consensus clustering, embeddings generated from SnapCESS and conventional VAE ensemble method were used for generating clustering results and then combined for deriving consensus results. Since the embeddings generated by these methods can be coupled with various clustering algorithms for creating consensus clustering results, we have included three different clustering algorithms for testing their effectiveness. These include a simple k-means clustering algorithm, a more sophisticated spectral clustering method, and SIMLR, a kernel-based clustering method designed for scRNA-seq data analysis (Wang et al., 2017). In particular, for the simple k-means clustering, we utilised the `kmeans` function in the `stats` package with the default settings. For the spectral clustering algorithm, we employed the `spectralClustering` function from the `CiteFuse` package (Kim et al., 2020) with the default parameters to perform spectral clustering. Lastly, we used the `SIMLR_Large_Scale` function in the `SIMLR` package with the number of principal components set to 20 as recommended. For all clustering algorithms, the number of clusters was set to be the same as the number of cell types in each dataset based on the cell type annotation from the original study. After obtaining individual clustering output P_t from each clustering method, a fixed-point iteration algorithm for obtaining hard least squares Euclidean consensus partitions was applied to compute the consensus

partition P_E of individual partitions using clue package `cl_consensus` function (Hornik, 2005):

$$\arg \min_{P_E \in \mathcal{P}} \sum_{t=1}^T w_b D(P_t, P_E)^2$$

where w_b denotes the weight assigned to each individual clustering output (set to 1 in our case), and $D(P_t, P_E)^2$ is the squared Euclidean distance used to measure the difference between an individual partition P_t and the consensus partition P_E . All clustering analyses were carried out in the R programming environment.

3.2.3 Evaluation settings and datasets

3.2.3.1 Conventional VAE ensemble

Conventional ensemble deep learning typically relies on perturbing the initialisation and/or input data to train individual base models that can be used for creating the final ensemble model (Cao et al., 2020). To compare the performance of SnapCCESS with the conventional VAE ensemble method, we implemented a conventional VAE ensemble learning model by using random initialisation of the VAE neural networks. For a fair comparison, we used the same VAE model as in SnapCCESS to learn the latent space for integrating the data modalities in multimodal single-cell omics data. The same numbers of neurons for encoders and decoders were used as in SnapCCESS and the learning rate was fixed at 0.02. To obtain multiple embeddings, multiple VAEs were trained each with a different set of network initialisation weights on the input dataset.

3.2.3.2 Settings for other multimodal embedding generation methods

MOFA+. MOFA2 (v1.6.0) which implements MOFA+ (Argelaguet et al., 2020) was used to generate multi-modality embeddings of the six datasets. Following the author's tutorial (https://raw.githubusercontent.com/bioFAM/MOFA2_tutorials/master/R_tutorials/getting_started_R.html), pre-normalised datasets were

first used to create the mofa object using `create_mofa` function. The data were then used as input for the modal training using `run\_mofa` function with default parameters. We use the `get\_factors` function with `factors = 'all'` to obtain the embeddings for each input dataset.

Jvis-learn. Jvis-learn (v0.0.12) (Do and Canzar, 2021) was employed for generating multi-modality embeddings of the six datasets. The pre-normalised datasets were used as input for creating the j-SNE embeddings that joint multimodal omics data via the JTSNE function.

totalVI. totalVI is designed for analysing CITE-seq data (Gayoso et al., 2021). In this study, the totalVI procedure implemented in the scvi-tools package (v0.17.3) was used for generating multi-modality embeddings of the three CITE-seq datasets. Following the author's tutorial (<https://docs.scvi-tools.org/en/stable/tutorials/notebooks/totalVI.html>), the raw count matrices of RNA and ADT were first normalised using the `normalize_total` and `log1p` functions and then top 4,000 most variable genes were selected using the `highly\_variable\_genes` function. The data were input for model training using `scvi.model.TOTALVI.setup_anndata`, `scvi.model.TOTALVI`, and `train` functions in scvi-tools. The latent space of RNA and ADT modalities was generated using the `get_latent_representation` function.

MultiVI. MultiVI is a sibling of totalVI and specifically designed for analysing data with RNA and ATAC modalities (Ashuach et al., 2023). Here, the MultiVI procedure implemented in the scvi-tools (v0.17.3) was used for multi-modality integration of the SHARE-seq and SNARE-seq datasets. In accordance with the author's tutorial (https://docs.scvi-tools.org/en/stable/tutorials/notebooks/MultiVI_tutorial.html), the raw count matrices of RNA and gene activity score matrices from ATAC and the paired matrix of RNA and ATAC were utilised as input. These data were first concatenated using the `organize_multiome_anndatas` function in scvi-tools and then used for model training using `scvi.model.MULTIVI.setup_anndata`, `scvi.model.MULTIVI` and `train` functions in scvi-tools. The latent space of RNA and ATAC modalities was generated using the `get_latent_representation` function.

3.2.3.3 Datasets and pre-processing

Ramaswamy CITE-seq dataset (Ramaswamy et al., 2021). The raw RNA and ADT matrices of PBMC from three healthy donors were downloaded from NCBI GEO using the accession number GSE166489. We used the healthy donor (GSM-5073072) in our analysis. After filtering RNA and ADT expressed in less than 1% of the cells and genes, discarding cell types with fewer than 50 cells, we obtained 9,745 cells and 21 cell types, with 11039 RNA and 189 ADT features.

Stephenson CITE-seq dataset (Stephenson et al., 2021). The PBMC CITE-seq data of healthy individuals sequenced by NCL medical centre was used in this study. The raw matrices of RNA and ADT and the annotation of cells to their respective cell types from the original study were downloaded from the EMBL-EBI ArrayExpress database under the accession number E-MTAB-10026. RNA and ADT in this dataset were filtered by removing those that expressed in less than 1% of the cells and genes, cell types were filtered by removing those that have less than 50 cells. After filtering, 64,197 cells from 15 cell types (4999 RNA, 192 ADT) were kept for analysis.

Hao CITE-seq dataset (Hao et al., 2021). The raw RNA and ADT matrices from this CITE-seq dataset were downloaded from NCBI GEO under the accession number GSE164378. As the above, RNA and ADT in this dataset were filtered by removing those that expressed in less than 1% of the cells and genes, and cell types were filtered by removing those that have less than 50 cells. In total, 67,035 cells (11,451 RNA, 228 ADT) and 29 cell types in batch 1 of the dataset were used in the analysis.

Chen SNARE-seq dataset (Chen et al., 2019a). The SNARE-seq data that measures RNA and ATAC from matched cells in the adult mouse brain cortex sample (AdBrainCortex) was downloaded from NCBI GEO under the accession number GSE126074. The cell type information was obtained from the authors. For ATAC data, peaks with no expression across cells were removed. We then summarised the ATAC data from peak level into gene activity scores using the `CreateGeneActivityMatrix` function in Seurat. We filtered out RNA and ATAC

quantified in fewer than 1% of the cells and genes, and removed cell types that have less than 50 cells, resulting in a dataset with 9,930 cells (11,011 RNA, 16,443 ATAC peak features) and 20 cell types for the subsequent analyses.

Ma SHARE-seq dataset (Ma et al., 2020). The SHARE-seq data that measures RNA and ATAC from matched cells in mouse skin samples were downloaded from NCBI GEO under the accession number GSE140203. Similar to the above, we first removed peaks with no expression across cells, and then summarised the ATAC data from peak level into gene activity scores using the `CreateGeneActivityMatrix` function in Seurat. We filtered out RNA and ATAC quantified in fewer than 1% of the cells and genes, and remove cell types that have less than 50 cells, resulting in a dataset with 32,968 cells (8,765 RNA, 17,413 ATAC peak features) and 23 cell types for the subsequent analyses.

Swanson TEA-seq dataset (Swanson et al., 2021). TEA-seq enables simultaneous single-cell profiling of transcripts, epitopes, and chromatin accessibility. The processed matrices of TEA-seq data from measuring PBMC were downloaded from the NCBI Gene Expression Omnibus (GEO) under the accession number GSE158013, with raw RNA expression, ADT expression, and peak accessibility (ATAC) measured for the same cells in four data batches. Due to the low batch effect presented in the four datasets, we merged the four data batches. We summarised the matrix of ATAC from peak level to gene activity scores using the `CreateGeneActivityMatrix` function in the Seurat package. Genes with fewer than 1% quantifications across all cells and all genes in the three modalities are removed. This resulted in a dataset with 25,286 cells and 9 cell types, including 9,772 RNA, 46 ADT and 16,520 ATAC peak features.

Each data modality was first normalised by a z-score transformation and then fed into each of their modality-specific encoders in SnapCCESS. The integrated embeddings were jointly learned across all modalities using outputs of modality-specific encoders (see Section 3.2.1).

3.2.4 Performance evaluation criteria

3.2.4.1 Clustering concordance performance evaluation

For evaluating clustering performance, we used adjusted Rand index (ARI) and normalised mutual information (NMI) to evaluate the clustering concordance with respect to pre-defined cell type annotations from their original studies (Kim et al., 2019). Let S be a set of N cells, then a clustering U on S is a way of partitioning S into non-overlap subset $U_1, U_2, \dots, U_R$. Here, we define $U = U_1, U_2, \dots, U_R$ as the real cell type labels with R cell types, $V = V_1, V_2, \dots, V_C$ is a partition with C clusters generated by a clustering. Pair counting based measures can be used for counting pairs of items on which the partition U and V agree or disagree. Specifically, the $N/2$ item pairs in S can be classified into one of the four types: (i) N_{11} : the number of pairs that are in the same partition in both U and V ; (ii) N_{00} : the number of pairs that are in different partitions in both U and V ; (iii) N_{01} : the number of pairs that are in the same partition in U but in different partitions in V ; (iv) N_{10} : the number of pairs that are in different partitions in U but in the same partition in V . Following this, ARI, NMI can be defined as follows:

$$ARI(U, V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}$$

$$NMI(U, V) = \frac{I(U; V)}{H(U) + H(V)}$$

where $I(U; V)$ is the mutual information between U and V , defined as:

$$I(U; V) = \sum_{i=1}^R \sum_{j=1}^C \frac{|U_i \cap V_j|}{N} \log_2 \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

and $H(\cdot)$ is the entropy of partitions, in which:

$$H(U) = - \sum_{i=1}^R \frac{|U_i|}{N} \log_2 \frac{|U_i|}{N}; \quad H(V) = - \sum_{j=1}^C \frac{|V_j|}{N} \log_2 \frac{|V_j|}{N}$$

To investigate the performance of SnapCCESS in terms of major and minor cell type identification, we split the cell types in each dataset into two sets. The first set included major cell types, defined as those with a number of cells greater than or equal to the median value of the number of cells per cell type. The second set included minor cell types, defined as those with a number of cells less than the median value of the number of cells per cell type.

3.2.4.2 *Assessment of the run time usage*

To evaluate the computation speed of SnapCCESS and the conventional VAE ensemble clustering method, all benchmark tasks were allocated to a research server with an NVIDIA GPU GeForce RTX 2080 Ti. The elapsed run time was calculated by using the Python function `time.per_counter`. Time for each method only takes into account the deep learning network building and model training steps.

3.3 RESULTS

3.3.1 *SnapCCESS is an effective and efficient ensemble deep learning method for clustering multimodal single-cell omics data*

We first evaluated the performance of SnapCCESS and compared its performance with a conventional VAE ensemble clustering method on the six multimodal single-cell omics datasets generated from different biotechnological platforms. The concordance of the k-means clustering output from each method with respect to the cell type annotation from the original study of each dataset was quantified using ARI (Figure 3.2a) and NMI (Figure S1) and the procedure was repeated 20 times to account for the variability in the clustering results. Notably, we found that ensemble learning improves the clustering performance of both methods and the clustering concordance increases while the variability re-

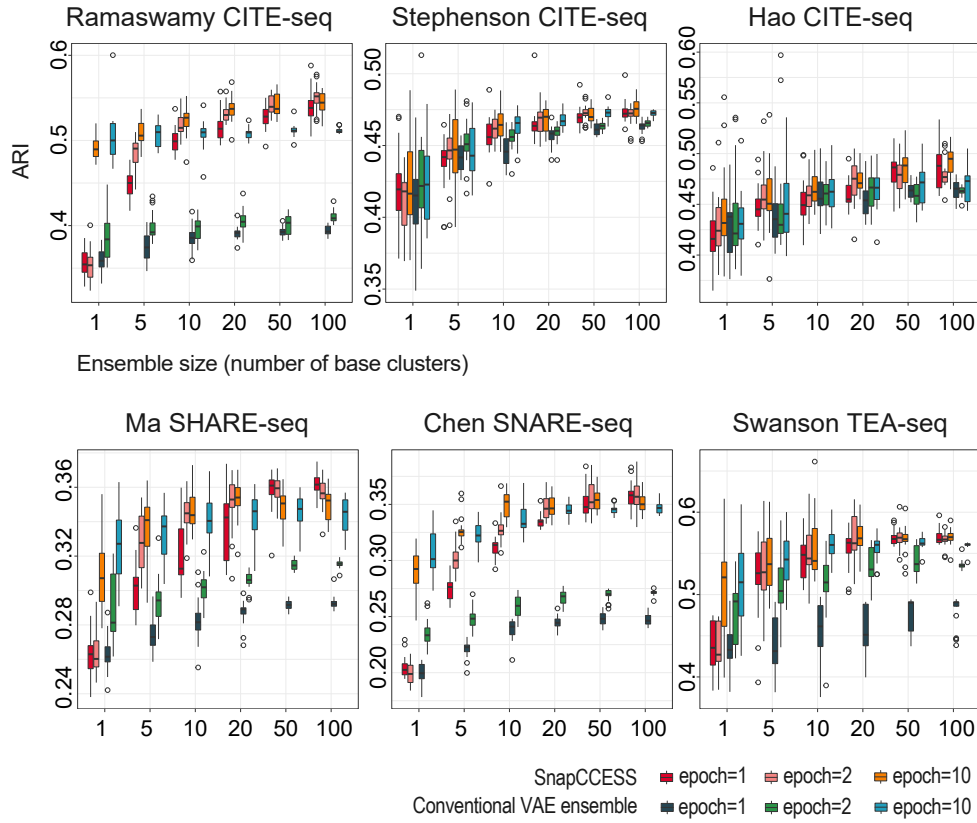


Figure 3.2: Clustering performance of SnapCCESS and conventional cluster ensembles trained by different numbers of epochs. Concordance of cell type clustering on six multimodal single-cell omics data. The x-axis is the number of base clusters for the ensemble and the y-axis is the concordance of the clustering output and the cell type annotation in the original studies quantified by ARI. The k-means clustering algorithm was used for clustering the embeddings generated from each method. The entire procedure was repeated 20 times for capturing the performance variability.

duces with the ensemble size (i.e. the number of base clusters). We also found that in general the improvement plateau at around the ensemble size of 50.

A key advantage of SnapCCESS is its ability to generate informative embeddings from multiple local minima and therefore requires much fewer epochs during the ensemble learning process (Figure 3.1b). To validate this, we tested using epochs of 1, 2, and 10 in SnapCCESS and the conventional VAE ensemble clustering. As expected, in most cases the conventional VAE ensemble clustering requires a high number of epochs to achieve high performance (Figure 3.2 and S1). In comparison, SnapCCESS achieves comparable performance using only one epoch in training the ensemble model at the sizes of 50 and 100. Since

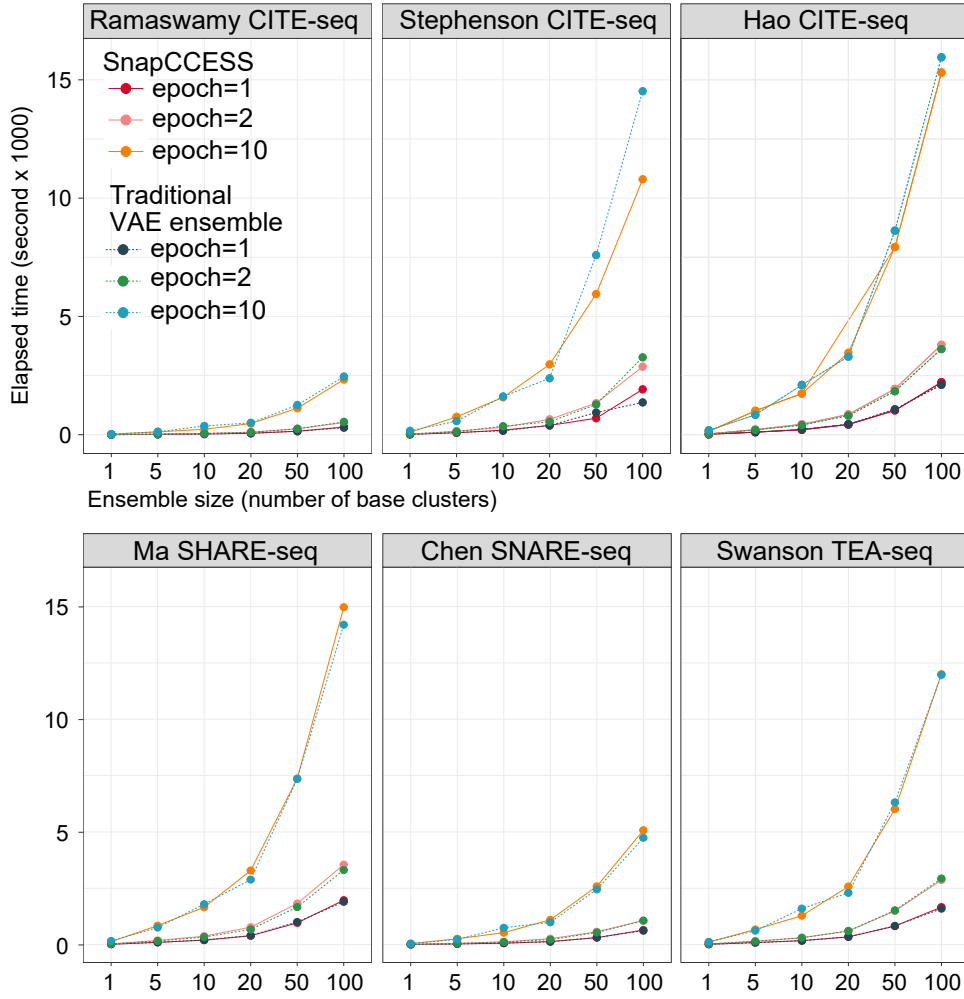


Figure 3.3: Clustering performance of SnapCCESS and conventional cluster ensembles trained by different numbers of epochs. Comparison of computation time for SnapCCESS and conventional VAE cluster ensembles. The x-axis is the number of base clusters included in the cluster ensembles and the y-axis is the elapsed time in the unit second. Results from SnapCCESS are presented as solid lines and those from conventional VAE cluster ensembles are presented as dashed lines. Epochs of 1, 2, and 10 were tested and denoted using different colours.

the epoch is a key parameter that defines the number of times the VAE will work through an input dataset, in general training on more epochs requires more computing time. To evaluate this, we recorded the computation time of SnapCCESS and the conventional VAE ensemble on each dataset. Indeed, we found that, for both methods, fewer epochs resulted in significantly faster computation, especially with large ensemble sizes (Figure 3.3b). Under the same number of epochs, the computation time of both methods is very similar. Nev-

ertheless, from our above analyses, only SnapCCESS could achieve high performance in clustering cells with a low training epoch and significantly outperforms the conventional VAE ensemble with an epoch of 1. Taken together, these findings demonstrate that combining individual clustering results derived from multiple embeddings does lead to more accurate and reproducible consensus clustering of multimodal single-cell omics data, and the SnapCCESS framework for ensemble deep learning of embeddings can achieve high-performance cell clustering using significantly less computational time.

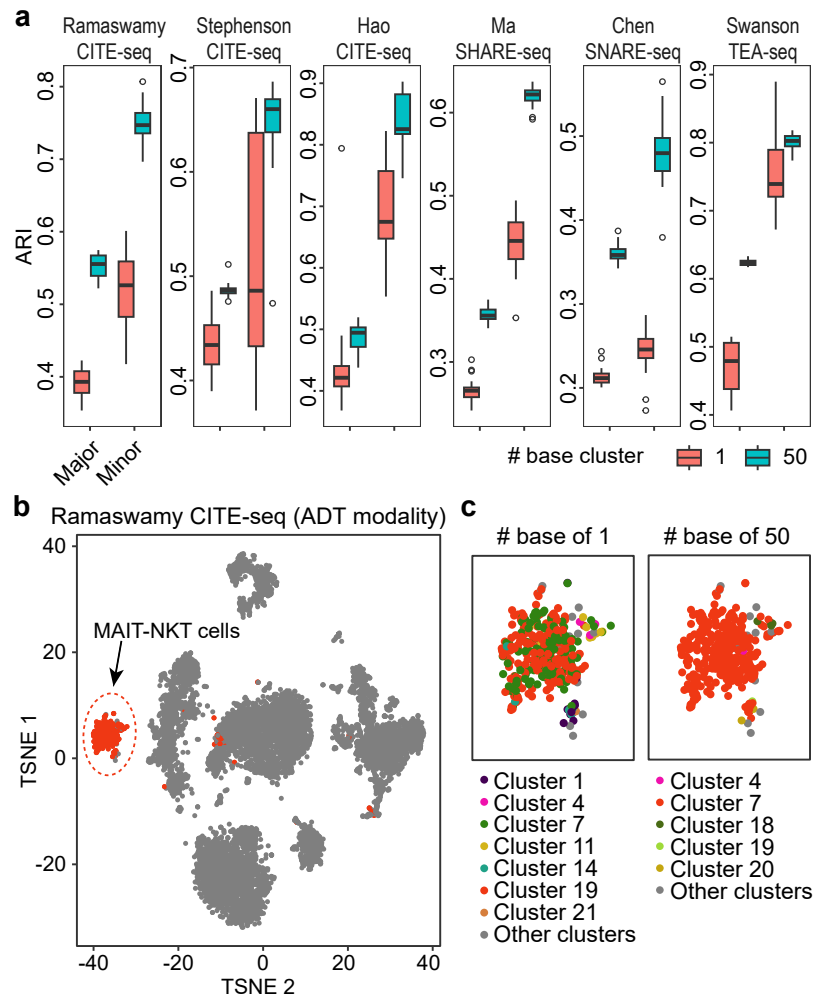


Figure 3.4: Corcordance of clustering output and the pre-defined cell type labels on (a) major and minor cell type quantified by ARI in each of the six datasets. (b) TSNE visualisation of Ramaswamy CITE-seq data using ADT modality. MAIT-NKT cells are highlighted in red. (c) Zoom in of MAIT-NKT cell clustering using SnapCCESS with 1 or 50 base clusters.

3.3.2 *Diagnostic analysis of SnapCCESS reveals its ability to improve major and minor cell type identification*

We evaluated the performance of SnapCCESS using different initial learning rates and varying numbers of cells in the dataset. Using the Swanson TEA-seq dataset, we found that the default learning rate of 0.02 resulted in high cell clustering performance whereas a larger learning rate of 0.2 led to a significant reduction in performance (Figure S2a). Next, we subsampled the cells from the Swanson TEA-seq dataset to test the impact of the number of cells on the performance of SnapCCESS. We found that datasets with small numbers of cells have a significant impact on the performance of SnapCCESS when no ensemble was used (base of 1) (Figure S2b). In comparison, SnapCCESS with an ensemble size of 50 base clusters are much more robust to data with small cell sizes, highlighting a key advantage of the ensemble technique.

To further investigate the performance of SnapCCESS on minor cell type identification, we split the cell types in each dataset into two sets with one containing major cell types and the other containing minor cell types (see Section 3.2.4.1). We then assessed each set for its concordance of clustering output and the pre-defined cell type labels. We found in all six datasets a higher clustering performance of SnapCCESS (epoch=1) with an ensemble of 50 compared to those without using the ensemble in both the major and minor partitions of the cell types (Figure 3.4a and S3a). Lastly, as a case study, we visualised the Ramaswamy CITE-seq dataset and highlighted the MAIT-NKT cells using either the ADT modality (Figure 3.4a) or the RNA modality (Figure S3b). We found that SnapCCESS with a base of 50 led to much better identification of cells from this cell type than those from using a base of 1 (Figure 3.4c and S3c). Taken together, these results suggest that the use of the snapshot ensemble technique in SnapCCESS can improve the identification of both the major and minor cell types.

3.3.3 Integrated embedding of multimodality generally leads to more precise cell clustering

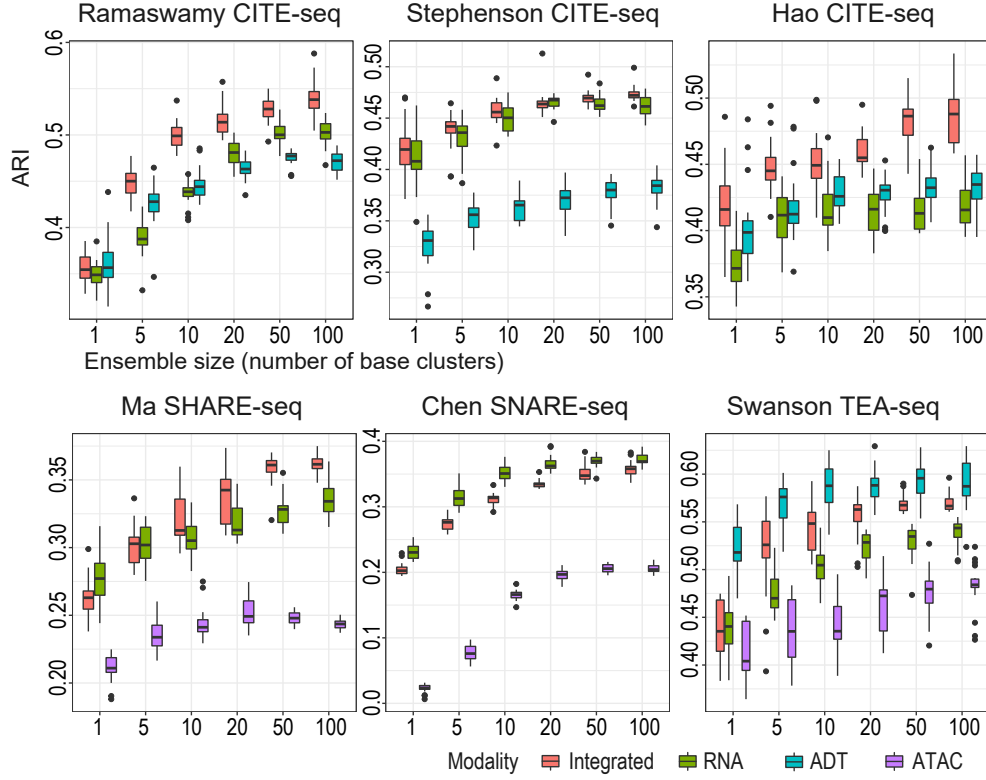


Figure 3.5: Comparison of embeddings learned from multimodal and unimodal approaches. The concordance of cell type clustering was quantified using the Adjusted Rand Index (ARI) on six multimodal single-cell omics datasets. Embeddings were generated by SnapCCESS from either all modalities combined or each modality alone. k-means clustering was used. The procedure was repeated 20 times to capture performance variability.

One of the key motivations in conducting multimodal single-cell omics experiments is the anticipation that the availability of multiple molecular features in individual cells will lead to more precise characterisation of cell identity and heterogeneity in complex multicellular organisms and biological systems (Zhu et al., 2020). To investigate this, for each dataset, we trained SnapCCESS (epoch=1) using either all available data modalities or each unimodality independently, and then performed cell clustering using the k-means clustering algorithm on either the integrated embedding of multimodality or embeddings from each unimodality. We found that in general clustering of cells using inte-

grated embeddings of multimodality leads to significantly better results than from using any unimodality alone (Figure 3.5 and S4). Among the clustering results using unimodal embeddings, those generated from RNA modality generally performed similarly or better compared to those from ADT modality. The clustering performance of ATAC modality appears to be lower compared to other modalities, which may be due to the higher dimensionality and data sparsity in ATAC data modality (Xiong et al., 2019). Together, these results support the expectation that taking into consideration of multiple molecular features of cells can lead to a more precise downstream characterisation of the biological systems, and further highlight the utility of modality integration methods for analysing multimodal single-cell omics data.

3.3.4 *SnapCCESS improves various clustering algorithms*

Since the ensemble deep learning of embeddings in SnapCCESS is independent of clustering algorithms, we next tested the performance of the SnapCCESS framework (epoch=1) by coupling it with a spectral clustering algorithm and SIMLR, a kernel-based clustering algorithm. Note that the two large CITE-seq datasets, Stephenson CITE-seq and Hao CITE-seq, were excluded due to the exponential growth of computational complexity with the number of cells in a dataset for the spectral clustering algorithm. Overall, we observed a clear increase in clustering performance with the increasing ensemble size, regardless of the types of clustering algorithms and concordance evaluation metrics (Figure 3.6 and S5). Nonetheless, compared to the simple k-means clustering algorithm, the application of more advanced SIMLR clustering and spectral clustering algorithms generally led to improved cell clustering as measured by their concordance to the cell type annotation. These findings are of particular interest to SIMLR, which was originally designed for analysing unimodal scRNA-seq data, as they demonstrate that embeddings learned by SnapCCESS from multimodal single-cell omics data can be used for a clustering algorithm designed for unimodal single-cell omics data.

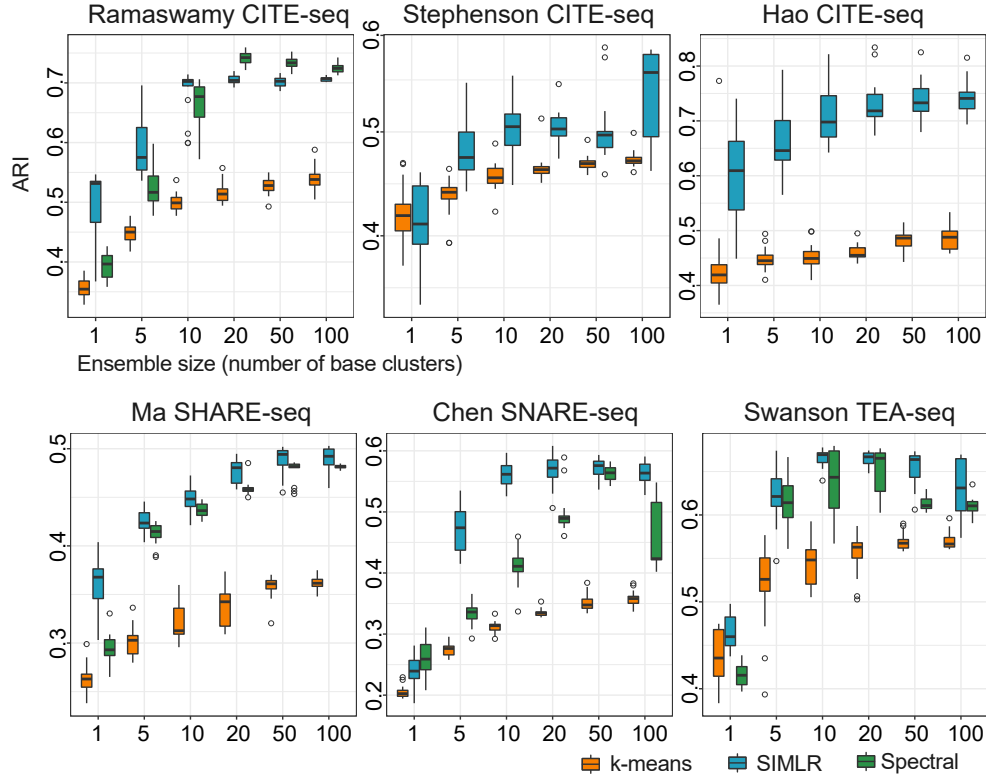


Figure 3.6: Evaluation of the SnapCCESS framework using alternative clustering algorithms on multimodal single-cell omics data. The concordance between cell type annotations and clustering outputs from each algorithm was assessed. The x-axis represents the number of base clusters in the ensemble, while the y-axis quantifies concordance using the ARI. The entire procedure was repeated 20 times to capture performance variability.

Similar to the results from k-means clustering, the improvement from SIMLR and spectral clustering also peaked around the ensemble size of 50 (Figure 3.6b and S5). This is also consistent with the results from scCCESS for scRNA-seq data analysis (Geddes et al., 2019), suggesting an ensemble size of 50 may be a suitable choice for the ensemble deep learning component in the SnapCCESS framework.

3.3.5 SnapCCESS performs competitively to the state-of-the-art embedding generating methods for multimodal single-cell clustering

Various methods exist for generating embeddings from multiple data modalities in single-cell multimodal omics data. Some of the state-of-the-art examples

include totalVI designed for combining RNA and ADT modalities in CITE-seq data and its sibling MultiVI for combining RNA and ATAC modalities such as in SHARE-seq and SNARE-seq, and Jvis-learn and MOFA+ which are generic and can be applied to all data modality combinations. Given the applicability of each method, we compared SnapCCESS (epoch=1 and ensemble size of 50) with TotalVI on the CITE-seq datasets, MultiVI on SHARE-seq and SNARE-seq datasets, and MOFA+ and Jvis-learn on all six datasets with two or three modalities. In particular, we used the multimodality integrated embeddings generated from each of these methods as input to k-means, SMILR, and spectral clustering algorithms and examined the concordance of clustering output with cell type annotation using ARI (Figure 3.7a) and NMI (Figure 3.7b).

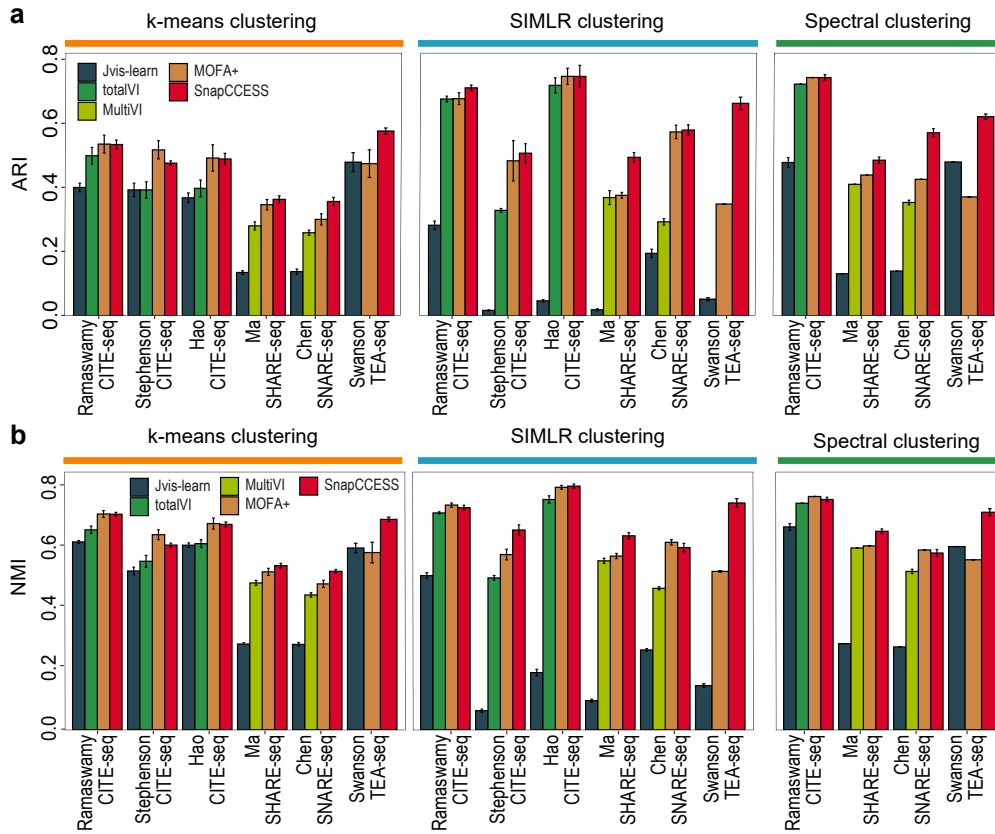


Figure 3.7: Comparison of SnapCCESS framework (epoch=1 and ensemble size of 50) with other multimodality embedding generation methods on cell clustering using k-means, SIMLR, and Spectral clustering algorithm. (a) Quantification of clustering concordance to cell type annotation using ARI. (b) Similar to (a) but quantifying clustering concordance using NMI. Each bar indicates the average performance across datasets with 20 repeats, and error bars represent the standard deviation.

We found that SnapCCESS performed competitively to other multimodality embedding generation methods. In most cases, its performance is significantly better than totalVI when applied to CITE-seq datasets and MultiVI when applied to SHARE-seq and SNARE-seq datasets (Figure 3.7). While MOFA+ also performed well especially on CITE-seq datasets, SnapCCESS appears to be slightly better than MOFA+ when used for generating embeddings and performing cell clustering on the SHARE-seq and the trimodal TEA-seq datasets. Interestingly, while clustering of the embeddings from Jvis-learn using k-means and Spectral clustering algorithms performed reasonably well, the use of SIMLR on Jvis-learn generated embeddings leads to poor results in many cases. These results may indicate the varying degree of generalisability of the multimodality embedding generation methods on different clustering algorithms. To this end, multimodality embeddings generated from the other four methods (i.e. SnapCCESS, totalVI, MultiVI, MOFA+) worked well regardless of the used clustering algorithm.

3.4 DISCUSSION AND CONCLUSION

A main challenge in single-cell omics data analysis is in handling the high-dimensionality of the feature space (Yang et al., 2021). Embedding learning is a popular approach for reducing feature dimension for subsequent analysis such as clustering of cells. While many methods have been designed to generate embeddings for unimodality scRNA-seq data, only a few methods are specifically designed for and can be applied to integrate data modalities in multimodal single-cell omics data. Our comparison to these methods highlights the utility of forming ensembles of embeddings for dimension reduction of multimodal single-cell omics data and their subsequent application in cell clustering.

A key innovation in SnapCCESS is the adaptation of the snapshot ensemble learning technique (Huang et al., 2017) which significantly reduces the computation time and resources for multiple embeddings compared to conventional VAE ensembles. The underlying idea of the snapshot ensemble is to save multi-

ple versions of a single model during the training process by using a cyclic learning rate scheduler. In SnapCCESS, this allows us to capture the multi-view of the latent space which leads to better clustering performance (Cao et al., 2020). Nevertheless, the clustering of each embedding generated from SnapCCESS is still performed individually. While parallelisation can be implemented to speed up the process, designing clustering algorithms that can cluster cells via multiple embeddings simultaneously could further improve computational efficiency. Related to this, there are various ensemble deep learning methods that aim to reduce computation time by using techniques such as model branching and neuron deactivation (Cao et al., 2020). The effectiveness of these alternative approaches for learning embeddings from multimodal single-cell omics data remains to be tested.

The utility of the embeddings generated from multimodal single-cell omics data is much wider than cell clustering. While clustering is one application that can make use of embeddings learned from such data, other tasks such as supervised cell type classification (Abdelaal et al., 2019) and unsupervised number of cell type estimation (Yu et al., 2022) that take the embeddings as input can also be applied for analysing multimodal single-cell omics data. Therefore, designing methods that can generate better embeddings will impact various downstream analyses and applications of multimodal single-cell omics data. To this end, the utility of ensemble deep learning methods for these applications (e.g. cell type classification, number of cell type estimation) should be investigated in future studies.

While the current study evaluates the SnapCCESS framework for clustering cells into discrete groups, the cell type structures from many biological systems are hierarchical, with subpopulations of cells existing in each major cell type (Wu and Wu, 2020). The development of multimodal single-cell omics technologies facilitates the characterisation of such hierarchical cell type relationships. Therefore, developing methods that are capable of multi-resolution clustering of cells on datasets with multimodal molecular attributes is a direction of future research. Another recent expansion in the single-cell omics field is the

increasing availability of spatial single-cell omics data produced by an array of new spatial profiling technologies (Larsson et al., 2021). Combining spatial data with other omics data types produced from the same cells and samples has the potential to uncover a wealth of information, including spatial-related cell type structure, and will help us gain a deeper understanding of cell and tissue development and disease progression. Developing clustering algorithms for integrating spatial data with other omics datasets is challenging and requires further methodological innovation.

In summary, our previous work demonstrated that ensemble learning of embeddings provides an effective approach for improving downstream clustering analyses by providing a multi-view of the input data (Geddes et al., 2019). Here we extend this idea for single-cell multimodal omics data analysis by introducing SnapCCESS, an efficient ensemble deep learning framework using VAE and snapshot techniques, gaining high performance in cell clustering while alleviating the limitation on computation efficiency in conventional ensemble learning methods. Since the clustering of individual embeddings can be performed independently from each other, the proposed framework can benefit from further speed up by parallelisation of embedding clustering. We expect SnapCCESS to serve as a useful tool and spark the future development of ensemble deep learning methods for multimodal single-cell omics data analysis.

INTERPRETABLE DIFFERENTIAL ABUNDANCE SIGNATURE

4.1 BACKGROUND AND CURRENT APPROACH

The recent increase in multi-sample and multi-condition single-cell studies offers significant potential to uncover molecular pathways and cell types crucial for the onset, progression, and treatment of diseases (Stephenson et al., 2021; Burdziak et al., 2023; Aissa et al., 2021). These studies enable researchers to identify and analyze phenotype-specific subpopulations and marker genes associated with various conditions. The standard approach to achieve this generally involves two steps. The first step involves the identification of subgroups of cells that change in abundance in response to a given phenotype (e.g., Milo, DA-seq). The second step involves annotating these subpopulations by identifying markers that characterize these groups. Together, this enables scientists to reach certain conclusions regarding cell states and their corresponding markers which drive these changes.

Recently, a suite of algorithms have been developed to analyze changes in cell populations in response to disease progression or experimental interventions. Notable methods include Milo (Dann et al., 2021), CNA (Reshef et al., 2021), DA-seq (Zhao et al., 2021), MELD (Burkhardt et al., 2021), and Scissor (Sun et al., 2021). Most of these methods begin by constructing a KNN graph and summarizing it into an abundance matrix. Statistical methods such as generalized linear models, logistic ridge regression, and kernel density estimation are then applied to link this matrix with sample-level phenotype labels, ultimately identifying phenotype-specific subpopulations by providing perturba-

tion scores representing the significance of differential abundance (DA) tests or probabilities related to the phenotype. After identifying these subgroups, the next step involves selecting distinct signatures of these differentially abundant groups for more detailed analysis, typically using classic differential expression analysis methods. Differential expression (DE) analysis on single cells is currently approached from two perspectives. The first is the sample-level view, where expression is aggregated to create "pseudobulk" samples that are then analyzed using methods originally designed for bulk RNA samples, such as edgeR (Robinson et al., 2009), DESeq2 (Love et al., 2014) or Limma-voom (Ritchie et al., 2015; Law et al., 2014). The second is the single-cell level view, where cells are modelled individually using generalized mixed effects models, such as MAST (Finak et al., 2015) or glmmTMB (Brooks et al., 2017). Together, these approaches help understand the gene features most affected by pathological or experimental conditions, providing valuable insights into the underlying biological processes.

Although these workflows do not usually require cell type annotation in the initial differential abundant groups identification, the subsequent marker gene identification often becomes complicated as it is often required to align or account for cell types or other related factors. For instance, many DA analysis methods are clustering-free, allowing them to detect DE genes between differential abundance regions, which may include multiple cell types or experimental conditions, thus complicating the interpretation of these DE genes. To make these genes more interpretable, it is critical to jointly examine the effects of cell types and other potentially related factors.

To address these challenges, we developed a computational framework to identify signatures generated from differential abundance tests, while taking into account the effects of cell types and other potentially relevant factors. We build an "ANOVA"-like model to identify the gene signatures in the subpopulation groups identified by the DA test. The model was used to assess whether gene signatures exhibit significant responses to factors of interest. For example, our approach aims to separate biomarkers or biosignatures that are common to all

individuals or only specific to a given cell type. We believe that this analytical approach will improve the clarity of data analysis by better classifying the signal from the data through the analysis of variance (ANOVA), thereby leading to better understanding of the relation between gene signature and different factors.

In this chapter, we first describe an ANOVA-based framework to categorise signatures into four groups. Then, we demonstrate the power of our approach by identifying gene signatures associated with perturbed cellular states and specific cell types from publicly available datasets in the context of melanoma immunotherapy treatment response status and validate it again in external data. We next extend the framework to identify gene signatures in multi-factor longitudinal and spatially resolved datasets. Finally, we finish this chapter with a discussion and conclusion.

4.2 DATA

4.2.1 *Single-cell and bulk RNA sequencing*

In this study, we used three public datasets of melanoma immunotherapy patients to demonstrate the iDAS framework: one single-cell RNA-seq dataset and two bulk RNA-seq datasets.

Pozniak2024: The single-cell RNA-seq data is from a longitudinal profiling study of metastatic melanoma under immune checkpoint blockade (ICB) (Pozniak et al., 2024). Biopsies were taken from cutaneous, subcutaneous, or lymph node metastases before treatment (BT) and early on-treatment (OT; before the second ICB infusion) and processed for scRNA-seq using the 10x Genomics Chromium platform. For two-way iDAS analysis, we used the pre-treatment data, filtering out cell states named Patient-specific-A and Patient-specific-B and only including protein coding genes. This resulted in a dataset of 18 patients with 8,789 cells and 14,346 genes for the differential abundance analysis to identify significant DA groups. For the three-way iDAS analysis, we used

the pre- and on-treatment data together, which includes 13,789 cells and 13,887 genes. The single-cell dataset was mapped to the GRCh38 genome reference.

Liu2019: This bulk RNA-seq data comes from a cohort of patients with advanced melanoma who received PD1 blockade as palliative treatment (Liu et al., 2019). FFPE tissues were obtained before PD-1 blockade. Patients were classified into five clinical statuses in the original study: Complete Response (CR), Partial Response (PR), Minor Response (MR), Progressive Disease (PD), and Stable Disease (SD). For our analysis, we selected CR and PR as responders ($n = 47$) and PD and SD as non-responders ($n = 72$) resulting in a dataset of 119 individuals. The data was aligned using STAR with GRCh38 reference and quantified with RSEM to yield gene-level expression in transcripts per million (TPM). We directly used the TPM matrix for our analysis.

Hugo2016: This bulk RNA-seq data comes from a cohort of patients with metastatic melanoma who received either pembrolizumab or nivolumab as their anti-PD-1 therapy (Hugo et al., 2016). Melanoma specimens before treatment with sufficient RNA quality were analyzed by RNA-Seq. Paired-end transcriptome reads were mapped to the UCSC hg19 reference genome using Tophat2. Normalized expression levels of genes were reported in FPKM values. We directly used the FPKM matrix of 26 patients for our analysis.

4.2.2 *Spatially resolved transcriptomics data*

We used the GHIST (Fu et al., 2024) prediction results applied to a set of TCGA breast cancer dataset. We select a set of samples focused on the HER2+ breast cancer subtype from TCGA-BRCA, we selected the HER2+ patients based on a positive entry in the “lab_proc_her2_neu_immunohistochemistry_receptor_status” metadata column resulting in 92 TCGA samples with corresponding matched H&E images. The in-silico prediction was describe in detailed from GHIST paper. After filtering out NA values in the AJCC pathological stage, we selected only ER/PR positive (ER+/PR+; $n = 54$) and ER/PR negative (ER-/PR-; $n = 22$) in-silico spatially resolved transcriptomics samples. In total, 274 genes were

predicted using the GHIST package. We then applied scFeatures (Cao et al., 2022) to calculate the nearest neighbor correlation (nnCorrelation) between these genes and their most adjacent neighboring cells, resulting in a nnCorrelation matrix of 274 genes across 76 samples.

4.3 INTERPRETABLE DIFFERENTIAL ABUNDANCE SIGNATURE USING AN ANOVA-BASED FRAMEWORK

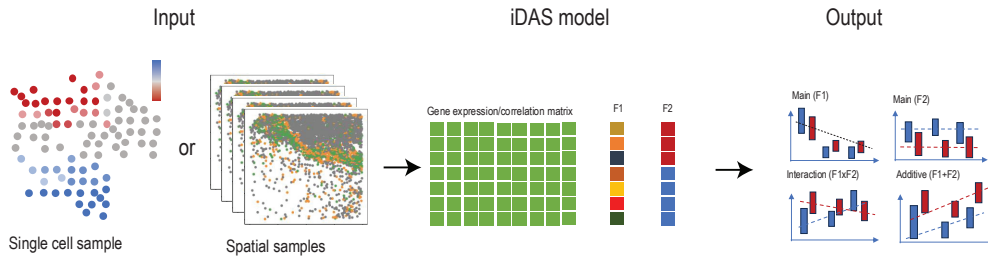


Figure 4.1: The proposed iDAS (Interpretable Differential Abundance Signature) framework identifies interpretable differential abundance signatures. iDAS can incorporate any upstream differential abundance method or spatial samples analysis results as a starting point. It creates cell type-by-sample pseudobulk profiles to perform ANOVA analysis, distinguishing genes significantly related to any main effect, F_1 and F_2 . These genes are then further categorized into four groups: main effect (F_1 , F_2), interaction effect ($F_1 \times F_2$), and additive effect ($F_1 + F_2$). To further interpret the gene signatures, additional differential expression analyses are performed to determine which category of effect each gene signature is most associated with.

We present a two-step ANOVA-based post-analytic approach to improve interpretation for perturbation studies, called Interpretable Differential Abundance Signature (iDAS), as shown in Figure 4.1. iDAS involves two main components, the first component uses a series of nested ANOVA-statistics to classify genes into different group, and the second component performs relevant differential expression analysis for signature identification and interpretation. Each step is described in the following subsections.

4.3.1 ANOVA-based test to classify genes

4.3.1.1 Two-way fix effect model

The development of the ANOVA-like tests in our framework was inspired by the nonparametric ANOVA (NANOVA) method (Zhou and Wong, 2011), which was originally proposed for classifying genes based on their factor effects in microarray data. In this approach, we model gene expression as the response variable and the cell type or experiment factors as explanatory variables. Specifically, in two-way factorial experiments, we can summarise gene expression by one of the following ANOVA models.

- Model 1: $y_{ijm} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijm}$
- Model 2: $y_{ijm} = \mu + \alpha_i + \beta_j + e_{ijm}$
- Model 3: $y_{ijm} = \mu + \alpha_i + e_{ijm}$
- Model 4: $y_{ijm} = \mu + \beta_j + e_{ijm}$
- Model 5: $y_{ijm} = \mu + e_{ijm}$

In these models, $\alpha_i (i = 1, \dots, I)$ and $\beta_j (j = 1, \dots, J)$ denote the two factors of interest at levels i and j , respectively, and $y_{g,ijm}$ represents the expression of gene g under the condition defined by (α_j, β_j) . Here, $m (m = 1, \dots, M_{ij})$ is a subscript for replicates. The 'interactive' model (Model 1) includes μ representing the baseline gene expression and $(\alpha\beta)_{ij}$ as the interaction term, indicating that genes are influenced by both factors and the effect of one factor depends on the level of the other. The 'additive' model (Model 2) shows that genes are affected by both factors independently, without any interaction effect. 'Single-factor' models (Models 3 and 4) illustrate genes affected by only one factor, either α_i or β_j . Lastly, the 'null' model (Model 5) indicates that gene expression is not influenced by either factor.

Next, we classify genes into five groups ($C_{F1}, C_{F2}, C_{F1 \times F2}, C_{F1+F2}, C_{non-sig}$), each corresponding to one of the models described earlier. This classification is based on a series of ANOVA tests as detailed below.

1. No effect Test: In this step, we test whether the gene expression is affected by any factor, corresponding to the null hypothesis $H_0 : \alpha_i = \beta_j = (\alpha\beta)_{ij} = 0; (i = 1, \dots, I, j = 1, \dots, J)$ (Model 5) vs alternative hypothesis $H_1 : \forall \alpha_i \neq 0, \beta_j \neq 0, (\alpha\beta)_{ij} \neq 0; (i = 1, \dots, I, j = 1, \dots, J)$ (Model 1). We set the significance level of the p-value by two methods: either by ordering the p-values and defining the top 2% as the significance level, or by using a predefined significance level, such as 0.05 or 0.01. Genes with p-values larger than this significance level are considered part of the non-significant group ($C_{\text{non-sig}}$). The significant genes (C_{sig}) are further classified into four groups ($C_{F1}, C_{F2}, C_{F1 \times F2}, C_{F1+F2}$).

2. Interaction effect Test: We further classify (C_{sig}) into interaction effect and non-interaction effect groups. The null hypothesis is $H_0 : (\alpha\beta)_{ij} = 0; (i = 1, \dots, I, j = 1, \dots, J)$ (Model 2), indicating no interaction effects between factor 1 and factor 2. The alternative hypothesis is $H_1 : (\alpha\beta)_{ij} \neq 0; (\forall i = 1, \dots, I, j = 1, \dots, J)$, which corresponds to Model 1. Follow the same idea from “No effect test”, the set of gene signatures with significant interaction effects is defined as having p-values less than significance level and is denoted as $C_{F1 \times F2}$.

3. Main effect Test: To test the main effect of α ($F1$), the null hypothesis is $H_0 : \alpha_i = 0 (i = 1, \dots, I)$, indicating no effect of $F1$ on the gene signatures, which corresponds to Model 4. The alternative hypothesis is $H_1 : \alpha_i \neq 0; (\forall i = 1, \dots, I)$, corresponding to Model 2. Similarly, to test the main effect of β_j ($F2$), the null hypothesis is $H_0 : \beta_j = 0; (j = 1, \dots, J)$, indicating the set of gene signatures with no factor 2 effect, which corresponds to Model 3. And the alternative is still Model 2, $H_1 : \beta_j \neq 0; (\forall j = 1, \dots, J)$. We applied the main effect test on the genes set $C_{\text{sig}} \setminus C_{F1 \times F2}$ to identify genes having α ($F1$) and β ($F2$) effect respectively. Denote these two sets as C_α and C_β , then we can get $C_{F1} = C_\alpha \setminus (C_\alpha \cap C_\beta)$ and $C_{F2} = C_\beta \setminus (C_\alpha \cap C_\beta)$.

4. Additive effect: After performing the three tests described above, the remaining genes in C_{sig} are classified into C_{F1+F2} , representing the additive effect group. This set is defined as $C_{\text{sig}} \setminus (C_{F1 \times F2} \cup C_{F1} \cup C_{F2})$

4.3.1.2 Three-way random effect model

In this section, we build a model of three-way ANOVA classification with random effect. Specifically, in three-way factorial experiments, we can summarise gene expression by one of the following models.

- Model 1: $y_{ijknm} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk} + u_n + e_{ijknm}$
- Model 2: $y_{ijknm} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + u_n + e_{ijknm}$
- Model 3: $y_{ijknm} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + u_n + e_{ijknm}$
- Model 4: $y_{ijknm} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + u_n + e_{ijknm}$
- Model 5: $y_{ijknm} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + u_n + e_{ijknm}$
- Model 6: $y_{ijknm} = \mu + \alpha_i + \beta_j + \gamma_k + u_n + e_{ijknm}$
- Model 7: $y_{ijknm} = \mu + \alpha_i + u_n + e_{ijknm}$
- Model 8: $y_{ijknm} = \mu + \beta_j + u_n + e_{ijknm}$
- Model 9: $y_{ijknm} = \mu + \gamma_k + u_n + e_{ijknm}$
- Model 10: $y_{ijknm} = \mu + u_n + e_{ijknm}$

In these models, $\alpha_i (i = 1, \dots, I)$, $\beta_j (j = 1, \dots, J)$ and $\gamma_k (k = 1, \dots, K)$ denote the three factors of interest at levels i, j and k , respectively, and $y_{g,ijknm}$ represents the expression of gene g under the condition defined by $(\alpha_j, \beta_j, \gamma_k)$. Here, $n (n = 1, \dots, N_{ijk})$ is a subscript for random effect factor, $m (m = 1, \dots, M_{ijk})$ is a subscript for replicates. The baseline gene expression is denoted as μ . $(\alpha\beta\gamma)_{ijk}$ is the interaction term, indicating that genes are influenced by three factors and the effect of one factor depends on the level of the others. $(\alpha\beta)_{ij}$, $(\beta\gamma)_{jk}$, and $(\alpha\gamma)_{ik}$ are the interaction term of two factors. Model 1 is the full model, including all three factors with two-way and three-way interaction effects, and a random effect u_n . Model 2 only includes two-way interaction effects. Models 3, 4, and 5 include only two-way interaction effects, but each model excludes one

of the two-way interaction effects. Model 6 illustrates that genes are affected by three factors independently, without any interaction effect. Model 7, 8, and 9 are single factor model, shows that genes are affected by only one factor. Lastly, Model 10 indicates that gene expression is not influenced by any factor.

Next, we classify genes into ten groups (C_{F1} , C_{F2} , C_{F3} , $C_{F1 \times F2}$, $C_{F1 \times F3}$, $C_{F2 \times F3}$, $C_{\text{two-waycombine}}$, C_{additive} , $C_{F1 \times F2 \times F3}$, $C_{\text{non-sig}}$), each corresponding to one of the models described earlier. This classification is based on a series of ANOVA tests as detailed below.

1. No effect Test: In this step, we test whether the gene expression is affected by any factor, corresponding to the null hypothesis $H_0 : \alpha_i = \beta_j = \gamma_k = (\alpha\beta)_{ij} = (\beta\gamma)_{jk} = (\alpha\gamma)_{ik} = 0, (\alpha\beta\gamma)_{ijk} = 0; (i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K)$ (Model 10) vs alternative hypothesis $H_1 : \forall \alpha_i \neq 0, \beta_j \neq 0, \gamma_k \neq 0, (\alpha\beta)_{ij} \neq 0, (\beta\gamma)_{jk} \neq 0, (\alpha\gamma)_{ik} \neq 0, (\alpha\beta\gamma)_{ijk} \neq 0; (i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K)$ (Model 1). We set the significance level of the p-value by two methods: either by ordering the p-values and defining the top 2% as the significance level, or by using a predefined significance level, such as 0.05 or 0.01. Genes with p-values larger than this significance level are considered part of the non-significant group ($C_{\text{non-sig}}$). The significant genes (C_{sig}) are further classified into other groups.

2. Interaction effect Test: We further classify (C_{sig}) into interaction effect and non-interaction effect groups. The null hypothesis is $H_0 : (\alpha\beta)_{ij} = 0, (\beta\gamma)_{jk} = 0, (\alpha\gamma)_{ik} = 0, (\alpha\beta\gamma)_{ijk} = 0; (i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K)$ (Model 6), indicating no interaction effects among factors. The alternative hypothesis is $H_1 : (\alpha\beta)_{ij} \neq 0, (\beta\gamma)_{jk} \neq 0, (\alpha\gamma)_{ik} \neq 0, (\alpha\beta\gamma)_{ijk} \neq 0; \forall i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$, which corresponds to Model 1. The set of gene signatures with significant interaction effects is denoted as C_{Int} , while set of gene signatures without significant interaction effects is denoted as C_{NotInt} .

3. Main effect Test: To test the main effect of α (F_1), the null hypothesis is $H_0 : \alpha_i \neq 0, \beta_j = 0, \gamma_k = 0; (i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K)$, indicating no effect of β_j (F_2) and γ_k (F_3) on the gene signatures, which corresponds to Model 7. The alternative hypothesis is $H_1 : \alpha_i \neq 0, \beta_j \neq 0, \gamma_k \neq 0; \forall i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$, corresponding to Model 6.

Similarly, to test the main effect β_j (F2), the null hypothesis is $H_0 : \alpha_i = 0, \beta_j \neq 0, \gamma_k = 0; (i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K)$, indicating the set of gene signatures with only F2 effect, which corresponds to Model 8. And the alternative is still Model 6, $H_1 : \alpha_i \neq 0, \beta_j \neq 0, \gamma_k \neq 0; \forall i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$.

Lastly, to test the main effect of γ_k (F3), the null hypothesis is $H_0 : \alpha_i = 0, \beta_j = 0, \gamma_k \neq 0; (i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K)$, indicating the set of gene signatures with only F3 effect, which corresponds to Model 9. And the alternative is still Model 6, $H_1 : \alpha_i \neq 0, \beta_j \neq 0, \gamma_k \neq 0; \forall i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$.

We applied the above tests on the genes set C_{NotInt} to identify genes having α (F1), β (F2) and γ_k (F3) effect, respectively. C_{F1} is the gene set that fails the main effect test for F1 but passes the main effect tests for F2 and F3. Conversely, C_{F2} fails the main effect test for F2 but passes the main effect tests for F1 and F3. And C_{F3} fails the main effect test for F3 but passes the main effect tests for F1 and F2.

4. Additive effect: After performing the three tests described in main effect test, the remaining genes in C_{NotInt} are classified into C_{additive} , representing the additive effect group, which includes $F1 + F2$, $F1 + F3$, $F2 + F3$ and $F1 + F2 + F3$.

5. Three-way interaction effect test: For the interaction gene set C_{Int} , we further classified the genes into groups based on three-way interaction effects and two-way interaction effects. The null hypothesis is $H_0 : (\alpha\beta\gamma)_{ijk} = 0; i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$, indicating there is no three-way intersection effect, which corresponds to Model 2. And the alternative is $H_1 : (\alpha\beta\gamma)_{ijk} \neq 0; \forall i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$, which corresponds to Model 1. With this test, C_{Int} can be partitioned into $C_{F1 \times F2 \times F3}$ when p-value less than the significance level, and $C_{\text{two-way}}$ when p value greater than cutoff value.

6. Two-way interaction effect test: To further interpret the two-way interaction effect genes, we classified them into four groups: $F1F2$, $F1F3$, $F2F3$, and a general two-way combination group.

To test the two-way interaction effect of $(\alpha\beta)_{ij}$ ($F1F2$), the null hypothesis is $H_0 : (\alpha\beta)_{ij} = 0; i = 1, \dots, I, j = 1, \dots, J$, indicating no effect of $(\alpha\beta)_{ij}$ ($F1F2$) on

the gene signatures, which corresponds to Model 3. The alternative hypothesis is $H1 : (\alpha\beta)_{ij} \neq 0; \forall i = 1, \dots, I, j = 1, \dots, J$, corresponding to Model 2.

Similarly, to test the two-way interaction effect of $(\beta\gamma)_{jk}$ (F2F3), the null hypothesis is $H0 : (\beta\gamma)_{jk} = 0; j = 1, \dots, J, k = 1, \dots, K$, indicating no effect of $(\beta\gamma)_{jk}$ on the gene signatures, which corresponds to Model 4. The alternative hypothesis is $H1 : (\beta\gamma)_{jk} \neq 0; \forall j = 1, \dots, J, k = 1, \dots, K$, corresponding to Model 2.

Lastly, to test the two-way interaction effect of $(\alpha\gamma)_{ik}$ (F1F3), the null hypothesis is $H0 : (\alpha\gamma)_{ik} = 0; i = 1, \dots, I, k = 1, \dots, K$, indicating no effect of $(\alpha\gamma)_{ik}$ (F1F3) on the gene signatures, which corresponds to Model 5. The alternative hypothesis is $H1 : (\alpha\gamma)_{ik} \neq 0; \forall i = 1, \dots, I, k = 1, \dots, K$, corresponding to Model 2. We applied the tests on the genes set $C_{\text{two-way}}$ to identify genes having $(\alpha\beta)_{ij}$ (F1F2), $(\beta\gamma)_{jk}$ (F2F3) and $(\alpha\gamma)_{ik}$ (F1F3) interaction effect respectively. $C_{F1 \times F2}$ is the gene set that passes the two-way interaction effect test for F1F2 but fails the main effect tests for F2F3 and F1F3. Conversely, $C_{F2 \times F3}$ passes the main effect test for F2F3 but fails the main effect tests for F1F2 and F1F3. And $C_{F1 \times F3}$ passes the main effect test for F1F3 but fails the main effect tests for F1F2 and F2F3.

After performing the three tests described above, the remaining genes in $C_{\text{two-way}}$ are classified into $C_{\text{two-way combine}}$, representing the two-way combination effect group.

4.3.2 Differential expression analysis for signature identification and interpretation

We illustrate the second step using two-way fix effect model. After classifying the gene list into four categories, the iDAS framework performs differential gene expression analysis at each level of the factors for both main effects and interaction effects. In our case study, where factor 1 is cell state and factor 2 is phenotype, the process is as follows:

1. Main Effect F₁ (Cell State): We perform differential expression (DE) analysis using limma (Ritchie et al., 2015) on the main effect-related gene signatures to identify cell markers characterizing each cell state.

2. Main Effect F₂ (Phenotype): We perform DE analysis using limma on the main effect-related gene signatures to find signatures that are highly differentially expressed between the responding and non-responding cell groups.

3. Interaction Effect: For interaction effect, we are most interested in cell state specific responding or non-responding signatures. First, we identify the top 10 marker genes of each cell state in the interaction signatures. Then, we repeat the DE analysis using these marker genes on the two phenotypes to determine whether any cell state marker genes are also differentially expressed between the two phenotype groups. Then we selected genes that are specifically expressed in one phenotype of a cell state and denoted them as cell state-specific responding or non-responding markers.

4.4 APPLICATION ON MELANOMA DATA ILLUSTRATE IDAS CAPACITY TO IDENTIFY CELL STATE SPECIFIC IMMUNOTHERAPY RESPONSE GENE SIGNATURES

To demonstrate our framework, we applied iDAS to a public dataset measuring immunotherapy response in melanoma cells (Pozniak et al., 2024). The dataset includes 20 patients, with samples taken before and during treatment. We used the before treatment dataset of 18 patients, filtering out patient-specific cell states not commonly present in melanoma. This resulted in 8,789 cells and 14,346 genes across 9 cell states being used in the differential abundance (DA) analysis using Milo. Milo was selected because it is a clustering-free differential abundance analysis method, which allows us to work directly with k-nearest neighbour graphs without relying on predefined clustering. Additionally, a benchmarking study (Yi et al., 2024) shows that Milo performs better than other methods for datasets with technical or biological noise. Patient-level phenotypes (responder and non-responder) were used as the conditions of interest

to calculate cell presence in each neighbourhood (see Supplementary Figure S1a). In total, we identified 416 DA subgroups by applying an FDR cutoff of ≤ 0.2 and $\log_2\text{FC} \geq 5$ (see Supplementary Figure S1b). Selected DA in neighbourhood graph, along with $\log_2$ fold change of differential abundance cells across patients, cell states, and patient phenotypes, are shown in Supplementary Figure S1c-f. Specially, the DA groups of responding and non-responding cells co-appeared in several cell states, such as mitochondrial, melanocytic, and antigen-presenting cell states, this suggests that there may be gene signature associated with both cell state and phenotype (see Supplementary Figure S1e).

4.4.1 Identification of cell states specific marker (main effect F1)

Next, we aimed to discover the gene signatures that affect the immunotherapy phenotype and cell states by applying iDAS to the selected 416 DA subgroups of cells, totalling 4,547 cells (Supplementary Figure S1c). To avoid the high false discovery rate commonly associated with single-cell analysis, usually due to the large number of replicates and violation of independence assumptions that can lead to false signals, we created pseudobulk samples instead of directly using single cells in our analysis. We calculated the average gene expression per cell state per patient. Additionally, we filtered out pseudobulk samples that only contained one cell in any cell state. We then performed the iDAS analysis on the pseudobulk matrix, setting factor 1 to cell state and factor 2 to patient-level phenotype. Our result shows that 158 genes were classified as factor 1 (cell state), 27 genes were classified as factor 2 (patient-level phenotype), 63 genes were classified as interaction effect genes, and 33 genes were classified as additive effect genes (see Supplementary Table S1). The remaining 13,791 genes were not significantly associated with either cell state or phenotype.

To assess whether the iDAS approach is able to identify known cell state markers, we examine the gene markers associated with the various cell states (F1, see Figure 4.2a) based on differential gene expression analysis on each cell state. Figure 4.2b shows that these top markers are highly expressed in specific

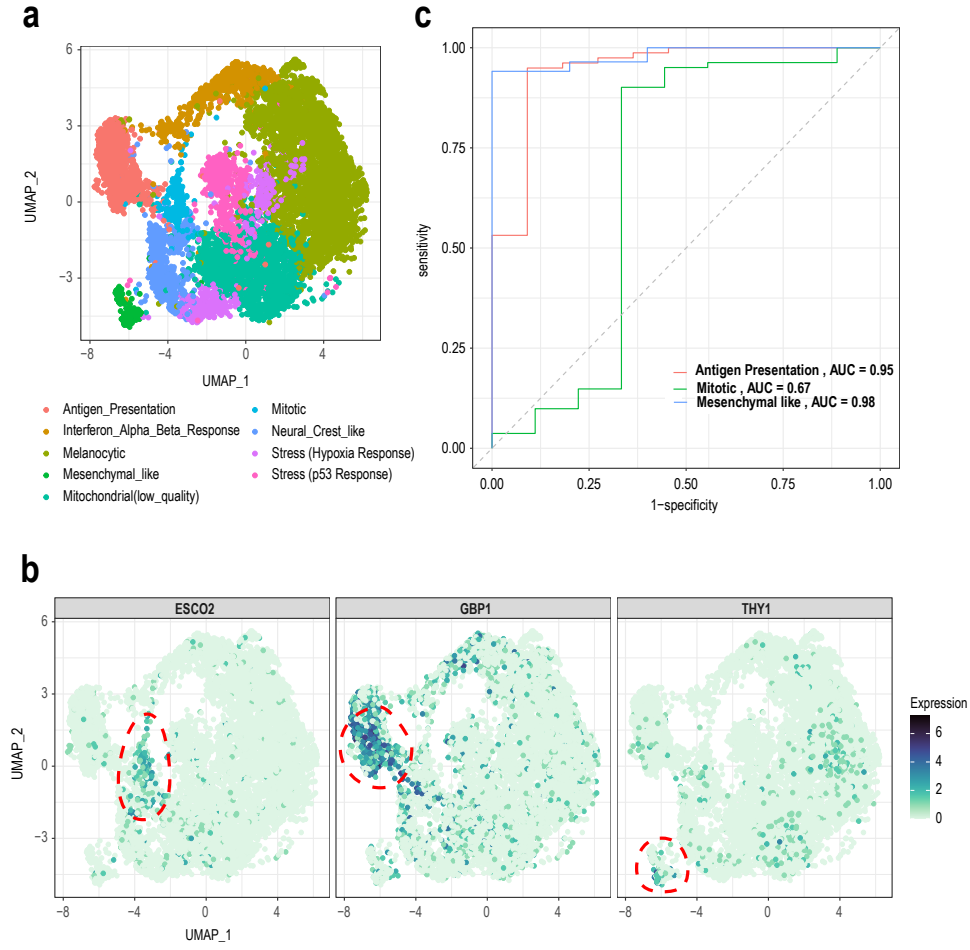


Figure 4.2: iDAS cell state associated main effect signatures. **(a)** UMAP plot of cell state annotations in Pozniak2024 data. **(b)** UMAP plot of single-cell data, highlighting the top main effect signatures for three cell states. Gene expression are shown for *GBP1* in antigen-presenting cells, *ESCO2* in mitotic cells, and *THY1* in mesenchymal-like cells. **(c)** The ROC curves depict the performance of classification models distinguishing each cell state (Antigen-Presentation, Mesenchymal-like, Mitotic) from all other cell states in the dataset. The area under the curve (AUC) scores are as follows: Antigen-Presentation (AUC = 0.97), Mesenchymal-like (AUC = 0.86), and Mitotic (AUC = 0.87).

cell states as expected and align with the original study (Pozniak et al., 2024). For example, we identify *ESCO2* in mitotic cells ($\log_2\text{FC} = 1.02$; adjusted p-value $< 10^{-10}$), *GBP1* in antigen-presenting cell state ($\log_2\text{FC} = 2.27$; adjusted p-value $< 10^{-10}$), and *THY1* in mesenchymal-like cells ($\log_2\text{FC} = 0.78$; adjusted p-value $< 10^{-10}$). In parallel, we also assess the predictability of these cells using a qualitative metric and this is achieved by calculating the average area under the ROC curve (AUC) based on 10-fold nested cross-validation on a pre-

diction model. The discrimination capacity of cell state-specific gene signatures with AUC from 0.52 to 0.99 (see Supplementary Figure S2). Figure 4.2c shows a strong discrimination capacity for the example cell states, and the highest cell state is antigen presentation. We also performed an analysis of hallmark gene enrichment for all cell state positive markers ($\log_2FC > 0$), and six cell states have markers enriched from the hallmark gene list (see Supplementary Figure S3). Positive markers of antigen presentation and interferon alpha/beta response cell states are both enriched in the interferon gamma and alpha response hallmark gene sets, while markers of the mitotic cell state are enriched in the E2F targets, G2M checkpoint, and mitotic spindle hallmark gene sets (see Supplementary Table S2). These results are consistent with the original publication, demonstrating that our method can effectively identify signatures related to the main effect.

4.4.2 Identification of responding and non-responding signatures (main effect F2)

To identify gene signatures specific to either the responding group or the non-responding group, we examine the genes associated with the treatment factor (F2). From the identified gene markers associated with the main effect F2, we found 11 genes significantly up regulated in the responding groups and 5 genes significantly upregulated in the non-responding groups (see Figure 4.3a), the other genes showed only small $\log_2$ fold changes. Within sample cross-validation shows these genes are able to predict the treatment response label of pseudobulk samples with AUC of 0.976 (see Figure 4.3b). Figure 4.3c shows the expression of four example genes between the responding and non-responding groups. *TUBB4A* is significantly upregulated in the responding group, and has been linked to melanoma-associated pathways in a previous study Zhang et al. (2023a), suggesting its potential as a treatment response marker. Another example of upregulation in the responding group is *B2M* gene, previous study found it expressed in either tumor or stroma are associated with a positive response to immunotherapy in a metastatic melanoma patient cohort (Martinez-Morilla et al., 2021). We also found that *ABL2* and *APLP2* are the most significantly up-

regulated genes in the non-responding cell group (see Figure 4.3c). This aligns with a study showing *ABL1/2* drives resistance to BRAF/MEK inhibitors in melanoma Tripathi et al. (2020) and *APLP2* decreases HLA class I expression (Pandey et al. 2016), which is linked to de-differentiation and resistance to PD-1 inhibition (Lee et al. 2020). Additionally, Figure 4.3d shows that the expression patterns of these genes in the responding and non-responding groups are consistent across all cell states, suggesting that these genes are mainly associated with the treatment phenotype rather than cell state.

4.4.3 Identification of cell state specific responding signature (interaction effect $F1 \times F2$)

In addition to the main effect which represent overall response signature across all cell states, the power of iDAS is the ability to identify cell state specific signatures, which are formulated as gene signatures associated with the interaction effect in the iDAS framework. Figure 4.4a and 4.4b highlight two genes showing interaction effects between cell state and treatment phenotype. We found *PDGFRB* is highly expressed in the non-responding group of mesenchymal-like cells (Figure 4.4a and Supplementary Figure S4a), previous studies have shown PDGFR family play a significant role in brain, breast, colorectal , melanoma and lung cancers (Farooqi and Siddik 2015; Shi et al. 2011), suggesting that it may be potential marker for treatment response. Another example is the gene *WDR76* which is observed to be highly expressed in the responding group of mitotic cells (Figure 4.4b, Supplementary Figure S4b). Current literatures have shown that *WDR76* function as tumor suppressor in colorectal, bladder, colon and liver cancer (Ro et al. 2019; Jeong et al. 2019), showing that it may also be potential marker for treatment response in melanoma.

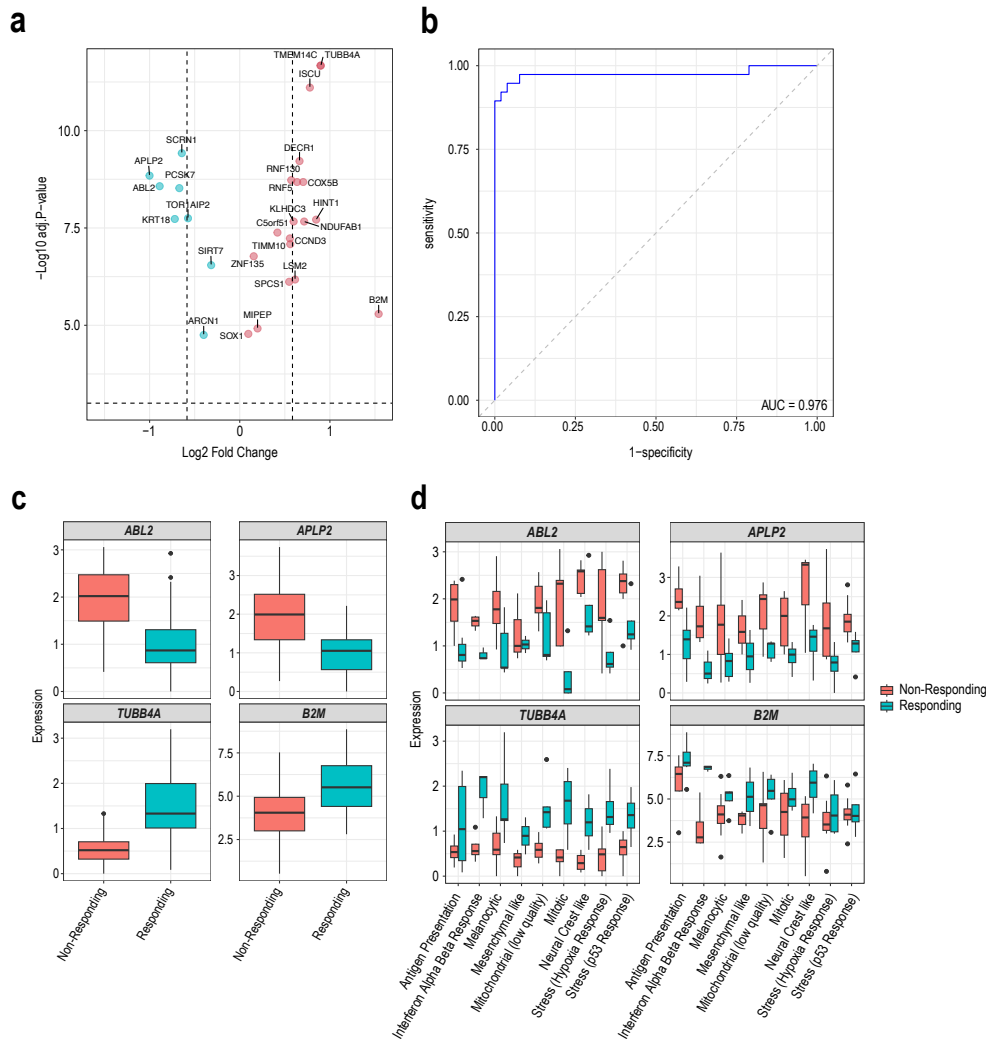


Figure 4.3: iDAS identifies the phenotype associated main effect signatures between phenotype and cell states. **(a)** Volcano plot illustrates the phenotype main effect signatures. The horizontal dashed line indicates the significance threshold at $p\text{-value} < 0.001$ ($-\log_{10}(p\text{-value}) > 3$). The vertical dashed lines mark the fold change thresholds at $\log_2 \text{fold change} > \log_2(1.5)$ and $\log_2 \text{fold change} < -\log_2(1.5)$. Upregulated genes in responding cell groups are colored red, and in non-responding cell groups are colored blue. **(b)** ROC curve evaluates the performance of a predictive model using phenotype associated main effect signatures to distinguish between responding and non-responding samples. The area under the curve (AUC) is 0.976. **(c)** The box plots display the expression levels of selected genes (*ABL2*, *APLP2*, *TUBB4A*, *B2M*) in responding and non-responding samples. **(d)** Boxplot of expression levels of the selected genes (*ABL2*, *APLP2*, *TUBB4A*, *B2M*) between the responding and non-responding groups.

4.5 IDENTIFICATION OF OVERALL AND CELL-SPECIFIC IMMUNOTHERAPY RESPONSE GENE SIGNATURES IN BULK EXPRESSION DATA FROM INDEPENDENT COHORTS

After demonstrating that the signatures are meaningful within the single-cell dataset, we aimed to see the translational potential of these signatures by exam-

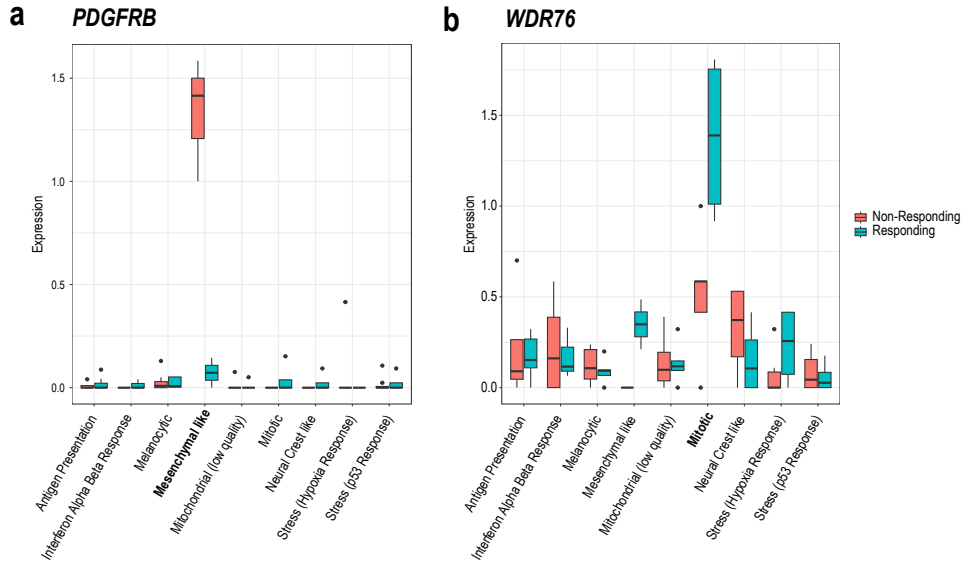


Figure 4.4: iDAS identifies the phenotype associated interaction effect signatures between phenotype and cell states. Boxplot of (a,b) are interaction effect signatures expression (*PDGFRB* and *WDR76*) across all cell states. (a) *PDGFRB* is highly expressed in non-responding groups of mesenchymal-like cells. (b) *WDR76* is highly expressed in responding groups of mitotic cells.

ining the fold-change of these markers in two other melanoma bulk RNA-seq data, Liu2019 (Liu et al., 2019) and Hugo2017 dataset (Hugo et al., 2016). In general, we note that the expression of these markers are not always high in the bulk samples. Thus, to further evaluate the predictive ability of phenotype-associated main effect signatures, we only selected genes with $\log_2\text{FC}$ greater than $\log_2(1.3)$ as features. Here, we built a generalized linear model to predict the phenotype of bulk samples. The AUC of 10-fold nested cross-validation is 0.675 from the Hugo2016 dataset (Figure 4.5a) and 0.595 from Liu2019 dataset (Figure 4.5b). This result suggests that the signatures discovered in the single-cell context are also predictive markers in a bulk RNA sequencing setting, reinforcing their potential utility as robust biomarkers.

To assess the cell state markers identified from the single-cell data in the bulk we examine the deconvoluted bulk RNA-seq data. We applied the BayesPrism method (Chu et al., 2022) to deconvolute the bulk RNA data into a cell state-level gene expression matrix. The examination the deconvoluted bulk RNA-seq data confirmed that the marker genes, highly expressed in specific cell

states in single-cell data, retain similar expression profiles in the deconvoluted bulk RNA data. Consequently, these markers are validated as conserved cell-markers across various datasets, as illustrated in Figure 4.5c and Supplementary Figure S5. We next examine the expression profile of the interaction effect signatures on the bulk deconvoluted dataset. Specifically, we examined the expression levels of *WDR76* and *PDGFRB* in the context of patient response to immunotherapy. Our results show that *WDR76* is slightly more highly expressed in the responding group than in the non-responding group in the mitotic cell state in the Hugo2016 dataset (see Supplementary Figure S6a). Its differential expression is more pronounced in the Liu2019 dataset (see Supplementary Figure S6c). Although supplementary Figure S6b shows a significant differential expression of *PDGFRB* in mesenchymal-like cells in the Hugo2017 dataset, the Liu2019 data does not support it as a non-responsive marker gene of mesenchymal-like cells (see Supplementary Figure S6d). Taken together, these results suggest that some of the cell state specific markers (interaction effect signatures) identified in the single-cell data are also reflected in the bulk data.

4.6 EXTENSIBILITY OF IDAS TO ACCOUNT FOR LONGITUDINAL STUDIES AND TO SPATIALLY RESOLVED OMICS DATA

The flexibility of iDAS model can be illustrated by its ability to handle multi-factor longitudinal data. For the same melanoma dataset, we can also model by accounting for pre- and on-treatment status for a given individual. Here, we define pre- and on-treatment as factor F1, treatment phenotype status (responding / non-responding) as factor F2, and cell state as factor F3. This three-way model allows us not only to identify the main effects associated with F1, F2, and F3, but also to detect two-way interaction effects, such as F1F2, F1F3, and F2F3. By running iDAS using a three-way random effects model, we identified a total of 350 genes that may be associated with one of the factors. We found 19 interaction effect signatures, including 18 three-way effects and 1 two-way effect, along with 331 main effect signatures (see Supplementary Table S3). The only two-way interaction effect gene is *NDUFA4L2*, which is a marker for the

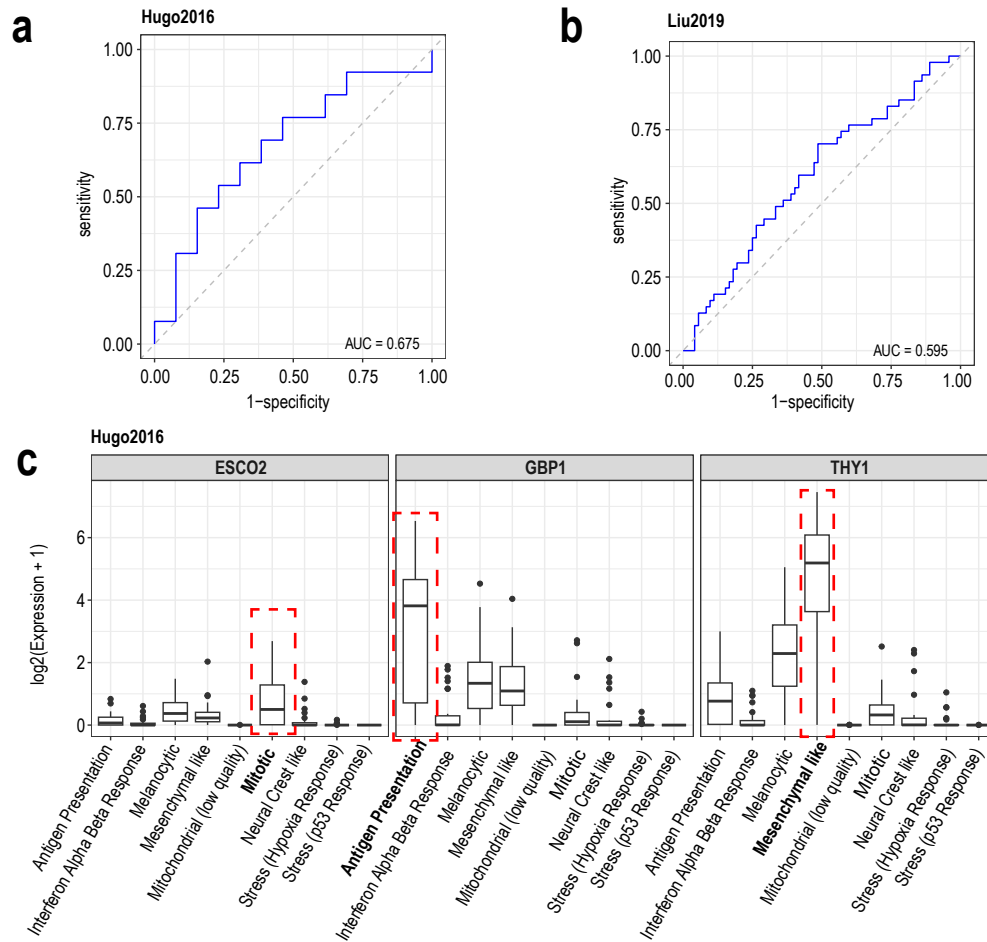


Figure 4.5: Main effect signature validation using independent cohort from bulk data (Hugo2016 and Liu2019). (a,b) AUC plots shows 10-fold cross-validation to predict bulk sample responses using gene signatures derived from iDAS, analyzed via generalized linear models. (c) shows cell state main effect signatures are transferable to external bulk data. The boxplot shows the expression of genes *ESCO2*, *GBP1* and *THY1*, mapped onto cell state specific deconvolutions of external bulk RNA dataset. Consistent with the observations in Figure 4.2b, these genes exhibit high expression in particular cell states, underscoring their role as specific markers across different datasets.

stress (hypoxia response) cell state in responding group, regardless of pre- or on-treatment status (see Figure 4.6). For the main effect signatures, one gene is associated with factor F1, 321 genes are associated with factor F3, and 9 genes are classified under additive effect categories. We found that the cell state-associated genes, specifically the positive markers of seven cell states (see Supplementary Figure S7), are enriched in the hallmark gene lists, with most of them consistent with the original publication (see Supplementary Table S4).

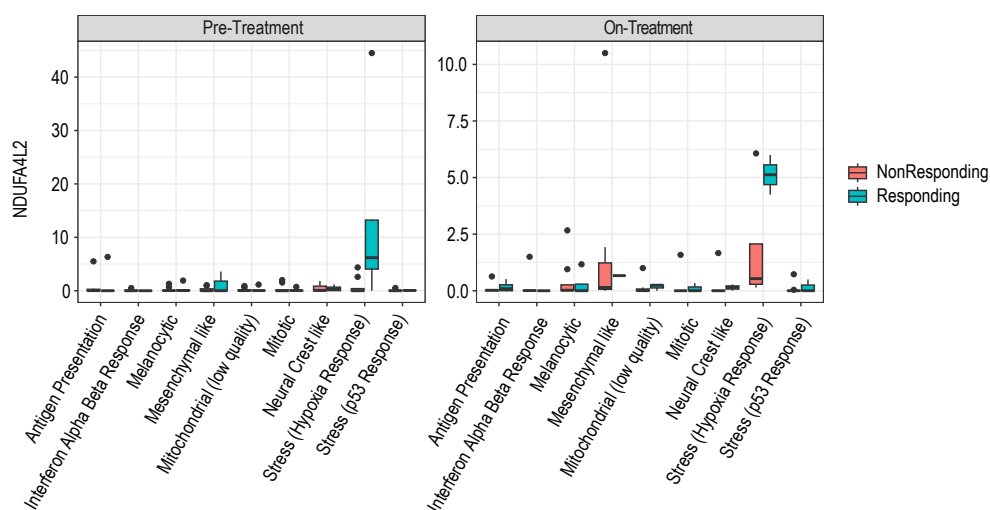


Figure 4.6: Application of iDAS to longitudinal data. Boxplot shows the expression levels of an interaction effect (F2F3) signature across various cell states in both pre-treatment and on-treatment conditions. *NDUFAL2* is highly expressed in stress (hypoxia response) specific cell state among responding groups.

Beyond the analysis of single-cell samples, iDAS framework can also be applied to examine spatially resolved transcriptomics data. For example, we used GHIST-predicted TCGA breast cancer spatial transcriptomics data to calculate the nearest neighbourhood correlation for each gene across all cell pairs. We then modeled the spatial metric as an input to investigate the spatial relationships of gene expression in relation to ER/PR status and AJCC pathological stage. As shown in Figures 4.7a and 4.7b, we found that the correlation of expression for *CLIC6* and *RAB30* in neighboring cells is higher in the ER+/PR+ group, while the correlation of expression for *KLRD1* and *LTB* is higher in the ER-/PR- group, regardless of whether the tumor is detected in the early or late stage. To explore tumor stage-associated genes, we discovered that correlation of expression for *PDGFRA* and *SOX17* in neighboring cells are higher in early-stage tumor patients but not in late-stage patients (see Figure 4.7c). Besides, we identified two genes, *ADIPOQ* and *CTH*, that exhibit a mixed pattern across tumor stages and ER/PR status. These genes are highly correlated in neighboring cells of ER+/PR+ patients in the early-stage group but not in ER-/PR- patients (see Figure 4.7d).

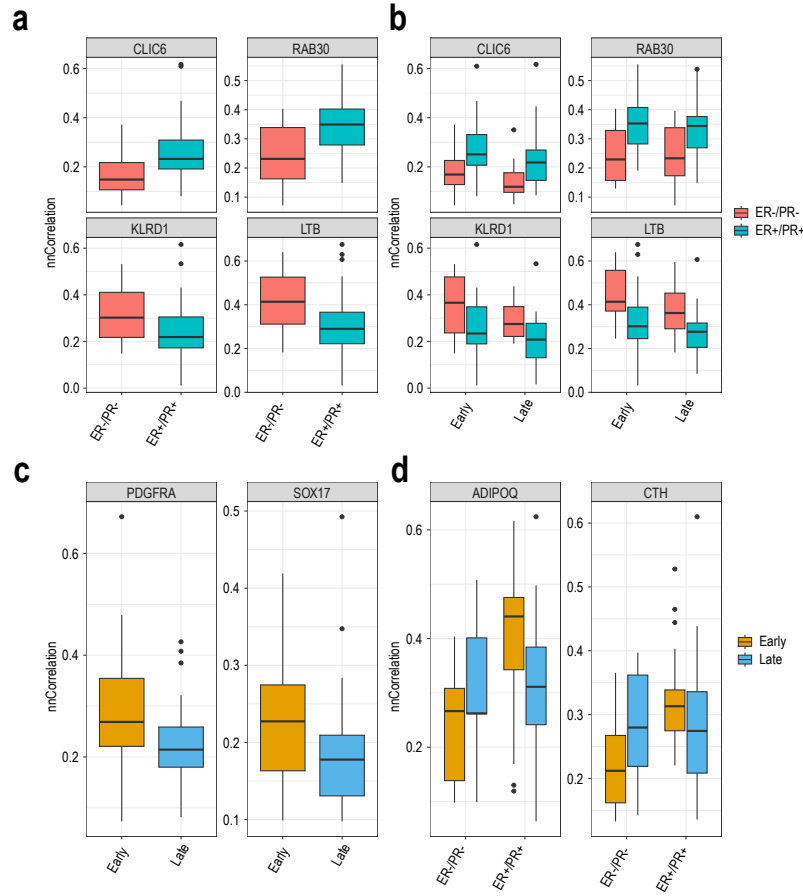


Figure 4.7: Application of iDAS to spatial data. **(a)** Boxplots showing nearest neighbourhood correlation of *CLIC6* and *RAB30*, indicating they are specificity to ER/PR status. These genes demonstrate a strong correlation with ER+/PR+ group than ER-/PR- group. **(b)** Boxplots showing that the nearest neighbourhood correlation of the same two genes in panel (a) is consistent across both early and late stages, indicating that their expression is not dependent on the stage of disease progression. **(c)** Boxplots illustrating the nnCorrelation of two stage-specific genes, showing differential expression between early and late AJCC pathological stage of the disease. **(d)** Boxplots depicting interaction effect between ER/PR status and AJCC pathological stage for *ADIPOQ* and *CTH*, with varying expression levels depending on both factors.

4.7 DISCUSSION AND CONCLUSION

In this study, we introduced iDAS, an ANOVA-based post-DA framework designed to interpret gene signatures from differential abundance analysis. Using a Melanoma dataset, we demonstrated that iDAS effectively identifies both over-all responding signature (main effect) and cell-state specific responding signature (interaction effect). We identified cell state-specific signatures that have

the power to distinguish between different cell states. We found upregulated and downregulated genes in both responding and non-responding groups, indicating potential markers for treatment response and non-response. Signatures were further validated on bulk datasets, demonstrating the transferability of the signature selected by iDAS. Lastly, we demonstrate the flexibility of the model by applying iDAS to more complex study design including longitudinal data with pre- and on-treatment measurements from the same individual using three-way analysis and in-silico spatially resolved transcriptomics data.

One of the strengths of iDAS's results is that the gene signatures found in a smaller single-cell dataset are transferable to a larger bulk RNA-sequencing dataset demonstrating its translational capacity. Validating single-cell-derived gene signatures in bulk RNA-seq data is essential because bulk platforms are often more accessible for translational applications. Demonstrating the transferability of these signatures highlights their potential to bridge experimental single-cell studies with broader translational applications. However, the validation strategies associated with both overall and cell-type specific markers are imperfect as they may depend on the difference between platforms as well as the performance of deconvolution approaches. For example, while our study shows very similar expression trends between single-cell and bulk datasets, we notice that some genes of interest found in single cells are not "highly expressed" in bulk. Hence validation across platforms should condition on "expressed genes". In addition, the actual fold-change estimate in the deconvoluted bulk dataset is much lower than single-cell data. Possible explanations include masking of signals from rare cell states (Ziegenhain et al., 2017), cell states in the reference differing from the bulk data (Avila Cobos et al., 2018), and lack of raw counts as input from bulk samples. Thus, improving deconvolution methods will increase the reliability of translating single-cell insights to bulk RNA-seq data, enabling more accurate validation. It is also important to note that in practice deconvolution is not required for cell state implementation as there exists other biotechnological platform strategies to select appropriate cell state.

Besides, there is still a need for some further work on the transfer signature validation approach. When performing external bulk validation, we train a new model with the signature expression profile of bulk data to predict the treatment phenotype on the same data. However, if we train the model on a single cell signature profile and apply the model to predict phenotype in bulk data, it will not be successful. An obstacle to the transfer single-cell model approach is the batch effect between different datasets due to the different normalisation methods used in different studies. To address this issue, we can use the log-ratios method from CPOP (Wang et al., 2022) to convert RNA normalised counts from different studies to the same ratio scale, which may help in using the single-cell model to predict the result of the bulk data. Future research is needed to investigate the effectiveness of this approach.

The iDAS framework offers much greater flexibility in handling complex experimental designs than pairwise DE analyses. First of all, ANOVA within iDAS can evaluate multiple groups simultaneously compared to pairwise DE analysis. In addition, ANOVA can detect interaction effects between factors, whilst pairwise analyses generally lack this ability, unless they are specifically designed to explore such interactions. Furthermore, our framework can be easily extended to longitudinal studies. We can easily include random effects in the models when the data includes patients with pre- and post-treatment information. This addition provides a more accurate understanding of each effect in the model between different experimental factors.

In summary, to enhance the interpretability of differential abundance signatures, we introduce an ANOVA-based framework, iDAS, that is able to identify the signature from multiple conditions at once. Our findings confirm that iDAS framework can effectively identify robust cell-state specific gene signatures, and the framework is flexible to handle various complex study designs including multi-factor longitudinal and spatially resolved data.

CONCLUSION

This thesis explores and addresses several challenges by developing computational approaches for data analysis on single cell omics data.

Firstly, we addressed the lack of systematic evaluation of scRNA-seq data clustering methods for estimating the number of cell types in Chapter 2. This work allows scientists to efficiently select clustering methods with better performance in clustering concordance and cell type estimation. It also aids developers in identifying limitations in existing methods. We found that concordance of cell clusters did not always correlate with accurate estimation of cell types. It suggests scientists pay attention to evaluating which clustering methods should be used when cell type detection is the goal of scRNA-seq data analysis. This project was done in 2021 during the COVID period, we did not include any Python packages of deep learning clustering methods for benchmarking due to the limited computing resources I had access to at the time. With the rapid development in this field, there are already several review papers discussing the topic of deep learning clustering (Karim et al., 2020; Min et al., 2018). One question that remains unaddressed in our work is if deep learning based clustering methods outperform other commonly used single cell clustering methods in any situation (e.g. multimodal omics data, spatial data, large-scale data). Therefore, the comparison between traditional clustering methods and deep learning based clustering methods needs to be further explored. To address this, a large-scale benchmarking study could be designed to compare traditional methods (e.g., Seurat, SC3) with deep learning methods (e.g., scVI, totalVI, GNN-based tools) using datasets of varying complexity. By evaluating their performance on multimodal, spatial, and large-scale data, this work can identify where deep learning methods excel and where traditional approaches are still sufficient. In

addition, creating an open-access benchmarking framework would also encourage the research community to contribute new methods and datasets, driving ongoing improvements in clustering approaches.

Secondly, in Chapter 3, we developed SnapCCESS, an ensemble deep learning method that efficiently integrates different data modalities into embeddings, for consensus clustering of multimodal single cell omics data. An innovation in SnapCCESS is the use of the snapshot ensemble learning technique. It reduces the training time of the ensemble model to generate multiple embeddings compared to traditional ensemble deep learning methods. This work aims to inspire the future development of ensemble deep learning methods for multimodal single cell omics data analysis. Besides, an interesting question in single cell clustering is how to develop multi-scale or multi-resolution clustering to systematically annotate single cells by different levels of similarity. Several methods have addressed this issue by developing methods to automatically estimate the number of clusters at multiple resolutions in single cell RNA sequencing datasets, such as IKAP (Chen et al., 2019b) and MultiK (Liu et al., 2021). However, this problem remains unresolved for multimodal single cell omics data, presenting an opportunity for further methodological development in this area. The integration of multi-omics data, particularly through methods that generate embeddings, captures shared biological signals across modalities. It offers a foundation to extend scRNA-seq clustering approaches to multimodal datasets. Once the cluster numbers (k) are estimated, they enable researchers to uncover the structure and heterogeneity of cell populations, aiding in identifying novel subpopulations or states. Benchmarking studies, such as the one presented in this thesis chapter 2, provide practical guidance on choosing methods for estimating k . By combining multi-omics embeddings with modifications to existing scRNA-seq clustering methods, it becomes possible to adapt these approaches for multimodal data and pave the way for more accurate and biologically meaningful clustering at multiple resolutions.

Finally, we developed iDAS, an interpretable differential abundance (DA) signature framework and demonstrated its capacity in a case study using single

cell melanoma data. We used ANOVA-like hypothesis testing to separate gene signatures that are common to all individuals from those that are specific to a particular cell type or cell state. We validate iDAS signatures on external datasets to demonstrate that this method can capture transferable features. To our knowledge, different DA methods can produce different cell state results due to their different design approaches. To improve the results of the iDAS gene signature, further research should be conducted to develop an intermediate step to harmonise cell states generated by different methods and to obtain stable gene signatures across DA methods. One approach to improve the stability of cell state labels is to apply AdaSampling (Yang et al., 2019). It can iteratively reclassify cell state labels to reduce noise and correct misclassification, leading to more consistent and reliable cell states across different DA methods.

This thesis demonstrates a cohesive progression from clustering single cells to integrating multi-omics data and finally to gene signature classification, addressing key challenges towards extracting actionable insights from a typical single-cell data workflow. The clustering methods in Chapter 2 form the foundation by systematically benchmarking and evaluating approaches for estimating the number of cell types. This provides essential guidance for selecting appropriate clustering methods for scRNA-seq data. Building on this, Chapter 3 introduces SnapCCESS, an ensemble deep learning approach that integrates multi-omics data to generate embeddings for clustering. This innovation reduces training time through snapshot ensemble learning and extends clustering capabilities to multimodal datasets, addressing the complexity of integrating diverse data modalities. In Chapter 4, the classification of gene signatures using iDAS builds upon the ability to identify cell types established in Chapters 2 and 3. Together, these chapters form a unified narrative: clustering provides the structural insights needed for single-cell data, multi-omics integration broadens the scope of analysis, and gene signature classification builds upon the ability to identify cell types and translates these insights into actionable biological and clinical knowledge. The methods and frameworks developed across these chapters not only tackle technical challenges but also deliver practical tools for advancing disease research and informing precision medicine.

In Chapters 2 and 3 of this thesis, we assume that all cell types are discrete. However, an important aspect of single-cell analysis is recognising that cells don't always fall into clear-cut categories based on gene expression patterns. Traditional approaches often aim to identify distinct cell types, but this can overlook the continuous nature of cellular states and fates seen in many biological systems. These states and fates are dynamic, and shaped by factors like development, the environment, and disease. This perspective emphasizes that although clustering is a powerful tool, it is not the only method for understanding a cell's status. Approaches that incorporate trajectory analysis, such as pseudotime methods, provide complementary insights into the complex landscape of cellular behaviour. Recognising these differences is critical for interpreting results from clustering-based methods, such as those developed in this thesis.

In recent years, there has been a proliferation of spatial transcriptomic data at different levels of resolution. Knowledge generated from single cells, such as cell type information and its association with histology or spatial transcriptomics, can contribute to the understanding of cell-cell interactions and further help scientists to understand the microenvironment of disease response to therapy. There are specific problems with spatial cell clustering that have not been fully addressed. One problem is that cells can overlap in the image, making it difficult to distinguish individual cells. Another problem is the different shapes of cells, such as nerve cells, which are very difficult to segment correctly. A benchmark study showed that cell clustering methods using location information could not always improve clustering compared to methods using gene expression data alone (Cheng et al., 2022). This highlighted the challenge of how effectively spatial information is being used for cell clustering. The ensemble deep learning clustering described in this thesis may provide insights for future method development to improve clustering on spatial transcriptomics data. We believe that the ensemble technique can better handle the cell shape and overlap problem by integrating different models that can focus on detecting the different types of cell boundaries.

In addition, with the increasing use of sequencing technologies in clinical settings, scientists have access to a large amount of patient bulk data with treatment information. There has also been a proliferation of single cell data sets at the atlas level over the last few years. A key question these two types of datasets pose is how to map knowledge from the single-cell level to the bulk patient level. A potential application of such a mapping arises when scientists have both atlas level single cell data and cell response information on multiple drugs, thereby benefitting clinicians who can use the single cell information by using it as a reference to estimate a patient's drug response with bulk data, further guiding the treatment on the patient. This is particularly similar to the external validation we try to illustrate in Chapter 4. Thus, as we discussed in Chapter 4, there is an urgent need for scientists to develop improved methods of bulk deconvolution or classification models to guide a better transfer of such treatment knowledge between different types of samples. From the work presented in this thesis, several contributions can help address this need. First, the accurate estimation of the number of cell types using clustering algorithms, as evaluated in Chapter 2, can improve the single-cell reference profiles used for bulk deconvolution. Second, integrating multi-omics datasets using methods like SnapCCESS, discussed in Chapter 3, improves cell type identification by using multi-omics features instead of relying solely on scRNA-seq. This also helps increase the resolution and accuracy of bulk deconvolution results. Lastly, iDAS in Chapter 4 helps identify meaningful and interpretable gene signatures. This improves the signal used to characterize bulk cell types, ensuring that important cell type-specific features from single-cell data are accurately captured in bulk analyses.

Going forward, with increased efforts contributing to the analysis of single cell data, I believe that the knowledge gained from the large-scale single cell data can be translated into clinical settings. Examples include modelling the association between digital pathology images and single cell spatial data, and using multiple modalities in single cell data to better predict disease treatment response. We expect that contributions from data science, such as using spatial information to improve cell type clustering, identifying signatures associated

with complex studies, and more, will be critical to advancing the understanding of disease mechanisms and individualised healthcare.

APPENDIX FOR CHAPTER 2

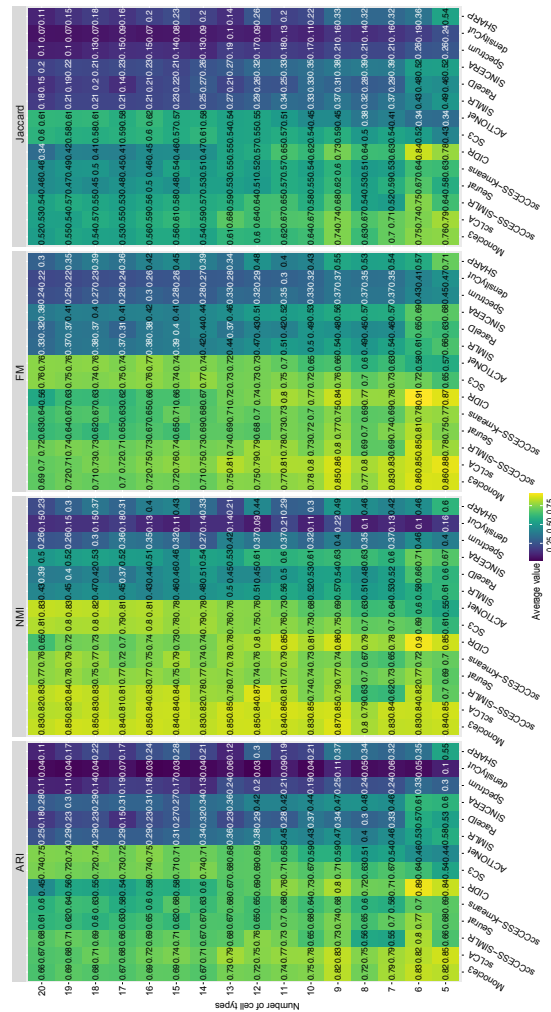


Figure S1: Concordance of clustering output and pre-defined cell type labels as quantified by four concordance measures, true number of cell types ranges from 5 to 20. Each tile represents the average score across 10 tests.

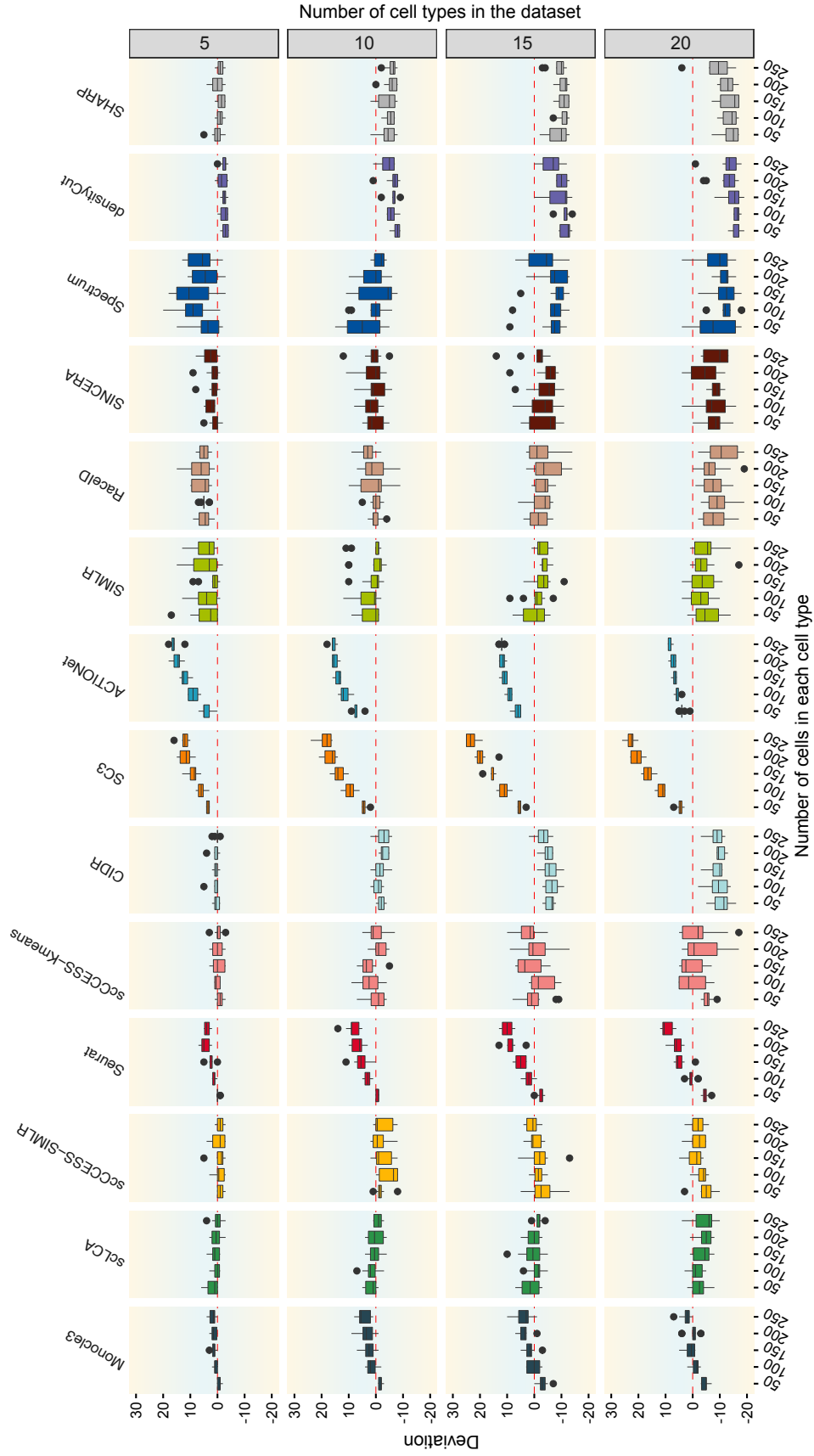


Figure S2: Deviation of the estimated and the true number of cell types for each of the 14 clustering methods on datasets with different number of cell types (i.e. 5, 10, 15, and 20) and different number of cells in each cell type (i.e. 50, 100, 150, 200, and 250). Each combination was repeated 10 times for estimating variability.

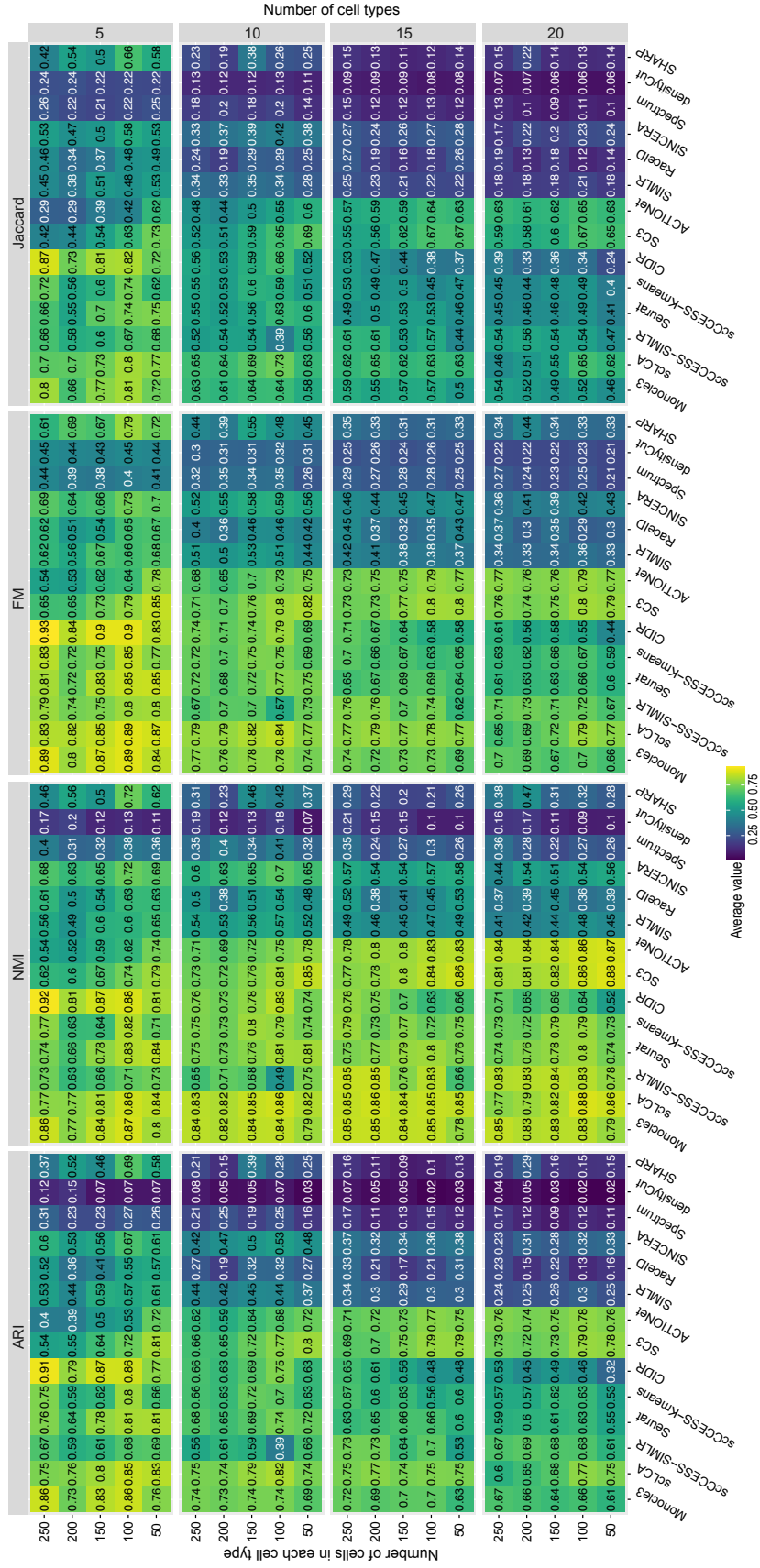


Figure S3: Concordance of clustering output and pre-defined cell type labels as quantified by four concordance measures with different true number of cell types ranges (i.e. 5, 10, 15, and 20) and different number of cells in each cell type (i.e. 50, 100, 150, 200, 250). Each tile represents the average score across 10 tests.

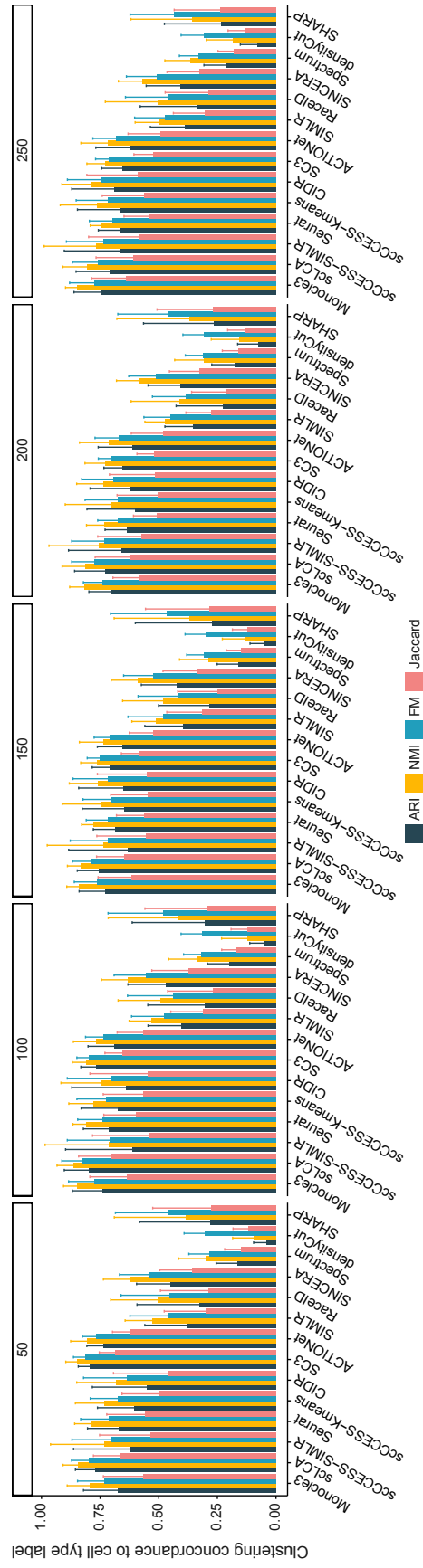


Figure S4: Concordance of clustering output and pre-defined cell type labels quantified by four concordance measures for datasets with different number of cells in each cell type (i.e. 50, 100, 150, 200, and 250). Error bars represent the standard deviation.

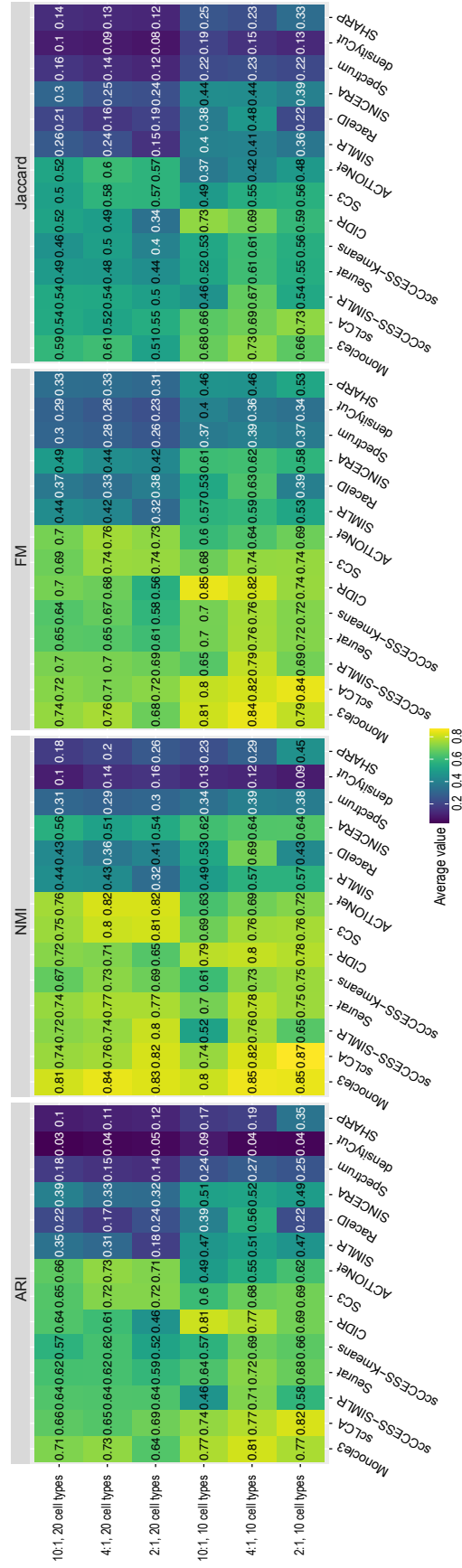


Figure S5: Comparison of cell clustering performance of 14 methods on datasets with different imbalance ratio (i.e., 2:1, 4:1, and 10:1) and different number of cell types (i.e., 10 and 20). Each tile represents the average score across 10 tests.

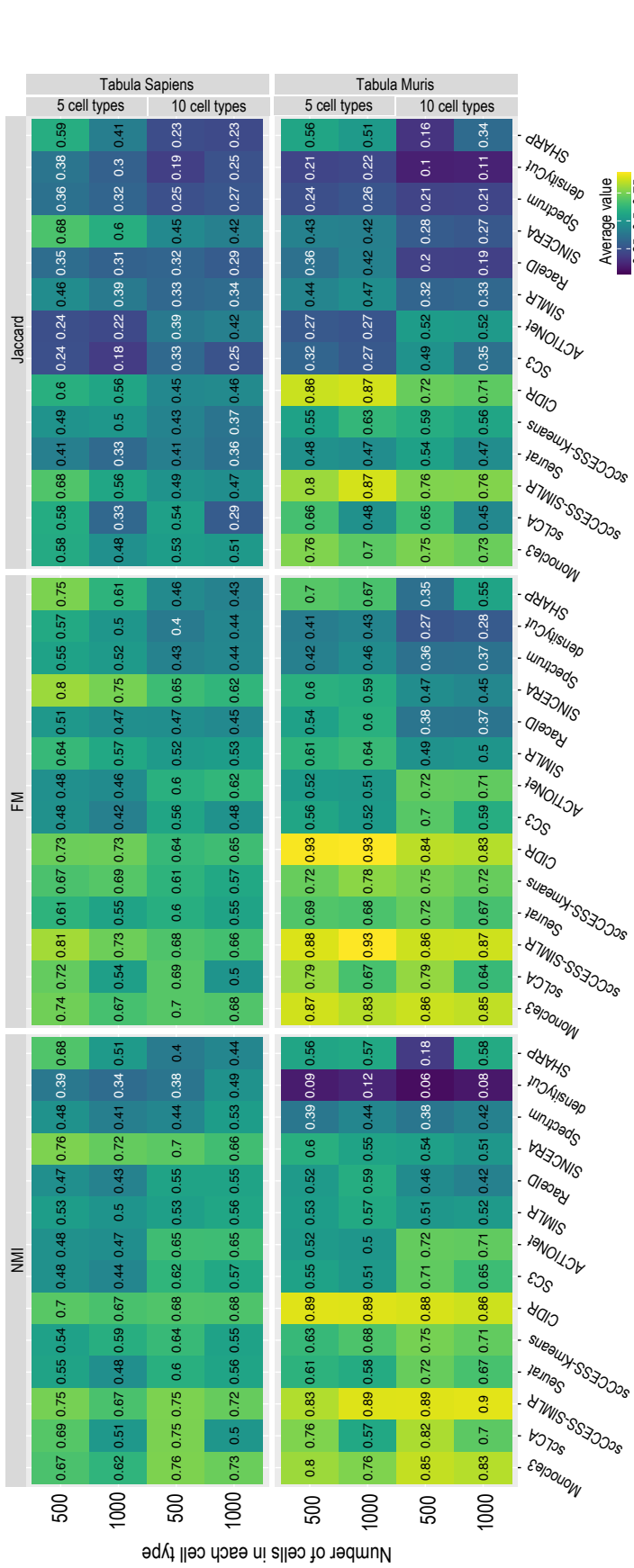


Figure S6: Comparison of cell clustering performance of 14 methods on datasets with large numbers of cells sampled from Tabula Sapiens and Tabula Muris. Each tile represents the average score across 10 tests.

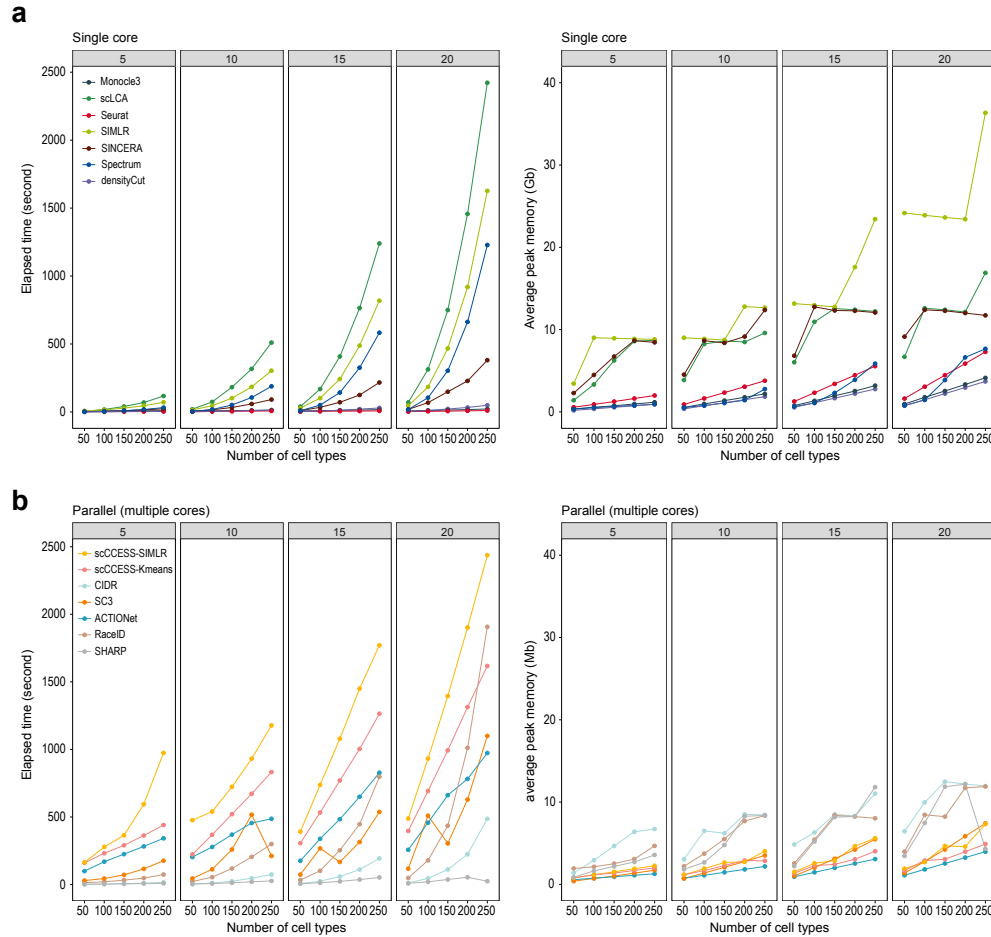


Figure S7: Benchmark of the elapsed time and peak memory usage on datasets with different number of cells (i.e., 50, 100, 150, 200, 250) and different number of cell types (i.e., 5, 10, 15, and 20). **(a)** The running time and peak memory usage of methods that uses only a single CPU core (i.e., densityCut, scLCA, SIMLR, Monocle3, Seurat, Spectrum, and SINCERA). **(b)** The running time and peak memory usage of methods that employ multi-cores for parallel computing (i.e., ACTIONet, RaceID, scCESS-Kmeans, SHARP, CIDR, SC3 and scCESS-SIMLR).

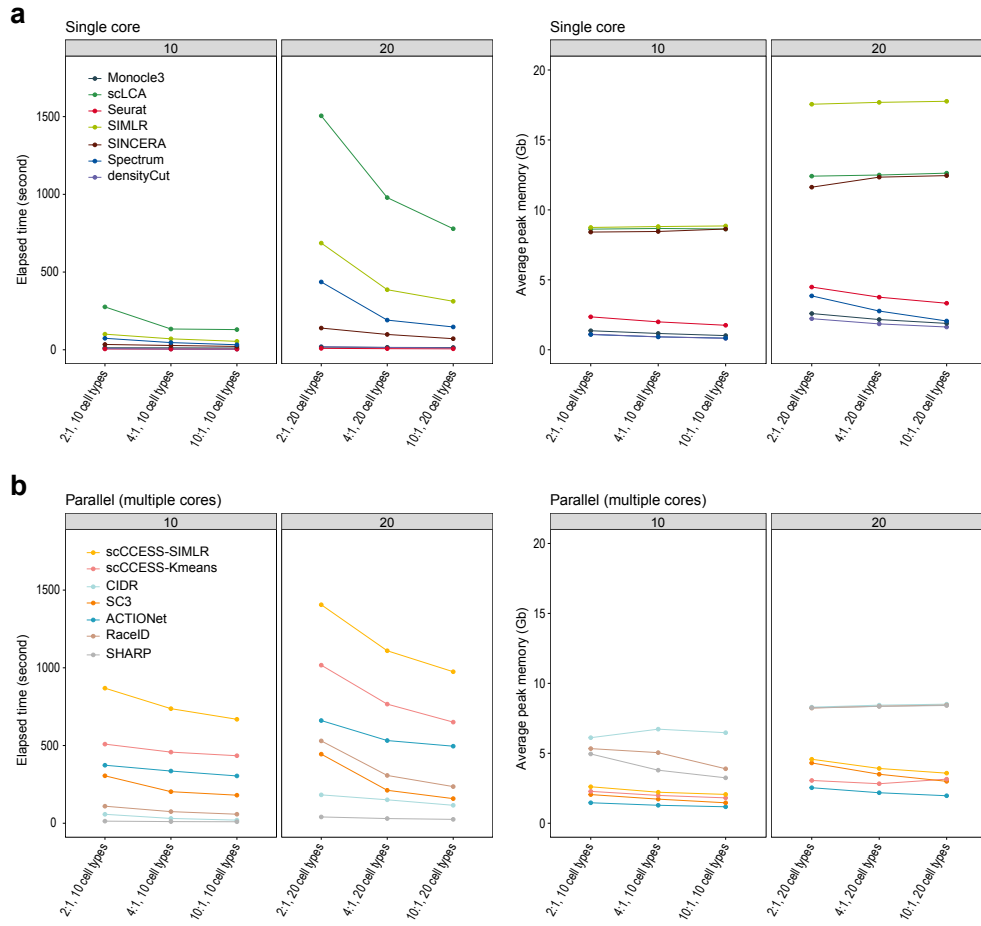


Figure S8: Benchmark of the elapsed time and peak memory usage on datasets with different imbalance ratios (i.e., 2:1, 4:1, and 10:1) and different number of cell types (i.e., 10 and 20). **(a)** The running time and peak memory usage of methods that uses only a single CPU core (i.e., densityCut, scLCA, SIMLR, Monocle3, Seurat, Spectrum, and SINCERA). **(b)** The running time and peak memory usage of methods that employ multi-cores for parallel computing (i.e., ACTIONet, RaceID, scCCESS-Kmeans, SHARP, CIDR, SC3 and scCCESS-SIMLR).

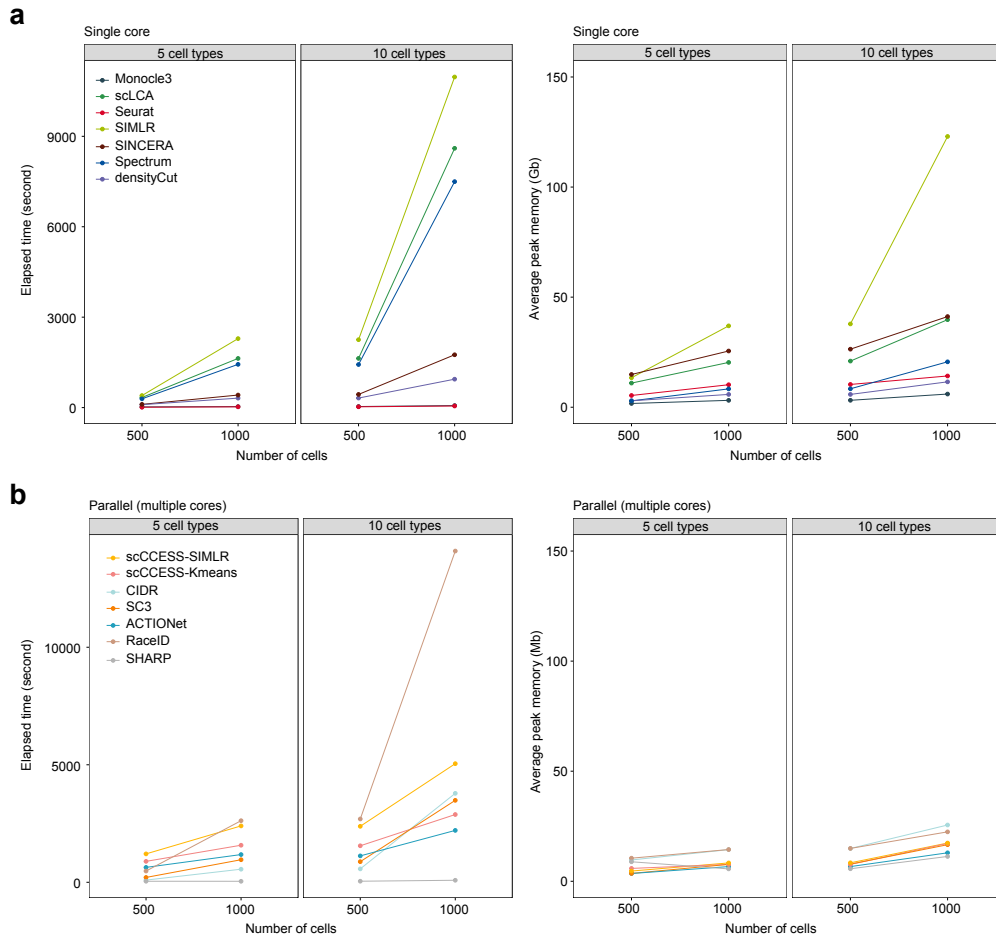


Figure S9: Benchmark of the elapsed time and peak memory usage on datasets with large number of cells. **(a)** The running time and peak memory usage of methods that uses only a single CPU core (i.e., densityCut, scLCA, SIMLR, Monocle3, Seurat, Spectrum, and SINCERA). **(b)** The running time and peak memory usage of methods that employ multi-cores for parallel computing (i.e., ACTIONet, RaceID, scCCESS-Kmeans, SHARP, CIDR, SC3 and scCCESS-SIMLR).

APPENDIX FOR CHAPTER 3

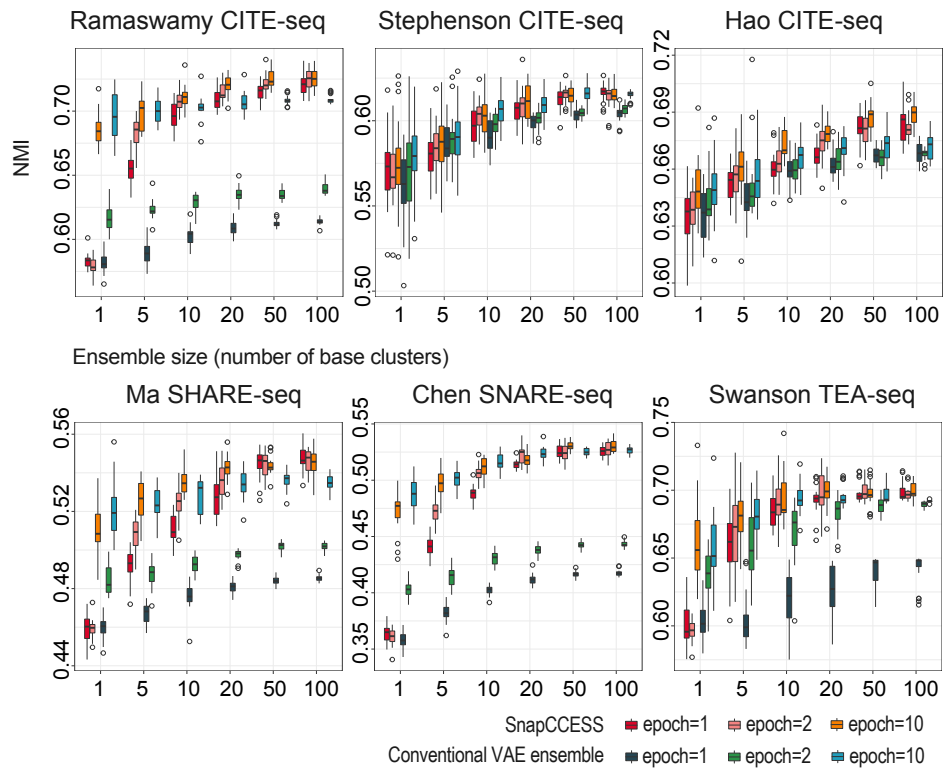


Figure S1: Concordance of cell type clustering on six multimodal single-cell omics data. The x-axis is the number of base clusters for the ensemble and the y-axis is the concordance of the clustering output and the cell type annotation in the original studies quantified by NMI. The k-means clustering algorithm was used for clustering the embeddings generated from each method. The entire procedure was repeated 20 times for capturing the performance variability.

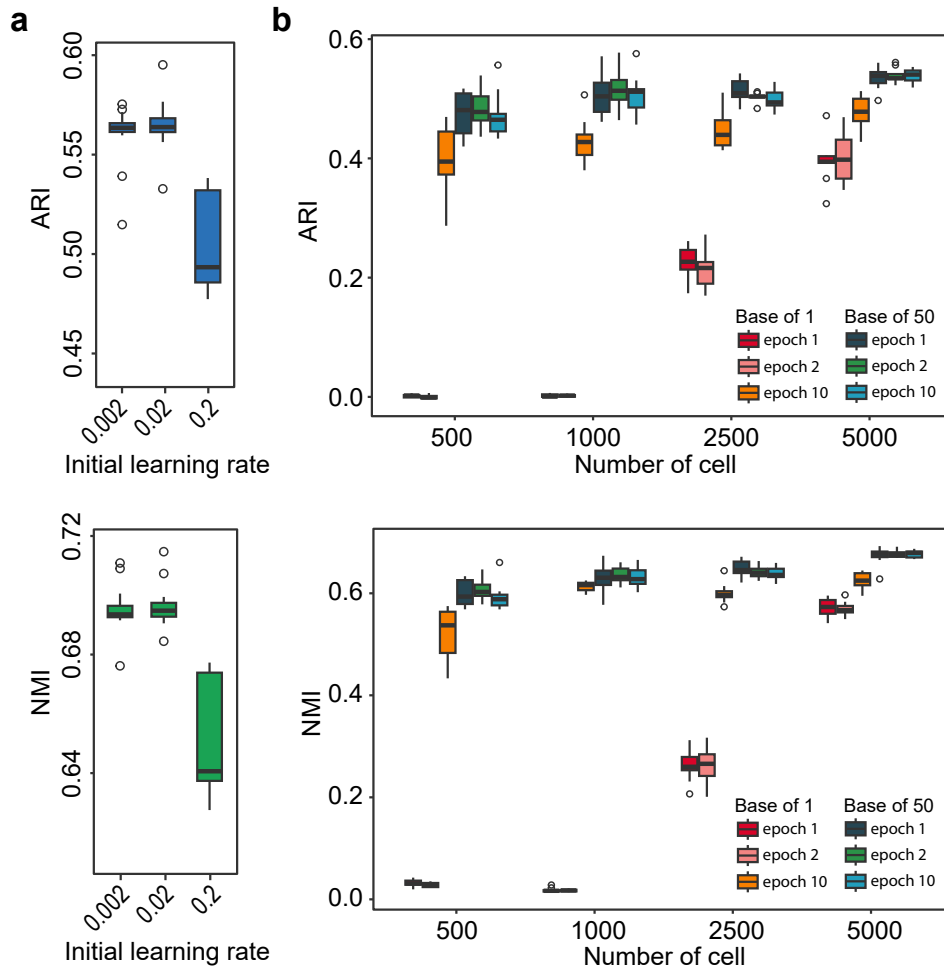


Figure S2: Evaluation of performance on (a) varying initial learning rates, and (b) the numbers of cells using the TEA-seq dataset by Swanson et al. (2021). Concordance of clustering output and the pre-defined cell type labels were quantified using ARI and NMI.

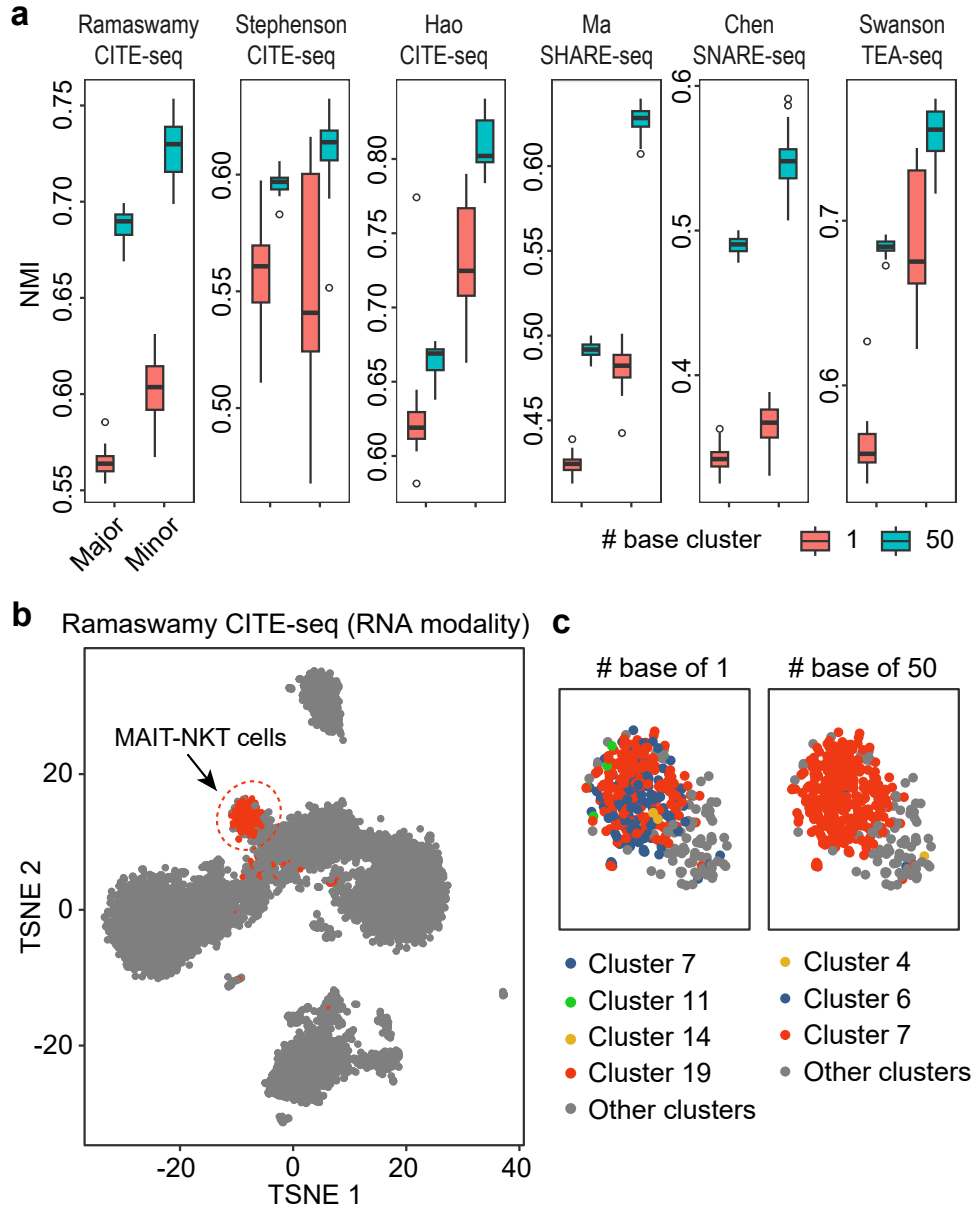


Figure S3: Evaluation of SnapCCESS (epoch=1) performance on **(a)** major and minor cell type clustering using each of the six datasets. Concordance of clustering output and the pre-defined cell type labels were quantified using NMI. **(b)** TSNE visualisation of Ramaswamy CITE-seq data using RNA modality. MAIT-NKT cells are highlighted in red. **(c)** Zoom in of MAIT-NKT cell clustering using SnapCCESS with 1 or 50 base clusters.

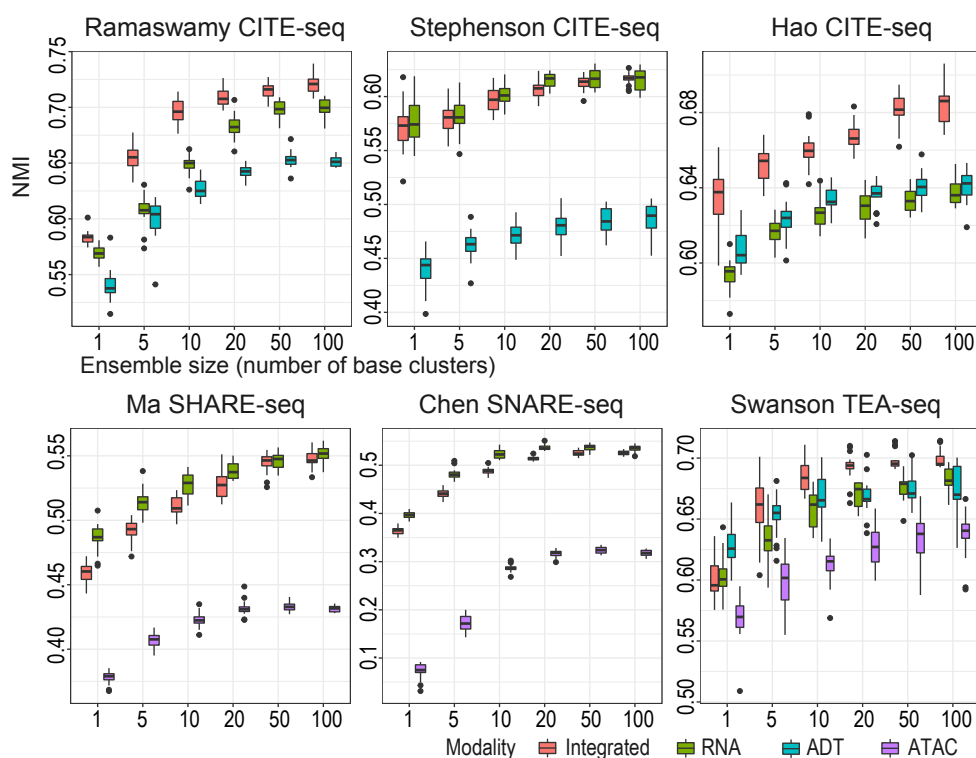


Figure S4: Comparison of integrated embedding of multimodality learned with those from unimodality using the k-means clustering algorithms. Concordance of cell type clustering quantified by NMI on six multimodal single-cell omics data using SnapCCESS generated embeddings from either all modalities in a dataset or each data modality alone. The entire procedure was repeated 20 times for capturing the performance variability.

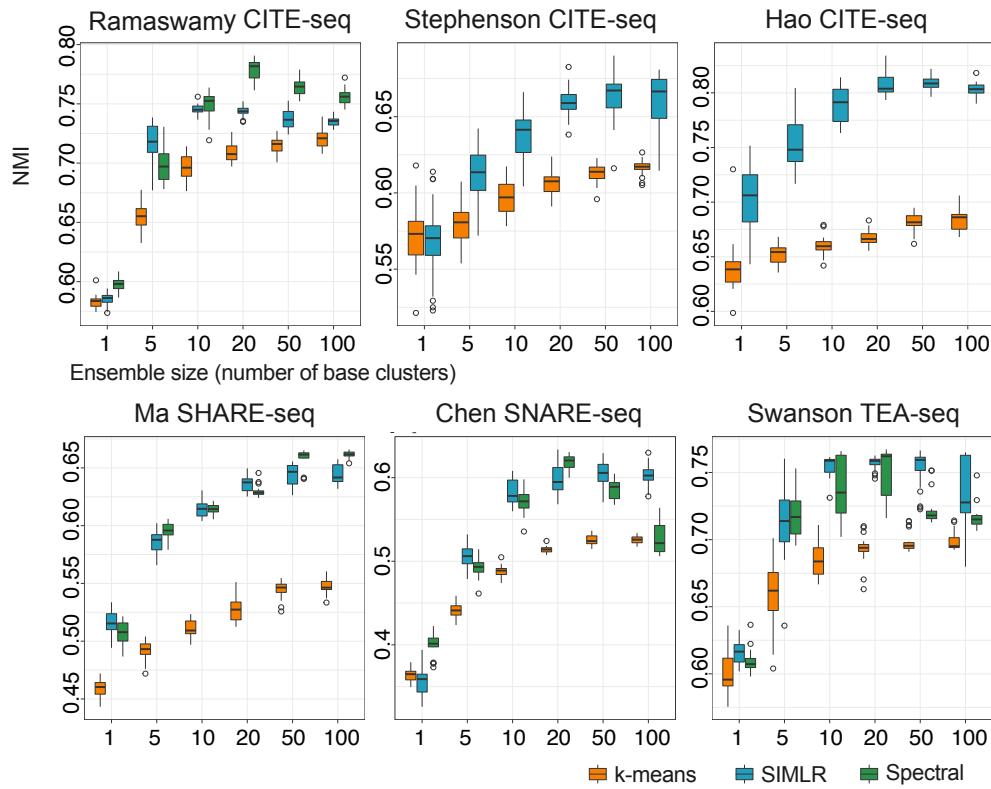


Figure S5: Evaluating the utility of the SnapCCESS framework with alternative clustering algorithms on multimodal single-cell omics data. Concordance of the cell type annotations and cell clustering output from each clustering algorithm. The x-axis is the number of base clusters in the ensemble and the y-axis is the concordance quantification by NMI. This procedure was repeated 20 times for capturing the performance variability.

APPENDIX FOR CHAPTER 4

C.1 SUPPLEMENTARY FIGURES

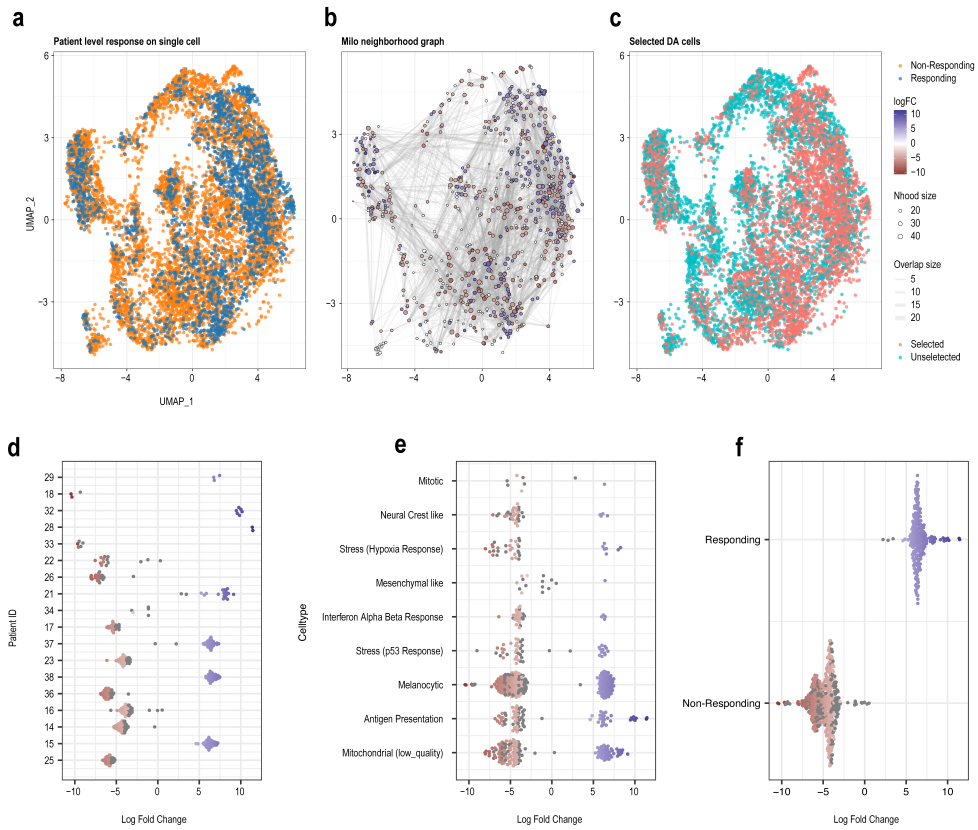


Figure S1: Using Milo to identify differentially abundant cells. **(a)** UMAP plot of patient-level response in the single-cell dataset. **(b)** Milo identified differentially abundant (DA) neighborhood with a Spatial FDR ≤ 0.2 . **(c)** UMAP plot of selected single cells from Milo DA neighbourhood groups with a Spatial FDR ≤ 0.2 and $\log_2$ fold change ≥ 5 . **(d)** Beeswarm plot visualizes the distribution of $\log_2$ fold changes across neighbourhood annotations of patient ID. **(e)** Beeswarm plot shows the distribution of $\log_2$ fold changes across cell states. **(f)** Beeswarm plot illustrates the distribution of $\log_2$ fold changes across phenotype.

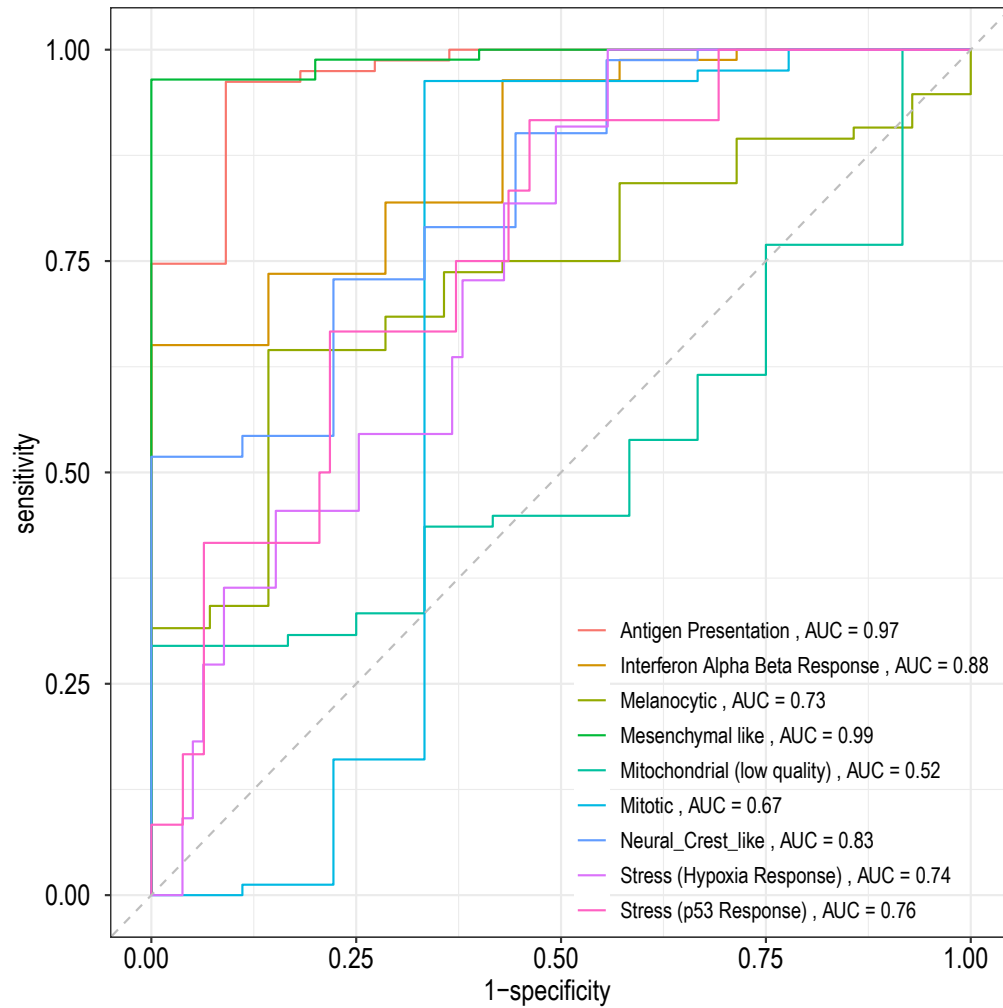


Figure S2: ROC curves for cell state-specific gene signatures. The AUC values for each signature are as follows: Antigen Presentation (AUC = 0.97), Interferon Alpha Beta Response (AUC = 0.88), Melanocytic (AUC = 0.73), Mesenchymal-like (AUC = 0.99), Mitochondrial (low quality, AUC = 0.52), Mitotic (AUC = 0.67), Neural Crest-like (AUC = 0.83), Stress (Hypoxia Response, AUC = 0.74), and Stress (p53 Response, AUC = 0.76). The dashed line represents the line of no discrimination (AUC = 0.5).

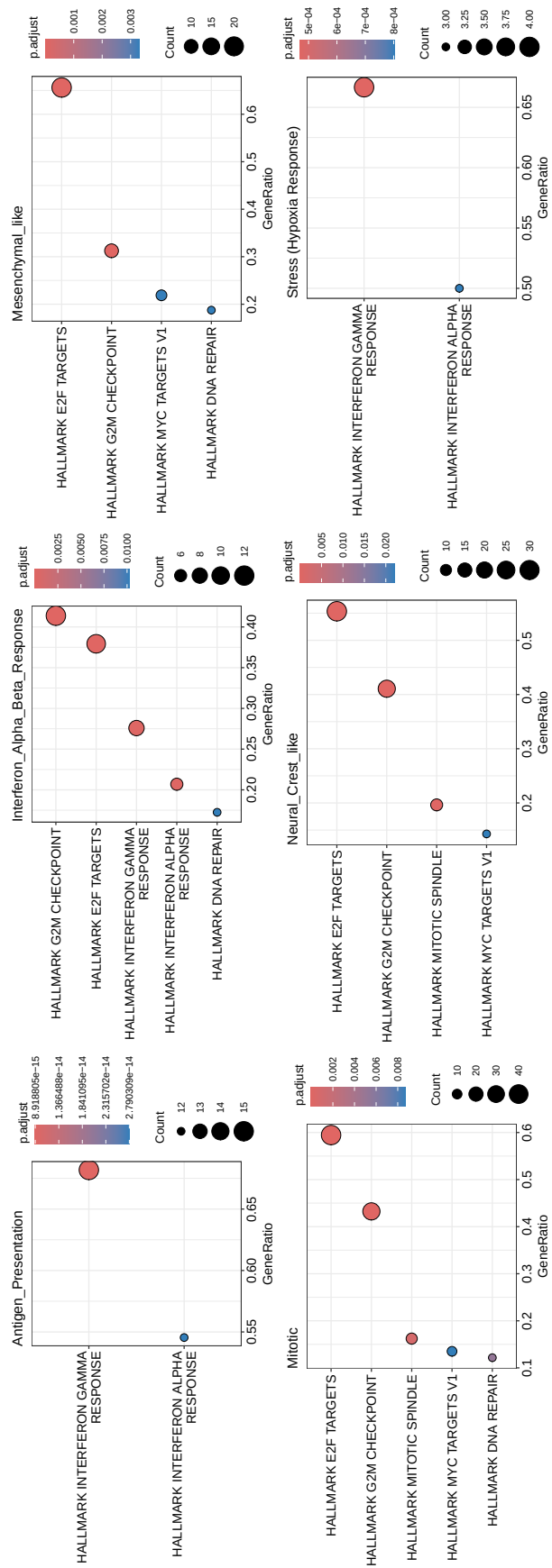


Figure S3: Hallmark gene set enrichment analysis of cell state-specific signatures (positive marker) obtained from the two-way iDAS model. Hallmark gene enrichment on six cell states. The gene ratio for each hallmark is plotted along the x-axis, and the size of each dot represents the count of genes in the corresponding gene set. Colors indicate the adjusted p-values, with a gradient from red (most significant) to blue (least significant).

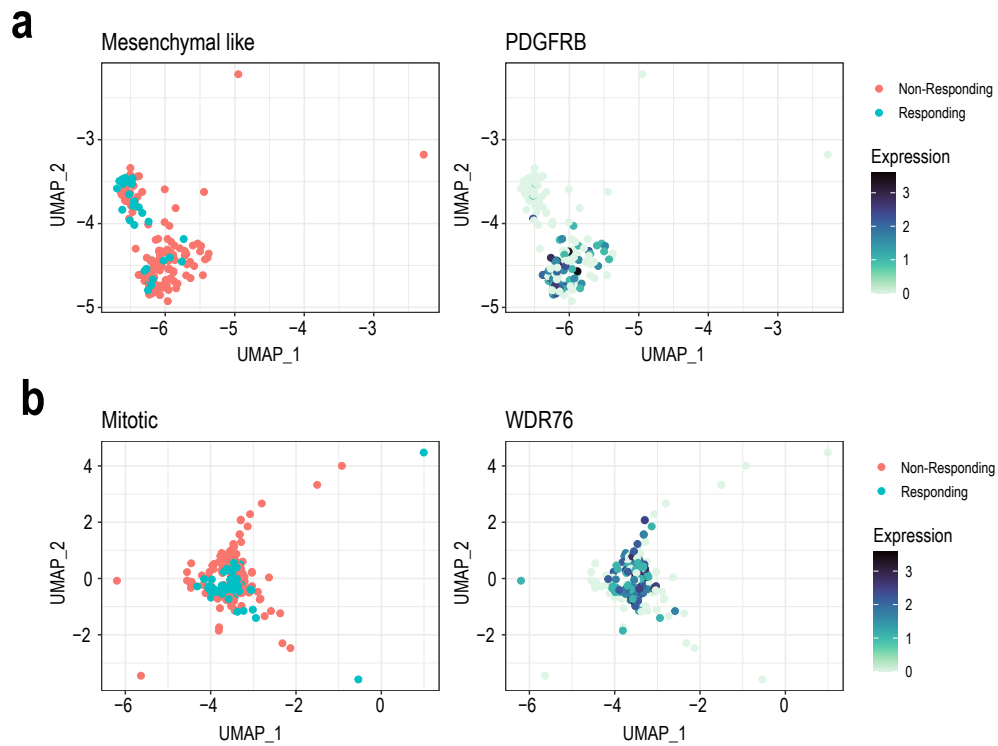


Figure S4: iDAS identifies the interaction effect signatures between phenotype and cell states. Cell state specific phenotype signatures were selected from mesenchymal-like cells and mitotic cells. **(a,b)** UMAP plots of mesenchymal-like cells and mitotic cells, respectively, with patient-level phenotype colored in each single cell in left panels and gene expression values of PDGFRB and WDR76 within their respective cell states in right panels.

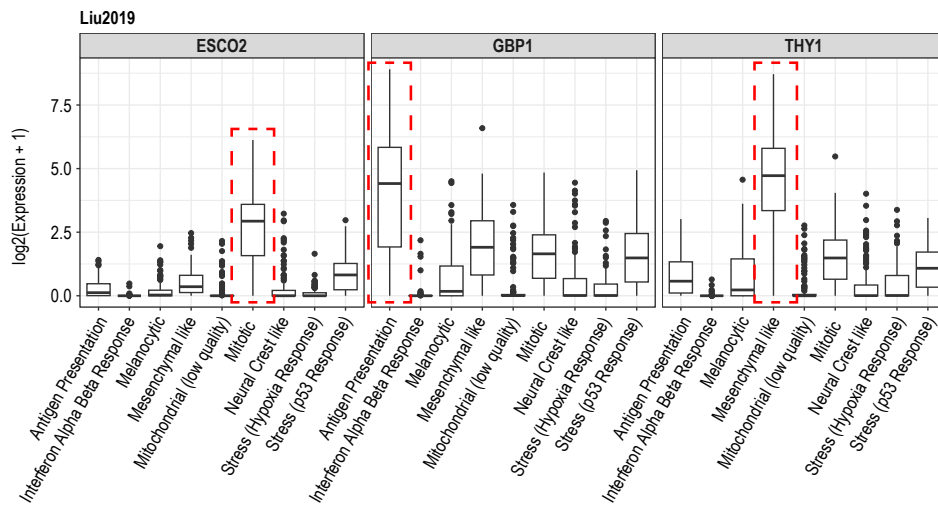


Figure S5: The cell state main effect signatures validation on external bulk data (Liu2019). The box plot shows the expression of genes *ESCO2*, *GBP1* and *THY1*, mapped onto cell state specific deconvolutions of external bulk RNA dataset. Consistent with the observations in Figure 4.2c, these genes exhibit high expression in particular cell states, underscoring their role as specific markers across different datasets.

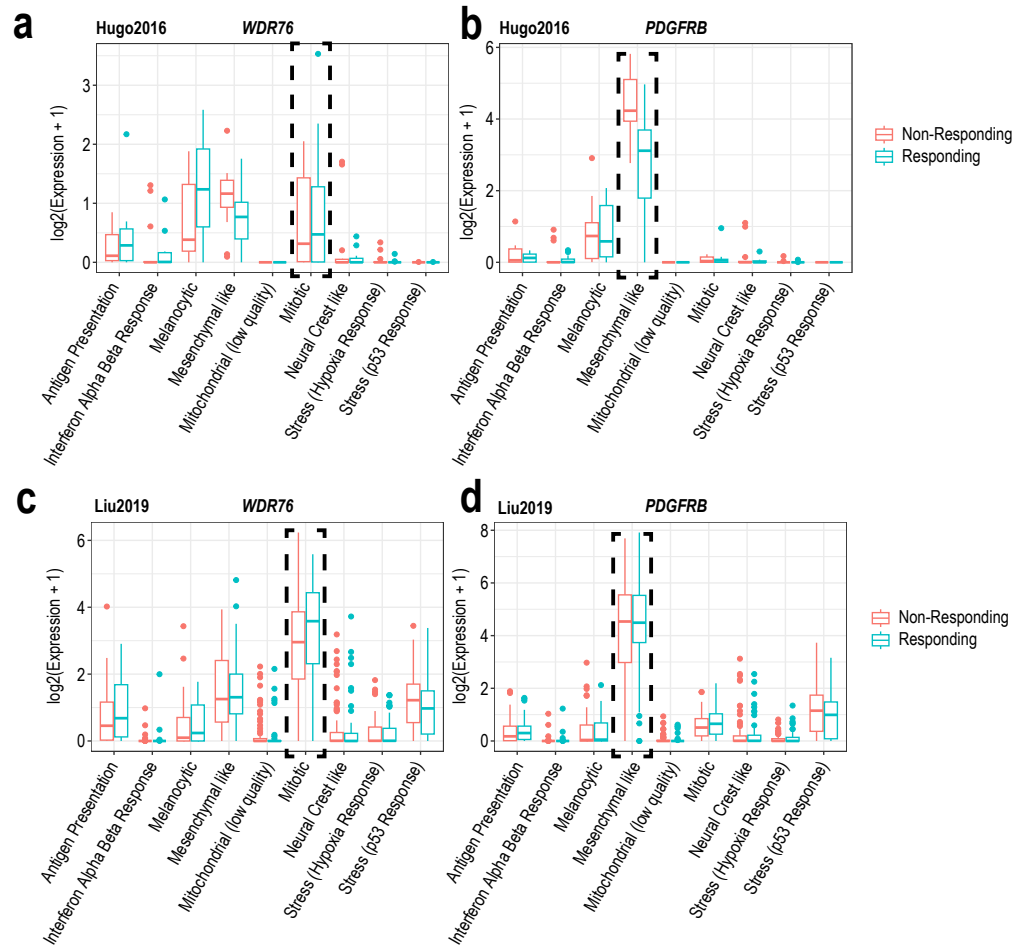


Figure S6: The interaction effect signatures, such as cell state specific responding signatures, are transferable to external bulk data. **(a,c)** Box plot showing *WDR76* gene expression levels in deconvoluted cell states from Hugo2016 and Liu2019 bulk samples. **(b,d)** Box plot showing *PDGFRB* gene expression levels in deconvoluted cell states from Hugo2016 and Liu2019 bulk samples. The colors indicate the phenotype of the bulk samples.

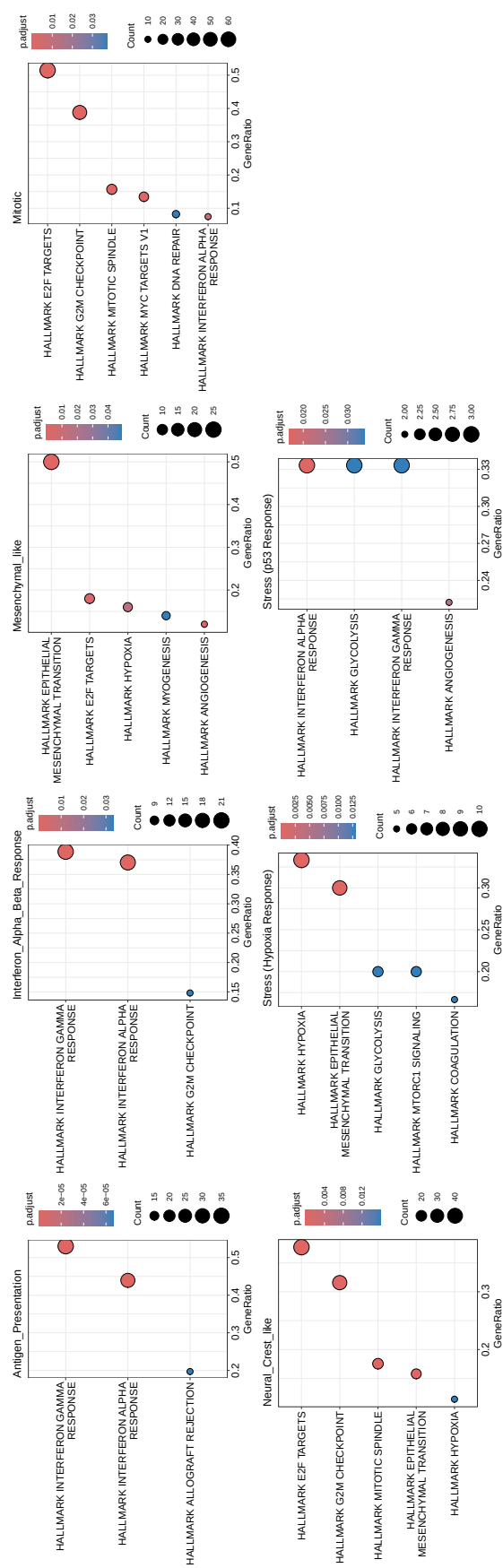


Figure S7: Hallmark gene set enrichment analysis of cell state-specific signatures (positive markers) obtained from the three-way iDAS model. Hallmark gene enrichment on seven cell states. The gene ratio for each hallmark is plotted along the x-axis, and the size of each dot represents the count of genes in the corresponding gene set. Colors indicate the adjusted p-values, with a gradient from red (most significant) to blue (least significant).

C.2 SUPPLEMENTARY TABLES

Table S1: Cell state signatures enriched in hallmark pathways (two-factor analysis)

Category	Gene	Gene	Gene	Gene
Immunotherapy_response	ABL2	CCND3	MIPEP	
Immunotherapy_response	TOR1AIP2	KLHDC3	B2M	
Immunotherapy_response	COX5B	SCRN1	NDUFAB1	
Immunotherapy_response	SPCS1	DECR1	SIRT7	
Immunotherapy_response	C5orf51	TIMM10	TUBB4A	
Immunotherapy_response	HINT1	PCSK7	ZNF135	
Immunotherapy_response	RNF130	ARCN1	SOX1	
Immunotherapy_response	TMEM14C	APLP2		
Immunotherapy_response	LSM2	KRT18		
Immunotherapy_response	RNF5	ISCU		
Cell_state	ISG15	SPC25	TRIP13	PKMYT1
Cell_state	E2F2	BARD1	CENPK	KIF22
Cell_state	CLSPN	HJURP	CENPH	ORC6
Cell_state	GBP3	FANCD2	DHFR	MIS12
Cell_state	GBP1	CDC25A	LMNB1	AURKB
Cell_state	GBP4	TRAIP	CDC25C	CDC6
Cell_state	CKS1B	POC1A	CD74	TOP2A
Cell_state	MPZ	POLQ	DEK	PSMC3IP
Cell_state	NUF2	PARP14	GMNN	BRCA1
Cell_state	UBE2T	MCM2	BTN3A1	NGFR
Cell_state	DTL	RFC4	NRM	SYNGR2
Cell_state	EXO1	NCAPG	HLA-DRB1	TK1
Cell_state	CMPK2	CXCL10	TAP2	TYMS
Cell_state	RRM2	ZGRF1	HLA-DPA1	NDC80
Cell_state	CHAC2	MAD2L1	TAPBP	RBBP8
Cell_state	NCAPH	CCNA2	KIFC1	SKA1
Cell_state	BUB1	MND1	MTFR2	MCM8

Continued on next page

Category	Gene	Gene	Gene	Gene
Cell_state	CKAP2L	DDX60	FBXO5	TPX2
Cell_state	MCM6	NEIL3	NUDT1	E2F1
Cell_state	NMI	CENPU	RPA3	DSN1
Cell_state	GGCT	RARRES3	OAS3	RBL1
Cell_state	RFC2	CDCA5	KNTC1	FAM83D
Cell_state	MCM7	THY1	POLE	UBE2C
Cell_state	XRCC2	DHTKD1	BRCA2	UHRF1
Cell_state	NCAPG2	MASTL	RFC3	DNMT1
Cell_state	CDCA2	ZWINT	EPSTI1	SPC24
Cell_state	ESCO2	CDK1	CKAP2	ASF1B
Cell_state	PBK	IFIT2	DIAPH3	WDR62
Cell_state	MYBL1	IFIT3	PSME2	LIG1
Cell_state	RAD54B	FOXO1	LRR1	POLD1
Cell_state	DSCC1	RAD51AP1	CDKN3	CDC45
Cell_state	ATAD2	NCAPD2	UBR7	MCM5
Cell_state	MELK	CDCA3	VRK1	APOL1
Cell_state	CENPP	RERG	WARS	CENPM
Cell_state	ZNF367	TMEM106C	BUB1B	NCAPH2
Cell_state	WDR34	TUBA1B	NUSAP1	DONSON
Cell_state	RRM1	RACGAP1	HAPLN3	C21orf58
Cell_state	E2F8	TIMELESS	FANCI	HLA-DRB5
Cell_state	FAM111B	RBMS2	BLM	
Cell_state	FEN1	PARPBP	SNRNP25	
Interaction_effect	CDA	TCF19	WDR76	PRSS3
Interaction_effect	ALPL	HSD17B8	MNS1	XKR5
Interaction_effect	FAM110D	FILIP1	FAM192A	GDAP1L1
Interaction_effect	FCN3	TAGAP	ABI3	
Interaction_effect	GJA4	HGF	CNN1	
Interaction_effect	USP1	GIMAP4	C19orf57	
Interaction_effect	INTS7	GFRA2	PLVAP	

Continued on next page

Category	Gene	Gene	Gene	Gene
Interaction_effect	SUSD4	NECAB1	ZNF681	
Interaction_effect	HTRA2	GALNT18	JAM2	
Interaction_effect	PLA2R1	ARNTL	CYP4X1	
Interaction_effect	KIF15	MCM10	CD3E	
Interaction_effect	CLEC3B	TSPAN11	TGFA	
Interaction_effect	SLCO2A1	CABP1	HIST1H2AL	
Interaction_effect	PL4K2B	SKA3	CTNNA3	
Interaction_effect	EDNRA	HMGB1	PNMT	
Interaction_effect	C4orf46	SLC25A30	HIST2H2AB	
Interaction_effect	ARSB	TFDP1	FBN3	
Interaction_effect	PDGFRB	CLEC14A	PTPRCAP	
Interaction_effect	FLT4	POLE2	C4orf36	
Interaction_effect	HIST1H3B	OIP5	KCNK9	
Additive	RHOB	CENPQ	PSME1	SUGP2
Additive	SP110	MCM3	ERH	SGSM3
Additive	SP100	COL19A1	DUT	RCAN1
Additive	LAP3	PHIP	POLR3K	
Additive	COX7C	RHEB	RMI2	
Additive	HIST1H1E	CKS2	NUDT21	
Additive	TUBB	ZDHHC12	NUFIP2	
Additive	PSMB8	CCDC34	COL1A1	
Additive	PSMB9	RBM17	RNF213	
Additive	TAP1	RNASEH2B	PCNA	

Table S2: Cell state signatures enriched in hallmark pathways (two-factor analysis)

Cell State	Term	Adjusted P-value	Count
Neural Crest-like	E2F_TARGETS	1.400E-26	31
Neural Crest-like	G2M_CHECKPOINT	4.410E-16	23
Neural Crest-like	MITOTIC_SPINDLE	2.635E-04	11
Neural Crest-like	MYC_TARGETS_V1	2.195E-02	8
Interferon Alpha Beta Response	G2M_CHECKPOINT	2.930E-08	12
Interferon Alpha Beta Response	E2F_TARGETS	2.170E-07	11
Interferon Alpha Beta Response	INTERFERON_GAMMA_RESPONSE	1.504E-04	8
Interferon Alpha Beta Response	INTERFERON_ALPHA_RESPONSE	1.504E-04	6
Interferon Alpha Beta Response	DNA_REPAIR	1.019E-02	5
MES	E2F_TARGETS	4.312E-20	21
MES	G2M_CHECKPOINT	8.793E-06	10
MES	MYC_TARGETS_V1	3.297E-03	7
MES	DNA_REPAIR	3.308E-03	6
Antigen Presentation	INTERFERON_GAMMA_RESPONSE	8.919E-15	15
Antigen Presentation	INTERFERON_ALPHA_RESPONSE	2.790E-14	12
Mitotic	E2F_TARGETS	3.030E-40	44
Mitotic	G2M_CHECKPOINT	2.270E-23	32
Mitotic	MITOTIC_SPINDLE	8.396E-04	12
Mitotic	DNA_REPAIR	5.224E-03	9
Mitotic	MYC_TARGETS_V1	8.703E-03	10
Stress (hypoxia response)	INTERFERON_GAMMA_RESPONSE	4.700E-04	4
Stress (hypoxia response)	INTERFERON_ALPHA_RESPONSE	8.012E-04	3

Table S3: Cell state signatures enriched in hallmark pathways (three-factor analysis)

Category	Gene	Gene	Gene	Gene
Treatment_status	APOC1			
Cell_state	ISG15	DEPDC1	S100A4	EXO1
Cell_state	ICMT	IFI44L	S100A1	RRM2
Cell_state	ENO1	GBP1	CKS1B	PPM1G
Cell_state	STMN1	GBP2	NUF2	NCAPH
Cell_state	SH3BGR13	GBP4	ASPM	SLC20A1
Cell_state	HMGN2	GSTM3	UBE2T	CKAP2L
Cell_state	CLSPN	CTSS	NUCKS1	MCM6
Cell_state	SLC2A1	S100A10	DTL	IFIH1
Cell_state	NASP	S100A11	CENPF	SPC25
Cell_state	USP1	S100A6	LYST	COL3A1
Cell_state	COL5A2	NKTR	CHCHD6	IGFBP7
Cell_state	STAT1	KIF15	MCM2	CXCL10
Cell_state	BARD1	POC1A	GMPS	H2AFZ
Cell_state	FN1	FLNB	SMC4	MAD2L1
Cell_state	IGFBP5	DCBLD2	RFC4	CCNA2
Cell_state	SERPINE2	CCDC80	RNF168	MND1
Cell_state	HJURP	FAM162A	SLBP	HMGB2
Cell_state	COL6A3	PARP9	TACC3	NEIL3
Cell_state	DTYMK	DTX3L	LAP3	CENPU
Cell_state	FANCD2	PARP14	NCAPG	MYO10
Cell_state	ITGA1	CD74	NRM	PSMB8
Cell_state	CENPK	SPARC	TUBB	PSMB9
Cell_state	CENPH	HNRNPAB	IER3	TAP1
Cell_state	DHFR	DEK	TCF19	HLA-DMA
Cell_state	VCAN	GMNN	HLA-C	HLA-DPA1
Cell_state	LMNB1	HIST1H4C	HLA-DRA	HLA-DPB1
Cell_state	HINT1	HIST1H1E	HLA-DRB1	TAPBP
Cell_state	P4HA2	HLA-F	HLA-DQA1	KIFC1

Continued on next page

Category	Gene	Gene	Gene	Gene
Cell_state	IRF1	HLA-A	HLA-DQB1	VEGFA
Cell_state	PDGFRB	HLA-E	TAP2	MCM3
Cell_state	EEF1A1	RFC2	PBK	RAD21
Cell_state	CENPW	SAMD9L	GIN54	DSCC1
Cell_state	FBXO5	MCM7	PRKDC	ATAD2
Cell_state	FAM20C	PCOLCE	MCM4	NDRG1
Cell_state	NUDT1	HILPDA	SDCBP	LY6E
Cell_state	RPA3	EZH2	GGH	TONSL
Cell_state	NDUFA4	RARRES2	MYBL1	RECQL4
Cell_state	GGCT	XRCC2	CCNE2	MLANA
Cell_state	ANLN	NCAPG2	UQCRB	PSIP1
Cell_state	H2AFV	ESCO2	EIF3E	MELK
Cell_state	ANXA1	PHLDA2	POLD3	P4HA1
Cell_state	CKS2	RRM1	CADM1	DNAJC9
Cell_state	SMC2	ADM	THY1	IFIT2
Cell_state	TNC	CCDC34	CHEK1	IFIT3
Cell_state	PHF19	UBE2L6	NCAPD3	KIF20B
Cell_state	WDR34	SERPING1	MCM10	KIF11
Cell_state	COL5A1	FAM111B	VIM	HELLS
Cell_state	TUBB4B	FAM111A	MASTL	SMC3
Cell_state	IFITM1	FEN1	ZWINT	MKI67
Cell_state	IFITM3	CDCA5	CDK1	FOXM1
Cell_state	RAD51AP1	PFDN5	HMGB1	PSME1
Cell_state	GAPDH	PMEL	BRCA2	PSME2
Cell_state	CDCA3	TIMELESS	N4BP2L2	CDKN3
Cell_state	C1R	NACA	RFC3	WARS
Cell_state	MGP	E2F7	TPT1	SIVA1
Cell_state	PPFIBP1	LUM	CKAP2	CDCA4
Cell_state	TMEM106C	DCN	DIAPH3	NUSAP1
Cell_state	TUBA1B	TMPO	TNFSF13B	WDR76

Continued on next page

Category	Gene	Gene	Gene	Gene
Cell_state	TUBA1A	RAN	COL4A1	B2M
Cell_state	RACGAP1	SKA3	COL4A2	DUT
Cell_state	MNS1	NUPR1	CDT1	IFI35
Cell_state	KIF23	KIF22	PFN1	BRCA1
Cell_state	PKM	HIRIP3	AURKB	ITGA3
Cell_state	FANCI	FUS	WSB1	COL1A1
Cell_state	BLM	SHCBP1	TMEM97	VMP1
Cell_state	PRC1	ORC6	ATAD5	BRIP1
Cell_state	PKMYT1	MT2A	CDC6	TK1
Cell_state	PAQR4	CKLF	TOP2A	BIRC5
Cell_state	TNFRSF12A	CENPN	IGFBP4	LGALS3BP
Cell_state	CARHSP1	GIN52	TUBG1	TYMS
Cell_state	NDC80	RBL1	ACTN4	CENPM
Cell_state	RBBP8	MYBL2	PLAUR	GTSE1
Cell_state	TCF4	UBE2C	LIG1	BTG3
Cell_state	CDH19	CHAF1A	CDC45	DONSON
Cell_state	PCNA	UHRF1	RANBP1	SLC5A3
Cell_state	MCM8	DNMT1	TIMP3	MX1
Cell_state	CST3	SPC24	MCM5	COL18A1
Cell_state	TPX2	ASF1B	APOL6	COL6A2
Cell_state	E2F1	BST2	APOL1	C21orf58
Cell_state	DSN1	LSM4	LGALS1	S100B
Cell_state	HLA-DRB5			
additive	IGFBP2			
Additive	MAP1B			
Additive	TOMM7			
Additive	COL1A2			
Additive	AHNAK			
Additive	RARRES3			
Additive	PMEPA1			

Continued on next page

Category	Gene	Gene	Gene	Gene
Additive	APOE			
Additive	COL6A1			
Twoways_interaction(F2F3)	NDUFA4L2			
Threeways_interaction	IFI6	DCT		
Threeways_interaction	CHI3L2	CASS4		
Threeways_interaction	RGS5	ANGPTL4		
Threeways_interaction	VIT	DSCR4		
Threeways_interaction	HIST1H3B	KCNK3		
Threeways_interaction	HLA-B	CD5		
Threeways_interaction	TAGAP	CHGB		
Threeways_interaction	IGFBP3	CAMP		
Threeways_interaction	IQUB			
Threeways_interaction	A2M			

Table S4: Cell state signatures enriched in hallmark pathways (three-factor analysis)

Cell State	Term	Adjusted P-value	Count
Neural Crest-like	E2F_TARGETS	3.398E-28	43
Neural Crest-like	G2M_CHECKPOINT	1.498E-20	36
Neural Crest-like	MITOTIC_SPINDLE	1.718E-06	20
Neural Crest-like	EPITHELIAL_MESENCHYMAL_TRANSITION	3.064E-05	18
Neural Crest-like	HYPOXIA	1.593E-02	13
Interferon Alpha Beta Response	INTERFERON_ALPHA_RESPONSE	6.560E-19	20
Interferon Alpha Beta Response	INTERFERON_GAMMA_RESPONSE	6.170E-14	21
Interferon Alpha Beta Response	G2M_CHECKPOINT	3.337E-02	8
MES	EPITHELIAL_MESENCHYMAL_TRANSITION	1.120E-19	25
MES	ANGIOGENESIS	4.260E-05	6
MES	E2F_TARGETS	4.103E-03	9
MES	HYPOXIA	1.468E-02	8
MES	MYOGENESIS	4.880E-02	7
Antigen Presentation	INTERFERON_ALPHA_RESPONSE	8.737E-31	29
Antigen Presentation	INTERFERON_GAMMA_RESPONSE	1.889E-29	35
Antigen Presentation	ALLOGRAFT_REJECTION	6.674E-05	13
Mitotic	E2F_TARGETS	6.269E-59	69
Mitotic	G2M_CHECKPOINT	1.071E-35	52
Mitotic	MITOTIC_SPINDLE	5.467E-06	21
Mitotic	MYC_TARGETS_V1	2.825E-04	18
Mitotic	INTERFERON_ALPHA_RESPONSE	5.189E-03	10
Mitotic	DNA_REPAIR	3.727E-02	11
Stress (p53 response)	INTERFERON_ALPHA_RESPONSE	1.683E-02	3
Stress (p53 response)	ANGIOGENESIS	2.392E-02	2
Stress (p53 response)	GLYCOLYSIS	3.365E-02	3
Stress (p53 response)	INTERFERON_GAMMA_RESPONSE	3.365E-02	3
Stress (hypoxia response)	HYPOXIA	4.226E-07	10
Stress (hypoxia response)	EPITHELIAL_MESENCHYMAL_TRANSITION	4.450E-06	9
Stress (hypoxia response)	GLYCOLYSIS	1.983E-03	6
Stress (hypoxia response)	MTORC1_SIGNALING	1.983E-03	6
Stress (hypoxia response)	COAGULATION	2.173E-03	5

BIBLIOGRAPHY

- Abdelaal, T., Michielsen, L., Cats, D., et al. (2019). A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome Biology*, 20:1–19.
- Adossa, N., Khan, S., Rytönen, K., et al. (2021). Computational strategies for single-cell multi-omics integration. *Computational and Structural Biotechnology Journal*, 19:2588–2596.
- Aicher, T. P., Carroll, S., Raddi, G., et al. (2019). *Seq-Well: A Sample-Efficient, Portable Picowell Platform for Massively Parallel Single-Cell RNA Sequencing*, page 111–132. Springer New York.
- Aissa, A. F., Islam, A. B. M. M. K., Ariss, M. M., et al. (2021). Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature Communications*, 12(1).
- Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., and Powell, J. E. (2019). scpred: accurate supervised method for cell-type classification from single-cell rna-seq data. *Genome Biology*, 20(1).
- Andrews, T. S. and Hemberg, M. (2018). Identifying cell populations with scRNAseq. *Molecular Aspects of Medicine*, 59:114–122.
- Aran, D., Looney, A. P., Liu, L., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172.
- Argelaguet, R., Arnol, D., Bredikhin, D., et al. (2020). Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):1–17.
- Argelaguet, R., Cuomo, A. S. E., Stegle, O., and Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39(10):1202–1215.
- Ashuach, T., Gabitto, M. I., Koodli, R. V., et al. (2023). Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231.
- Avila Cobos, F., Vandesompele, J., Mestdag, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*, 34(11):1969–1979.
- Bandura, D. R., Baranov, V. I., Ornatsky, O. I., et al. (2009). Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry*, 81(16):6813–6822.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

- Brennecke, P., Anders, S., Kim, J. K., et al. (2013). Accounting for technical noise in single-cell rna-seq experiments. *Nature Methods*, 10(11):1093–1095.
- Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., and Bolker, B. M. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, 9(2):378–400.
- Buenrostro, J. D., Wu, B., Litzenburger, U. M., et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.
- Buettner, F., Natarajan, K. N., Casale, F. P., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160.
- Burdziak, C., Alonso-Curbelo, D., Walle, T., Reyes, J., et al. (2023). Epigenetic plasticity cooperates with cell-cell interactions to direct pancreatic tumorigenesis. *Science*, 380(6645).
- Burkhardt, D. B., Stanley, J. S., Tong, A., Perdigoto, A. L., Gigante, S. A., Herold, K. C., Wolf, G., Giraldez, A. J., van Dijk, D., and Krishnaswamy, S. (2021). Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39(5):619–629.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420.
- Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27.
- Cao, J., Cusanovich, D. A., Ramani, V., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385.
- Cao, J., Spielmann, M., Qiu, X., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502.
- Cao, Y., Geddes, T., Yang, J. Y. H., et al. (2020). Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, 2(9):500–508.
- Cao, Y., Lin, Y., Patrick, E., et al. (2022). sfeatures: multi-view representations of single-cell and spatial data for disease outcome prediction. *Bioinformatics*, 38(20):4745–4753.
- Cao, Z.-J. and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466.
- Chalise, P. and Fridley, B. L. (2017). Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm. *PLOS ONE*, 12(5):e0176278.
- Chen, A., Liao, S., Cheng, M., et al. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21.

- Chen, K. H., Boettiger, A. N., Moffitt, J. R., et al. (2015). Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233).
- Chen, S., Lake, B. B., and Zhang, K. (2019a). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457.
- Chen, T.-Y., You, L., Hardillo, J. A. U., and Chien, M.-P. (2023). Spatial transcriptomic technologies. *Cells*, 12(16):2042.
- Chen, Y.-C., Suresh, A., Underbayev, C., et al. (2019b). Ikap—identifying k major cell population groups in single-cell rna-sequencing analysis. *GigaScience*, 8(10).
- Cheng, A., Hu, G., and Li, W. V. (2022). Benchmarking cell-type clustering methods for spatially resolved transcriptomics data. *Briefings in Bioinformatics*, 24(1).
- Cheng, C., Easton, J., Rosencrance, C., et al. (2019). Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell rna-seq data. *Nucleic acids research*, 47(22):e143–e143.
- Chu, T., Wang, Z., Pe’er, D., and Danko, C. G. (2022). Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology. *Nature Cancer*, 3(4):505–517.
- Clark, S. J., Argelaguet, R., Kapourani, C.-A., et al. (2018). scnm-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature Communications*, 9(1).
- Cusanovich, D. A., Daza, R. M., Adey, A., et al. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (New York, N.Y.)*, 348(6237):910–914.
- Dann, E., Henderson, N. C., Teichmann, S. A., et al. (2021). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253.
- de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T., and Holstege, F. C. P. (2019). Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic Acids Research*, 47(16):e95–e95.
- Ding, J., Shah, S. P., and Condon, A. (2016). densitycut: an efficient and versatile topological approach for automatic clustering of biological data. *Bioinformatics (Oxford, England)*, 32(17):2567–2576.
- Do, V. H. and Canzar, S. (2021). A generalization of t-sne and umap to single-cell multimodal omics. *Genome Biology*, 22(1):1–9.
- Domínguez Conde, C., Xu, C., Jarvis, L. B., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594).
- Dou, J., Liang, S., Mohanty, V., et al. (2022). Bi-order multimodal integration of single-cell data. *Genome Biology*, 23(1).

- Duò, A., Robinson, M. D., and Soneson, C. (2020). A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7:1141.
- Eng, C.-H. L., Lawson, M., Zhu, Q., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. *Nature*, 568(7751):235–239.
- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.
- Finak, G., McDavid, A., Yajima, M., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, 16(1).
- Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data. *F1000Research*, 7:1297.
- Fu, X., Cao, Y., Bian, B., et al. (2024). Spatial gene expression at single-cell resolution from histology using deep learning with ghist.
- Gayoso, A., Steier, Z., Lopez, R., et al. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18(3):272–282.
- Geddes, T. A., Kim, T., Nan, L., et al. (2019). Autoencoder-based cluster ensembles for single-cell rna-seq data analysis. *BMC Bioinformatics*, 20:1–11.
- Gong, W., Kwak, I.-Y., Pota, P., et al. (2018). Drimpute: imputing dropout events in single cell rna sequencing data. *BMC Bioinformatics*, 19(1).
- Grün, D., Lyubimova, A., Kester, L., et al. (2015). Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255.
- Guo, M., Wang, H., Potter, S. S., Whitsett, J. A., and Xu, Y. (2015). Sincera: A pipeline for single-cell rna-seq profiling analysis. *PLOS Computational Biology*, 11(11):e1004575.
- Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1).
- Hao, Y., Hao, S., Andersen-Nissen, E., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.
- Hashimshony, T., Senderovich, N., Avital, G., et al. (2016). Cel-seq2: sensitive highly-multiplexed single-cell rna-seq. *Genome Biology*, 17(1).
- Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature Biotechnology*, 37(6):685–691.
- Hornik, K. (2005). A clue for cluster ensembles. *Journal of Statistical Software*, 14:1–25.

- Hou, Y., Guo, H., Cao, C., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research*, 26(3):304–319.
- Hu, P., Zhang, W., Xin, H., and Deng, G. (2016). Single cell isolation and analysis. *Frontiers in Cell and Developmental Biology*, 4.
- Huang, G., Li, Y., Pleiss, G., et al. (2017). Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.
- Huang, M., Wang, J., Torre, E., et al. (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, 15(7):539–542.
- Hugo, W., Zaretsky, J. M., Sun, L., et al. (2016). Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165(1):35–44.
- Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.
- Ji, Z. and Ji, H. (2016). Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic Acids Research*, 44(13):e117–e117.
- Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome Biology*, 17(1).
- John, C. R., Watson, D., Barnes, M. R., et al. (2019). Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics (Oxford, England)*, 36(4):1159–1166.
- Johnson, T. S., Yu, C. Y., Huang, Z., et al. (2022). Diagnostic evidence gauge of single cells (degas): a flexible deep transfer learning framework for prioritizing cells in relation to disease. *Genome Medicine*, 14(1).
- Karim, M. R., Beyan, O., Zappa, A., et al. (2020). Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22(1):393–415.
- Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742.
- Kim, H. J., Lin, Y., Geddes, T., et al. (2020). Citefuse enables multi-modal analysis of cite-seq data. *Bioinformatics*, 36(14):4137–4143.
- Kim, H. J., Wang, K., Chen, C., et al. (2021). Uncovering cell identity through differential stability with cepo. *Nature Computational Science*, 1(12):784–790.
- Kim, T., Chen, I. R., Lin, Y., et al. (2019). Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in Bioinformatics*, 20(6):2316–2326.

- Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282.
- Kiselev, V. Y., Kirschner, K., Schaub, M. T., et al. (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486.
- Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell rna-seq data across data sets. *Nature Methods*, 15(5):359–362.
- Korsunsky, I., Millard, N., Fan, J., et al. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, pages 1–8.
- Korthauer, K. D., Chu, L.-F., Newton, M. A., et al. (2016). A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome Biology*, 17(1).
- Kothari, R. and Pitts, D. (1999). On finding the number of clusters. *Pattern Recognition Letters*, 20(4):405–416.
- Lähnemann, D., Köster, J., Szczurek, E., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1).
- Lange, T., Roth, V., Braun, M. L., and Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323.
- Larsson, L., Frisén, J., and Lundeberg, J. (2021). Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods*, 18(1):15–18.
- Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15(2):R29.
- Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):1–9.
- Lieberman, Y., Rokach, L., and Shay, T. (2018). Castle – classification of single cells by transfer learning: Harnessing the power of publicly available single cell rna sequencing experiments to annotate new experiments. *PLOS ONE*, 13(10):e0205499.
- Lin, P., Troup, M., and Ho, J. W. K. (2017). Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome biology*, 18(1):59–59.
- Lin, X., Tian, T., Wei, Z., et al. (2022). Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature Communications*, 13(1):7705.
- Lin, Y., Cao, Y., Kim, H. J., et al. (2020). scclassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular Systems Biology*, 16(6).

- Lin, Y., Ghazanfar, S., Wang, K. Y., et al. (2019). scmerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell rna-seq datasets. *Proceedings of the National Academy of Sciences*, 116(20):9775–9784.
- Liu, C., Huang, H., and Yang, P. (2023). Multi-task learning from multimodal single-cell omics with matilda. *Nucleic Acids Research*, 51(8):e45–e45.
- Liu, D., Schilling, B., Liu, D., et al. (2019). Integrative molecular and clinical modeling of clinical outcomes to pd1 blockade in patients with metastatic melanoma. *Nature Medicine*, 25(12):1916–1927.
- Liu, J., Gao, C., Sodicoff, J., et al. (2020). Jointly defining cell types from multiple single-cell datasets using liger. *Nature Protocols*, 15(11):3632–3662.
- Liu, S., Thennavan, A., Garay, J. P., et al. (2021). Multik: an automated tool to determine optimal cluster numbers in single-cell rna sequencing data. *Genome Biology*, 22(1).
- Lopez, R., Regier, J., Cole, M. B., et al. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058.
- Lord, E., Willems, M., Lapointe, F.-J., and Makarenkov, V. (2017). Using the stability of objects to determine the number of clusters in datasets. *Information Sciences*, 393:29–46.
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12).
- Luecken, M. D., Büttner, M., Chaichoompu, K., et al. (2021). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50.
- Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome Biology*, 17(1).
- Ma, S., Zhang, B., LaFave, L. M., et al. (2020). Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116.
- Macosko, E. Z., Basu, A., Satija, R., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.
- Martinez-Morilla, S., Villarroel-Espindola, F., Wong, P. F., Toki, M. I., Aung, T. N., Pelekanou, V., Bourke-Martin, B., Schalper, K. A., Kluger, H. M., and Rimm, D. L. (2021). Biomarker discovery in patients with immunotherapy-treated melanoma with imaging mass cytometry. *Clinical Cancer Research*, 27(7):1987–1996.
- Mathys, H., Davila-Velderrain, J., Peng, Z., et al. (2019). Single-cell transcriptomic analysis of alzheimer’s disease. *Nature*, 570(7761):332–337.

- McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186.
- McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction.
- Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). Desingle for detecting three types of differential expression in single-cell rna-seq data. *Bioinformatics*, 34(18):3223–3224.
- Miao, Z., Humphreys, B., McMahon, A., et al. (2021). Multi-omics integration in the age of million single-cell data. *Nature Reviews Nephrology*, 17(11):710–724.
- Min, E., Guo, X., Liu, Q., et al. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6:39501–39514.
- Missarova, A., Rosen, L. U., Dann, E., Satija, R., and Marioni, J. (2023). Sensitive cluster-free differential expression testing.
- Mohammadi, S., Davila-Velderrain, J., and Kellis, M. (2020). A multiresolution framework to characterize single-cell state landscapes. *Nature communications*, 11(1):5399–5399.
- Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1/2):91–118.
- Nagano, T., Lubling, Y., Stevens, T. J., et al. (2013). Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64.
- Nature (2020). Method of the Year 2019: Single-cell multimodal omics. *Nature, Methods*, 17(1):1–1.
- Nofech-Mozes, I., Soave, D., Awadalla, P., and Abelson, S. (2023). Pan-cancer classification of single cells in the tumour microenvironment. *Nature Communications*, 14(1).
- Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12):2074–2093.
- Peterson, V. M., Zhang, K. X., Kumar, N., et al. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10):936–939.
- Peyvandipour, A., Shafi, A., Saberian, N., and Draghici, S. (2020). Identification of cell types from single cell data using stable clustering. *Scientific Reports*, 10(1).
- Picelli, S., Björklund, Å. K., Faridani, O. R., et al. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1098.
- Pozniak, J., Pedri, D., Landeloos, E., et al. (2024). A tcf4-dependent gene regulatory network confers resistance to immunotherapy in melanoma. *Cell*, 187(1):166–183.e25.

- Ramaswamy, A., Brodsky, N. N., Sumida, T. S., et al. (2021). Immune dysregulation and autoreactivity correlate with disease severity in sars-cov-2-associated multisystem inflammatory syndrome in children. *Immunity*, 54(5):1083–1095.
- Regev, A., Teichmann, S. A., Lander, E. S., et al. (2017). The human cell atlas. *eLife*, 6.
- Reshef, Y. A., Rumker, L., Kang, J. B., et al. (2021). Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nature Biotechnology*, 40(3):355–363.
- Ritchie, M. E., Phipson, B., Wu, D., et al. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47.
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.
- Rodriques, S. G., Stickels, R. R., Goeva, A., et al. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467.
- Rotem, A., Ram, O., Shoresh, N., et al. (2015). Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.
- Skinneider, M. A., Squair, J. W., Kathe, C., et al. (2020). Cell type prioritization in single-cell data. *Nature Biotechnology*, 39(1):30–34.
- Smallwood, S. A., Lee, H. J., Angermueller, C., and other (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820.
- Soumillon, M., Cacchiarelli, D., Semrau, S., et al. (2014). Characterization of directed differentiation by high-throughput single-cell rna-seq.
- Stephenson, E., Reynolds, G., Botting, R., et al. (2021). Single-cell multi-omics analysis of the immune response in covid-19. *Nature Medicine*, 27(5):904–916.
- Stoeckius, M., Hafemeister, C., Stephenson, W., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868.
- Street, K., Risso, D., Fletcher, R. B., et al. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1).
- Stuart, T., Butler, A., Hoffman, P., et al. (2019). Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21.
- Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a dataset. *Journal of the American Statistical Association*, 98(463):750–763.

- Sun, D., Guan, X., Moran, A. E., et al. (2021). Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. *Nature Biotechnology*, 40(4):527–538.
- Swanson, E., Lord, C., Reading, J., et al. (2021). Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using tea-seq. *eLife*, 10:e63632.
- The Tabula Muris Consortium (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372. Publisher: Nature Publishing Group.
- The Tabula Sapiens Consortium (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896. Publisher: American Association for the Advancement of Science.
- Tian, T., Wan, J., Song, Q., and Wei, Z. (2019). Clustering single-cell rna-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4):191–198.
- Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(2):411–423.
- Traag, V. A., Waltman, L., and van Eck, N. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233–5233.
- Tracy, C. A. and Widom, H. (1994). Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159(1):151–174.
- Tran, H. T. N., Ang, K. S., Chevrier, M., et al. (2020). A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biology*, 21(1).
- Tripathi, R., Liu, Z., Jain, A., et al. (2020). Combating acquired resistance to mapk inhibitors in melanoma by targeting abl1/2-mediated reactivation of mek/erk/myc signaling. *Nature Communications*, 11(1).
- Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). Basics: Bayesian analysis of single-cell sequencing data. *PLOS Computational Biology*, 11(6):e1004333.
- Van den Berge, K., Roux de Bézieux, H., Street, K., et al. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, 11(1).
- van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.
- Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854.
- von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

- Wagner, S. and Wagner, D. (2007). Comparing Clusterings - An Overview. *Analysis*, 4769(001907):1–19. ISBN: 9783540748380.
- Wan, S., Kim, J., and Won, K.-J. (2020). Sharp: hyperfast and accurate processing of single-cell rna-seq data via ensemble random projection. *Genome research*, 30(2):205–213.
- Wang, B., Zhu, J., Pierson, E., et al. (2017). Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods*, 14(4):414–416.
- Wang, K. Y. X., Pupo, G. M., Tembe, V., et al. (2022). Cross-platform omics prediction procedure: a statistical machine learning framework for wider implementation of precision medicine. *npj Digital Medicine*, 5(1).
- Welch, J. D., Hartemink, A. J., and Prins, J. F. (2016). Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data. *Genome Biology*, 17(1).
- Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1).
- Wolock, S. L., Lopez, R., and Klein, A. M. (2019). Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8(4):281–291.e9.
- Wu, W., Zhang, W., and Ma, X. (2022). Network-based integrative analysis of single-cell transcriptomic and epigenomic data for cell types. *Briefings in Bioinformatics*, 23(2).
- Xie, K., Huang, Y., Zeng, F., et al. (2020). scaide: clustering of large-scale single-cell rna-seq data reveals putative and rare cell types. *NAR Genomics and Bioinformatics*, 2(4).
- Xiong, L., Xu, K., Tian, K., et al. (2019). Scale method for single-cell atac-seq analysis via latent feature extraction. *Nature Communications*, 10(1):4576.
- Yang, P., Huang, H., and Liu, C. (2021). Feature selection revisited in the single-cell era. *Genome Biology*, 22:1–17.
- Yang, P., Ormerod, J. T., Liu, W., et al. (2019). Adasampling for positive-unlabeled and label noise learning with bioinformatics applications. *IEEE Transactions on Cybernetics*, 49(5):1932–1943.
- Yang, Y. and Wang, H. (2018). Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2):83–107.
- Yi, H., Plotkin, A., and Stanley, N. (2024). Benchmarking differential abundance methods for finding condition-specific prototypical cells in multi-sample single-cell datasets. *Genome Biology*, 25(1).
- Yu, L., Cao, Y., Yang, J. Y., and Yang, P. (2022). Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biology*, 23(1):1–21. Publisher: BioMed Central Ltd.

- Yu, L., Liu, C., Yang, J. Y. H., and Yang, P. (2023). Ensemble deep learning of embeddings for clustering multimodal single-cell omics data. *Bioinformatics*, 39(6):btad382.
- Zappia, L., Phipson, B., and Oshlack, A. (2018). Exploring the single-cell rna-seq analysis landscape with the scrna-tools database. *PLOS Computational Biology*, 14(6):e1006245.
- Zappia, L. and Theis, F. J. (2021). Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. *Genome Biology*, 22(1).
- Zhang, J., Zhang, Y., Liu, J., et al. (2023a). Identification of tubb4a as a prognostic biomarker of melanoma by transcriptomic data and in vitro experiments. *Technology in Cancer Research & Treatment*, 22.
- Zhang, S., Li, X., Lin, J., et al. (2023b). Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *RNA (New York, N.Y.)*, 29(5):517–530.
- Zhao, J., Jaffe, A., Li, H., Lindenbaum, O., Sefik, E., Jackson, R., Cheng, X., Flavell, R. A., and Kluger, Y. (2021). Detection of differentially abundant cell subpopulations in scrna-seq data. *Proceedings of the National Academy of Sciences*, 118(22).
- Zheng, G. X. Y., Terry, J. M., Belgrader, P., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1).
- Zhou, B. and Wong, W. H. (2011). A bootstrap-based non-parametric anova method with applications to factorial microarray data. *Statistica Sinica*, 21(2):495–514.
- Zhu, C., Preissl, S., and Ren, B. (2020). Single-cell multimodal omics: the power of many. *Nature Methods*, 17(1):11–14.
- Ziegenhain, C., Vieth, B., Parekh, S., et al. (2017). Comparative analysis of single-cell rna sequencing methods. *Molecular Cell*, 65(4):631–643.e4.
- Zilionis, R., Nainys, J., Veres, A., et al. (2016). Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1):44–73.