What I talk about when I talk about integration of single-cell data

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Yang Xu
August 2022

Copyright © 2022 by Yang X. “What I talk about when I talk about integration of single-cell data” Yang
All rights reserved.

ACKNOWLEDGEMENTS

Coming from a rural region in the southwest of China, I am becoming the first person in my hometown who would earn a doctoral degree in the United States. I don't think I could go this far without support from my parents. Many people in my hometown didn't believe that knowledge would have big impact on life quality, and neither of my parent finished high school. However, both of my parents have firm faith that higher education can change my life and I could go far beyond with knowledge and have a wider world view. Their faith in education played a critical role in shaping my world view and shaping who I am today. My warmest and biggest appreciation would be for my parents.

Along my education journey, I have met many people. They showed up in different stages of this journey and kindly taught me their experiences and perspectives as my mentors. I have my special thanks to my doctoral advisor, Dr. Rachel Patton McCord. As my mentor, she gave me enough flexibility as well as guidance to help me find my research focuses and strength. She is also generous and supportive on my career development. I own her a lot in terms of professional growth. I also want to thank all my committee members and my graduate program, Genome Science & Technology, for your generous support.

Finally, a special gratitude goes to my partner, Gabriel Cole. He was always there for me and stand by me whenever I need, in the past years of my graduate school. I am happy to share every life moment with him.

ABSTRACT

Over the past decade, single-cell technologies evolved from profiling hundreds of cells to millions of cells, and emerged from a single modality of data to cover multiple views at single-cell resolution, including genome, epigenome, transcriptome, and so on. With advance of these single-cell technologies, the booming of multimodal single-cell data creates a valuable resource for us to understand cellular heterogeneity and molecular mechanism at a comprehensive level. However, the large-scale multimodal single-cell data also presents a huge computational challenge for insightful integrative analysis. Here, I will lay out problems in data integration that single-cell research community is interested in and introduce computational principles for solving these integration problems. In the following chapters, I will present four computational methods for data integration under different scenarios. Finally, I will discuss some future directions and potential applications of single-cell data integration.

PREFACE

I came up with this dissertation title when I reflected what I have explored so far and what I learned along the way. It reminds me of the book *What I Talk About When I Talk About Running* by Haruki Murakami, in which he shared an intimate journey he had with writing and running.

Just like writing and running for Haruki Murakami, my scientific journey in graduate school is also intimate and personal. I have explored a lot, and many projects I have tried failed. Fortunately, I narrowed down a path that I believe is right, and I navigated through failure to have some achievements to enjoy. Looking back, I found it's a process of building up momentum. The momentum pushes me to go higher and higher.

Just like writing and running for Haruki Murakami, I was also lucky to figure out how to balance life and research, and I am lucky to enjoy both life moments and research achievements, big and small. This could never happen, if I don't have passion on either regular life moments or my scientific mission. My journey in graduate school is finding and keeping passion for both, over struggles and difficulties.

TABLE OF CONTENTS

INTRODUCTION THE DECADE OF SINGLE-CELL TECHNOLOGY 1

The emergence of single-cell technologies 2

Batch effects! Batch effects! Batch effects! 5

From single modality to multi-modalities 8

CHAPTER I MARKER-BASED AUTOMATIC CELL-TYPE ANNOTATION FOR GENERAL SINGLE CELL EXPRESSION DATA 12

Abstract 13

Introduction 13

Method development 14

Result 20

Conclusion 30

CHAPTER II MARKER-BASED AND MODEL-FREE APPROACH FOR STANDARDIZATION AND INTEGRATION OF SINGLE-CELL TRANSCRIPTOMICS DATA 31

Abstract 32

Introduction 33

Result 34

Benchmark impacts of data processing on batch correction 34

Workflow of MASI for integrative analysis 38

Benchmarking cell-type annotation and data integration 40

Dependence on choice of reference dataset 42

Annotation of spatial transcriptomics data with MASI 46
Integrative temporal analysis using cell-type score matrix..... 46
Case 1: Using human heart atlas for integration of single-cell human heart across studies 51
Case 2: Using human kidney atlas for integration of single-cell human kidney across multiple conditions..... 53
Case 3: transferring human lung atlas for integration of single-cell COVID19 data across participants..... 55
Discussion..... 57
Methods..... 58
Data preprocessing..... 58
Marker rank aggregation..... 59
Weighing markers..... 59
Converting gene expression matrix to cell-type score matrix 60
Classification by linear and non-linear SVM..... 60
Transfer learning through MASI 60
Transfer learning using scNym and scArches 61
2D visualization using UMAP..... 61
Integrative lineage analysis..... 61
Evaluation metrics 62
CHAPTER III MUTUAL INFORMATION LEARNING FOR INTEGRATION OF SINGLE-CELL OMICS DATA 64

Abstract 65
Introduction 66
Method 69
Architecture of SMILE, p(paired)SMILE and mp(modified paired)SMILE 69
Cell pairing 73
Loss function 74
Data integration through SMILE/pSMILE/mpSMILE 75
Evaluation of data integration 76
Evaluation of label transferring 77
Processing of RNA-seq, ATAC-seq, Methyl, Hi-C, and histone marker data 77
Downstream analysis 78
Resutls 79
SMILE accommodates many single cell data types 79
SMILE eliminates batch effects in single-cell transcriptome data from multiple sources 81
Joint clustering through mpSMILE improves upon previous methods and reveals key biological variables 83
Application of p/mpSMILE in joint profiling DNA methylation and chromosome structure data 90
Combining SMILE and pSMILE for integration of more than 2 data modalities 92
Discussion 94

CHAPTER IV INTEGRATING SINGLE-CELL CHROMATIN ACCESSIBILITY AND GENE EXPRESSION DATA VIA CYCLE-CONSISTENT ADVERSARIAL NETWORK 99
Abstract 100
Introduction 101
Resutls 104
Overview of sciCAN and potential applications 104
Benchmark of sciCAN with existing integration methods 106
Integration learned by sciCAN preserves hematopoietic hierarchy 110
sciCAN identifies common responses after CRISPR perturbation 114
Conclusion 118
Methods 119
Representation learning 119
Domain adaptation 120
Cycle-consistent adversarial network 120
Data preprocessing 121
Model training 122
Integration via LIGER 123
Integration via Harmony 123
Integration via Seurat 124
Integration via ArchR 124
Integration via SMILE 125

Activity-expression velocity 125
Evaluation 125
CHAPTER V DIAGONAL INTEGRATION OF MULTIMODAL SINGLE-CELL DATA: AN ENCHANTING GOAL BUT A HAZARDOUS JOURNEY 128
Introduction 129
The enchanting goal 130
The hazardous journey 131
Searching for solid ground 134
CONCLUSION THE DIVERSITY OF MULTI-OMICS 138
1. Power single-cell transcriptome for diverse research projects 139
1.1. The diverse research projects 139
1.2. Towards data-driven integration 141
2. The diversity of multi-omics 142
2.1. Spatial transcriptome 143
2.2. 3D genome at single-cell resolution 145
3. Conclusion 149
REFERENCES 150
VITA 177

LIST OF TABLES

Table 1.1. Performance of MACA, CellAssign, SCINA, Cell-ID, and scCATACH in 6 scRNA-seq datasets, measured by ARI and NMI 22

Table 1.2. Runtime of 5 annotation tools across 6 benchmark datasets 24

Table 1.3. Mean accuracy of 5-fold SVM classifier 26

Table 1.4. Performance of MACA in 4 human-chamber single-nuclei RNA-seq datasets, measured by ARI, NMI, and Accuracy of SVM classifier 29

LIST OF FIGURES

Box 1. Key term explaination 9

Figure 1.1. Schematic workflow of MACA 17

Figure 1.2. Integrated annotation of human PBMC and pancreas data across different single cell platforms 28

Figure 2.1. Illustration of 10 analysis pipelines for scRNA-seq data 39

Figure 2.2. Benchmarking of impacts of 10 analysis pipelines on batch correction 22

Figure 2.3. Integrative annotation pipeline through MASI 39

Figure 2.4. Benchmarking of impacts of 12 DE tests on MACA-based cell-type annotation 41

Figure 2.5. Batch correction and label transferring benchmarks 43

Figure 2.6. Comparison of annotation resolution for MASI, SCCAF, and combination of SCCAF and MASI 45

Figure 2.7. Integration of scRNA-seq and Slide-seqV2 by MASI 47

Figure 2.8. Integrative lineage analysis using cell-type score matrix 49

Figure 2.9. Transferring human heart atlas for integration of single-cell human heart across research groups 52

Figure 2.10. Transferring human kidney atlas for integration of single-cell human kidney across conditions 54

Figure 2.11. Transferring human lung atlas for integration of single-cell COVID19 data across individuals 56

Figure 3.1. Architectures of 3 SMILE variants 70

Figure 3.2. Application of SMILE in single-source scRNA-seq, scATAC-seq and scHi-C 80

Figure 3.3. Integration of multi-source single-cell transcriptome data using SMILE. A, Evaluation of batch-effects correction 82

Figure 3.4. Integration of synthetic multimodal single-cell data through pSMILE 84

Figure 3.5. Integration of scRNA-seq and scATAC-seq through mpSMILE 87

Figure 3.6. Explanation of co-embedding of RNA-seq and ATAC-seq by cluster-specific differential genes and top 95-percentile genes identified through screening 89

Figure 3.7. Integration of multimodal human hematopoiesis and mouse kidney data through Seurat and mpSMILE 91

Figure 3.8. Integration of scMethyl and scHi-C through p/mpSMILE 93

Figure 3.9. Integration of Paired-Tag mouse brain data through SMILE 95

Figure 4.1. Overview of sciCAN and potential applications 105

Figure 4.2. Benchmarking of sciCAN against other 5 existing integration methods 109

Figure 4.3. Comparison of RNA-centered and ATAC-centered integration by sciCAN 111

Figure 4.4. Integration learned by sciCAN preserves hematopoietic hierarchy 113

Figure 4.5. Activity-expression velocity of the hematopoietic hierarchy 115

Figure 4.6. sciCAN identifies common response after CRISPR perturbation 117

Figure 5.1. Artificial alignments by diagonal methods in 5 scenarios 133

Figure 5.2. Conceptual models for integration of multimodal single-cell data 135

Figure 6.1. Unsupervised neural network model to learn spatial feature 146

Figure 6.2. Single-cell 3D genome data analysis with 3 different levels of features 148

INTRODUCTION THE DECADE OF SINGLE-CELL TECHNOLOGY

The emergence of single-cell technologies

The very first single-cell transcriptomics study came out in 2009 with a surprise that single-cell transcriptomics could reveal rare cell types that are obscured by the bulk transcriptome profiling (Tang et al., 2009). Soon, this pioneer study set off an initial wave of developing transcriptomics protocols at single-cell resolution (Hashimshony et al., 2012; Islam et al., 2011; Jaitin Diego et al., 2014), including SMART-seq (Goetz and Trimarchi, 2012) and SMART-seq2 (Picelli et al., 2013) that are still widely used nowadays. However, these protocols were only able to profile from 100 to 1,000 cells at once and still required a certain amount of human labor. In the following years, we witnessed a new wave of single-cell transcriptomics technologies, which really leveled up single-cell transcriptomics to a large scale. Two droplet-based high-throughput single-cell transcriptomics platforms, Drop-seq (Macosko et al., 2015) and inDrop (Klein et al., 2015), came out on *Cell* on the same day. A single cell is encapsulated into a droplet and the droplet flows in a microfluidics device. This elegant mechanic design scales profiling capacity to more than 10,000 (about 10 times improvement) with less human labor. Other high-throughput single-cell transcriptomics technologies also emerged by using different strategies to scale up output, for example depositing cells into micro or picolitre wells (Gierahn et al., 2017; Han et al., 2018; Rosenberg et al., 2018b). Furthermore, commercial platforms, like 10X and Parse Biosciences, are widely used by different research laboratories now and are becoming the major platforms for generating large-scale single-cell transcriptomics data (Zheng et al., 2017). With these single-cell transcriptomics platforms in hand, a global effort of building up single-cell databases in the past decade brought the concept of “cell

atlas" to life (Rozenblatt-Rosen et al., 2017). These atlas studies covered different tissues (Litviňuková et al., 2020; Muraro et al., 2016; Stewart Benjamin et al., 2019; Travaglini et al., 2020), multiple species (Baron et al., 2016; Han et al., 2018; Jin et al., 2020; Li et al.; Schaum et al., 2018), and different developmental stages (Almanzar et al., 2020; Cao et al., 2020a; Farrell Jeffrey et al., 2018; Haniffa et al., 2021; Wagner Daniel et al., 2018), and too many to be listed here. Collectively, they provide a valuable resource for biomedical research and could unravel comprehensive understanding of transcriptional definitions of cell types in cellular and molecular biology.

Our curiosity is not just limited to understanding transcriptomes, by asking which gene is turned on and off in a certain cell type at a certain time point, but also extends to why and how it happens, from genetics to epigenetics, and even to proteomics. One major ambition of this field is to predict cell fates by combining information we could gather from multiple views. Development of single-cell technologies for other modalities, like genomic variant and chromatin accessibility, benefited a lot from strategies we already learned in single-cell transcriptomics studies, and single-cell transcriptomics technologies were transferred well for high-throughput sequencing of other modalities (Buenrostro et al., 2015; Cusanovich et al., 2015; Lodato et al., 2015). The single-cell community started to give more and more attention to construct atlas data of these modalities beyond transcriptome. The first single-cell atlas of mouse chromatin accessibility became available and provided a more detailed regulatory landscape combined with mouse atlas of transcriptome (Cusanovich et al., 2018). A more recent collaboration effort across different institutes made a comprehensive cell atlas come true in mouse brain. This collaboration

ended up with an atlas data including transcriptome, chromatin accessibility, and DNA methylation (Liu et al., 2021; Yao et al., 2021). Till now, the single-cell community across the globe has built a comprehensive toolbox that enables us to study genetic variant, epigenetic modification, chromatin structure, gene expression, surface markers, and so on (Stuart and Satija, 2019).

The journey never stops with developing single-cell technologies for different modalities independently. Scientists are even more ambitious to profile multiple views of cell status at the same time, partially because computational integration of single-cell data from independent experiments could not serve as the ground truth for interactions between different levels in individual cells. This led to inventions of joint profiling single-cell technologies. For instance, sci-CAR, SNARE-seq, and SHARE-seq can simultaneously profile chromatin accessibility and gene expression for thousands of single cells (Cao et al., 2018; Chen et al., 2019; Ma et al., 2020b). scMethyl-HiC and sn-m3C-seq can profile DNA methylation and 3D chromatin structure at the same time at single-cell resolution (Lee et al., 2019; Li et al., 2019). Paired-Tag jointly pulls out information of different histone modifications and transcriptome from the same single cell (Zhu et al., 2021). These new joint-profiling single-cell technologies, nevertheless, lift up single-cell technology to a brand-new level and open a new door to diverse scientific questions as well as presenting challenges for data integration.

In this thesis, I will focus on addressing data integration for single modality data, specifically single-cell transcriptome data. Then, I will move on to multimodal data integration. Four chapters will cover 4 different computational methods I developed as the

first author. These 4 methods also fall into different categories of computational principles for data integration, and I am about to introduce their backgrounds in more detail.

Batch effects! Batch effects! Batch effects!

Looking back the evolution history of single-cell technology, we could summarize that method development for single-cell transcriptome is predominant over all single-cell modalities. As briefly introduced above, multiple single-cell transcriptome platforms exist and are widely used by different research groups. Over the past decade, single-cell transcriptome datasets were also generated with all these platforms. Instead of analyzing these datasets separately, the single-cell community is gradually realizing the importance of bringing all possible single-cell transcriptome data that come from the same tissue but under different biological conditions for integrative analysis. Thus, integration of single-cell transcriptome data becomes a trending topic in the single-cell community. When we deal with integration for single-cell transcriptome data, we often run into the problem that major differences among datasets come from platform differences. For instance, nanodroplet-based platforms are intrinsically different from their picolitre-well-based counterparts, in terms of how cells are captured. Not to mention, within nanodroplet-based platforms, methods like Drop-seq and inDrop, can produce biologically irrelevant variations due to other technical differences. More than platform differences, construction of cell atlas database is result of global collaboration across multiple institutes with many different personnel handling data generations. This kind of large-scale collaboration involving hundreds of and even thousands of people could inevitably return data with biologically irrelevant variations from many unknown sources. Indeed, this problem,

collectively called the batch effects, is quite common in integration of single-cell transcriptome data.

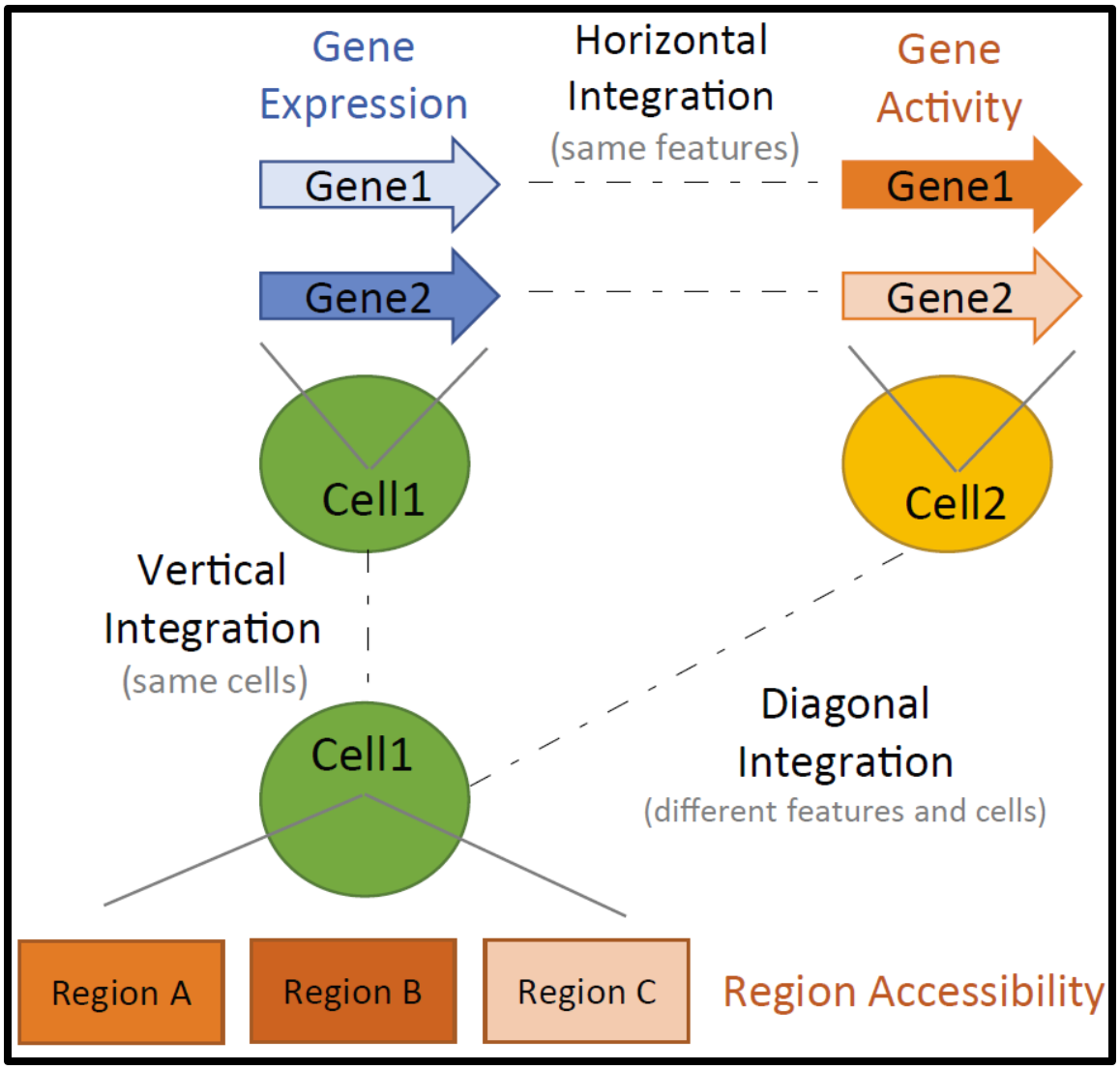
Batch effects are usually the prominent variation when data from multiple sources are compared. Removing batch effects is a critical and necessary step before performing any biological interpretation. Along the development timelines of single-cell transcriptome platforms, many computational approaches were proposed to address batch effects. In the first wave, we observed many batch-correction methods were based on conventional machine learning approaches. These methods work well in a range of cases of single-cell transcriptome data integration. Methods including Seurat, Harmony, and LIGER received tremendous success and are embraced greatly by single-cell community(Korsunsky et al., 2019; Liu et al., 2020; Stuart et al., 2019). For example, Harmony learns the joint representation through an iterative k-means clustering, and the outcome is a linear correction function that transforms the original principal components (PCs) to the batch-corrected PCs (Korsunsky et al., 2019). Seurat, on the other hand, uses canonical correlation analysis (CCA) to learn the shared latent space among batches. Seurat first identifies cell anchors between two batches to learn a mutual neighborhood graph. Then, it computes a projection that brings all other cells to this shared latent space. Because of its "anchor" design, Seurat needs pairwise computation of anchor points when datasets come from more than two sources (Butler et al., 2018). Both the iterative k-means clustering in Harmony and pairwise CCA need intensive calculation that consumes large computation resources. As the quantity of data grows exponentially, from handling data with thousands of cells to millions of cells, we start to demand methods that can handle large-scale data

quickly and efficiently. In recent years, we witnessed another wave of method development for batch correction, which is using neural networks. Benefiting from the power of GPU and training data in a mini-batch manner, deep learning models gained growing attention while showing the single-cell community its capability for large-scale data. Nowadays, you can search deep-learning-based batch-correction methods and end up getting a long list (Bahrami et al., 2020; Dincer et al., 2020; Kimmel and Kelley, 2021; Lakkis et al., 2021; Lopez et al., 2018; Lotfollahi et al., 2021; Shaham et al., 2017; Wang et al., 2019; Wang et al., 2021a; Xu et al., 2021a). Among all these deep-learning-based methods, variational autoencoder is a common neural network architecture, and they are trained in adversarial manner with batch information as labels.

Regardless of conventional machine learning or deep learning approaches, with either simple or sophisticated modeling, these batch-correction methods approach to the solution and address batch effects to varying degrees. One benchmark study shows that these methods are not applicable for all scenarios (Tran et al., 2020). For the purpose of generalization, we may need to figure out in what form the batch effects exist within single-cell transcriptome data. Resolving the mysterious form of batch effects, we can generalize a simple approach for batch correction. In Chapter I and II, I will describe two simple methods I developed to resolve batch effects. In Chapter III, I will introduce another simple framework for learning representation, and this framework also addresses batch effects for multi-source single-cell transcriptome data. Combining Chapter I to III, I want to bring out a potential explanation of why these methods work to discussion.

From single modality to multi-modalities

With revolution of multi-omics single-cell technologies, single-cell computational analysis also jumps into the multi-omics era. Single-cell community has been giving more and more attention to integrating multi-omics single-cell data, as this new research domain promises us to understand a complex cellular system from different viewpoints, such as gene expression, epigenetic modification, and chromosome structure. However, different types of 'omics data present the same cellular system in different data formats. They do not necessarily share the same features, though we should keep in mind that features across different modalities are often highly correlated. For instance, transcriptomics describes expression of genes, while epigenomics measures histone modifications or accessibility across all regions of the genome. This feature discrepancy presents the first barrier to bringing multimodal single-cell data together. Besides feature discrepancy, each modality preserves shared information as well as something distinct. How to wisely integrate multimodal data without loss of distinctness of each modality is another challenge we are facing. In recent years, many integration methods have been published to address different scenarios of multimodal single-cell data integration. A recent review summarized three categories of multimodal single-cell data integration (Argelaguet et al., 2021). Of these categories, "horizontal integration" methods require anchored features to align up different modalities, while "vertical integration" methods anchor different modalities with shared cells. The "diagonal integration" approach is claimed to require neither anchoring cells nor features for integration, presenting a distinct advantage over horizontal and vertical methods (Box 1).



Box 1
Key Terms:
Modality: A type of biological measurement, such as gene expression, chromatin accessibility, 3D chromosome contacts, or shape descriptors from imaging.
Feature: An entity to which measurements are assigned, such as genes, promoters, genomic bins, or positions in an image.
Horizontal, vertical, and diagonal integration: Creating a shared representation space for single cell measurements from multiple modalities anchoring on features (horizontal), cells (vertical), or neither (diagonal). See schematic above.

To accomplish horizontal integration, we can transfer the use of batch-correction methods. Methods I mentioned above, like LIGER, Harmony, and Seurat, were already extensively tested in the task of integrating single-cell transcriptome data with single-cell chromatin accessibility or single-cell DNA methylation data (Forcato et al., 2021; Liu et al., 2020; Stuart et al., 2019; Yang and Michailidis, 2016). To use these tools for multimodal integration, the common practice is converting chromatin accessibility and DNA methylation to a gene-expression-like format, for the purpose of matching features with single-cell transcriptome. Once features are matched, resolving modality difference wouldn't be different from removing batch effects, by considering modality difference as a form of batch effect. The reason why these horizontal methods would work is primarily exploiting correlations within shared features. However, each modality also preserves distinct features, and this conversion inevitably distorts, and discards information obtained from the original format.

In some cases, transforming different modalities to have shared features could discard informative features to a large degree. For example, relevant histone modifications can occur far from genes, and therefore assigning histone modifications to gene features to match transcriptomics will by necessity loss information. Meanwhile, feature matching could be not applicable in cases like integrating transcriptome data with chromatin structure data. To overcome this problem, we came to the second category, vertical integration, for a solution. Because vertical integration anchors modalities with shared cells, application of vertical integration is also limited to cases when multimodal information from the same cell is known. Fortunately, technological breakthroughs in joint

profiling make it possible to capture multiple data types from the same single cell. With these joint-profiling technologies, we are able to pull out paired gene expression/chromatin accessibility, paired DNA methylation/chromatin structure, and so on (Cao et al., 2018; Chen et al., 2019; Lee et al., 2019; Li et al., 2019; Ma et al., 2020b; Zhu et al., 2021). These joint-profiled data could serve as reference to train vertical integration methods (Jin et al., 2020; Wu et al., 2021).

Since horizontal and vertical methods require either anchored features or anchored cells, their applications are not applicable for all cases of multimodal integration. Therefore, there is extensive interest in diagonal integration because it doesn't use any prior knowledge. The existing diagonal integration methods split the task into two parts (Cao et al., 2020b; Cao et al., 2021; Demetci et al., 2020; Yang et al., 2021). One is learning lower representation for each modality. The lower representation needs to preserve biological variation through distinguishing different cell types in each modality. The other task is modality alignment. Methods should close the gap between modalities by aligning up the same cell types. However, these two tasks seem to be disjoint in most existing methods, raising a doubt about whether these diagonal methods reach to a solution that is biological rather than simply mathematically optimal.

In chapter III and IV, I will discuss two multimodal integration methods I have developed. One is vertical integration method, and the other falls into the category of horizontal integration. In the final chapter V, I will quantify pitfalls and discuss potential future directions of diagonal integration.

CHAPTER I
MARKER-BASED AUTOMATIC CELL-TYPE ANNOTATION FOR GENERAL SINGLE CELL EXPRESSION DATA

A version of this chapter is a manuscript by Yang Xu, Simon J. Baumgar, Christian M. Stegmann, and Sikander Hayat. This manuscript was published in *Bioinformatics*.

Y.X. and S.H. planned and designed the study. Y.X. performed the computational analysis. Y.X. and S.H. analyzed and interpreted the data and wrote the manuscript. J.B. and C.M.S. edited the manuscript and advised on data interpretation. All authors read and approved the manuscript.

Abstract

Accurately identifying cell-types is a critical step in single-cell sequencing analyses. Here, we present **m**arker-based **a**utomatic **c**ell-type **a**nnotation (**MACA**), a new tool for annotating single-cell transcriptomics datasets. We developed MACA by testing 4 cell-type scoring methods with 2 public cell-marker databases as reference in 6 single-cell studies. MACA compares favorably to 4 existing marker-based cell-type annotation methods in terms of accuracy and speed. We show that MACA can annotate a large single-nuclei RNA-seq study in minutes on human hearts with ~290k cells. MACA scales easily to large datasets and can broadly help experts to annotate cell types in single-cell transcriptomics datasets, and we envision MACA provides a new opportunity for integration and standardization of cell-type annotation across multiple datasets.

Introduction

Identifying constituent cell-types in a single-cell dataset is fundamental to understand the underlying biology of the system. Many computational methods have been proposed to automatically label cells, and a benchmark study shows that a standard Support Vector

Machine (SVM) classifier outperforms most other sophisticated supervised methods and can achieve high accuracy in cell-type assignment (Abdelaal et al., 2019). However, due to lack of ground-truth in most single cell studies, supervised classification approaches are not feasible and may not be generalized for new single cell studies with different experimental designs. Therefore, unsupervised clustering approaches are still the predominant options for single-cell data analysis (Lähnemann et al., 2020). Unsupervised approaches usually require human assistance in both defining clustering resolution and manual annotation of cell-types. This results in cell-type annotation being time-consuming and less reproducible due to human inference. As more single cell studies are available, summarizing markers identified in these studies to construct a marker database becomes an alternative approach for automatic cell-type annotation. For example, PanglaoDB (Franzén et al., 2019b) and CellMarker (Zhang et al., 2019b) are two marker databases that summarize markers found in numerous single cell studies and cover a broad range of major cell-types in human and mouse. Meanwhile, NeuroExpresso (Mancarci et al., 2017) is a specialized database for brain cell-types. Taking advantage of those databases for robust cell-type identification, we present MACA, a **m**arker-based **a**utomatic **c**ell-type **a**nnotation method and show how MACA automatically annotates cell-types with high speed and accuracy.

Method development

MACA takes as input expression profiles measured by single cell or nuclei RNA-seq experiments. MACA calculates two cell-type labels for each cell based on 1) an individual cell expression profile and 2) a collective clustering profile (Figure 1.1A). From these, a

final cell-type label is generated according to a normalized confusion matrix. MACA first computes cell-type scores for each cell, using a scoring method based on a marker database or user-defined marker lists. The scoring method uses the raw gene count to calculate a cell-type score for each cell, according to gene markers of this cell-type. This results in converting a gene expression matrix to cell-type score matrix. Then, MACA generates a label (Label 1) for each cell by identifying the cell-type associated with the highest score. Independently, using the matrix of cell-type scores as input, the Louvain community detection algorithm is applied to generate Label 2, which is a clustering label to which a cell belongs (Blondel et al., 2008). Since the number of cell types is usually unknown, MACA tries clustering at greater resolution to over-cluster cells into many small but homogeneous groups.

Both Label 1 and Label 2 serve complimentary functions. Label 1 is assigned on a per-cell basis which may result in incorrectly annotating many cells due to noisiness in the maximum cell-type score for each cell. This may occur when the putative cell-type feature is covered up by ambient RNAs from dominant cell-types (Pliner et al., 2019). On the other hand, Label 2 is likely to suffer from a common problem in single cell RNA-seq clustering analysis, where cells may share the same dominant features, even though they have been clustered into different groups because of subtle differences. Additionally, results from a clustering analysis can often vary since clustering is non-deterministic. Due to its dependence on user's decisions, mostly the choices of clustering resolution and neighborhood size.

To address these issues, MACA combines Label 1 and Label 2 to get a comprehensive cell-type annotation by mapping Label 2 to Label 1 through a normalized confusion matrix (Figure 1.1B). In the confusion matrix $\boldsymbol{C}$, $\boldsymbol{c_{i,j}}$ represents the number of cells that were clustered as the $\boldsymbol{i^{th}}$ cluster in Label 2 and labeled as the $\boldsymbol{j^{th}}$ cell-type in Label 1. The basic assumption of mapping Label 2 to Label 1 through a confusion matrix is that cells with the same clustering label (Label 2) should have the same cell-type label (Label 1). Ideally, if cells were identified to be in the same cluster, they should all share the same cell-type, and this cell-type has the highest score for cells in that cluster. However, in real data, this is rarely the case, as we argued above. Therefore, using a confusion matrix, we look for consensus between Label 1 and Label 2, by searching for the highest cell-type score in each cluster. Here, we compute the normalized confusion matrix $\boldsymbol{C_n}$ through dividing confusion matrix $\boldsymbol{C}$ by the size of the cluster: $\boldsymbol{c_{i,j}} = \frac{\boldsymbol{c_{i,j}}}{\sum_{j=1}^{N} \boldsymbol{c_{i,j}}}$, and we search for column number with the largest value for each row (Figure 1b). If $\boldsymbol{max_j(c_{i,j}) \geq 0.5}$, the $\boldsymbol{i^{th}}$ cluster would be assigned as the $\boldsymbol{j^{th}}$ cell-type, as more than 50% of cells in the $\boldsymbol{i^{th}}$ cluster are labeled as the $\boldsymbol{j^{th}}$ cell-type (Case 1). For cases where $\boldsymbol{max_j(c_{i,j}) < 0.5}$, it is likely that cell identities of some cells were covered up by ambient RNAs from dominant cell-types (Case 2). Therefore, MACA records significant or at least the top-3 cell-types for each cell in the $\boldsymbol{i^{th}}$ cluster based on cell-type scores. To find significant cell-types for each cell, we get a distribution of scores of all cell-types for each cell and define those cell-types as significant if their z-scores > 3. If the number of significant cell-types is less than 3, we would keep the top-3 cell-types. Doing this can retrieve more potential cell-type labels for this cluster, and each cell will contribute at least 3 candidates into a pool of candidate cell-types for this

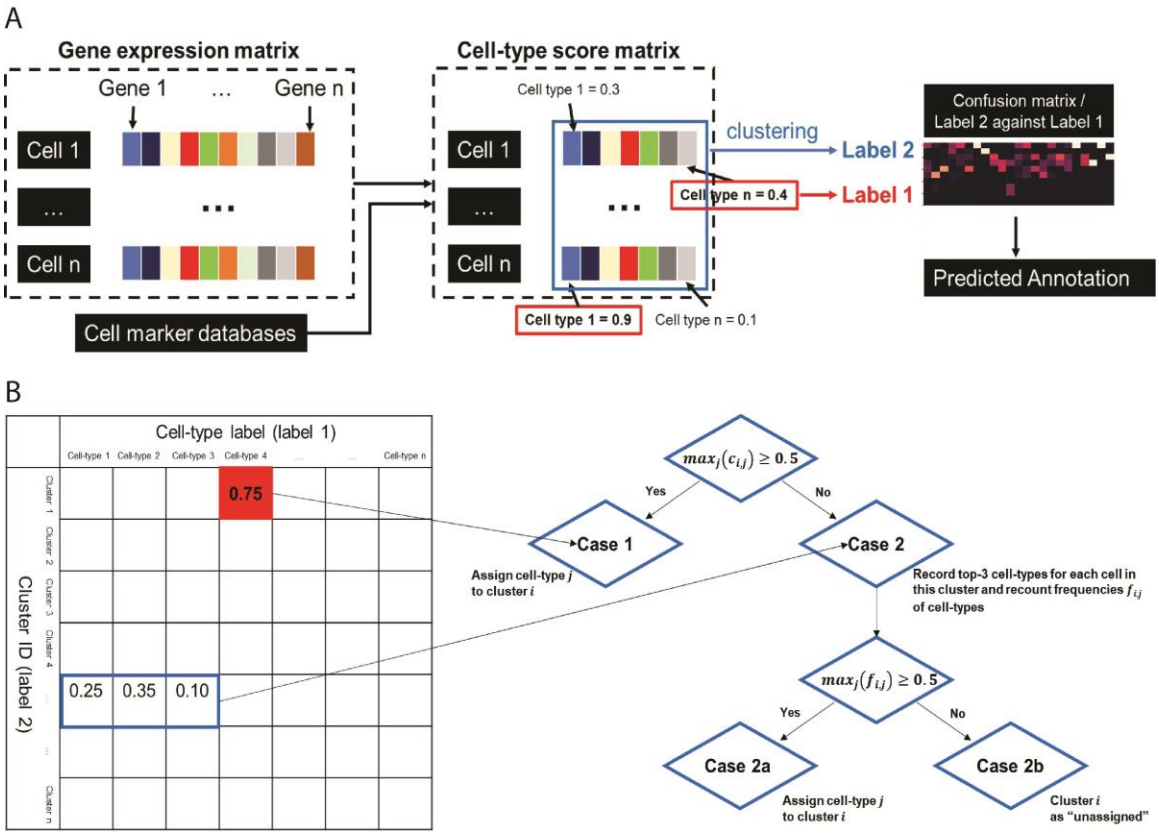


Figure 1.1. Schematic workflow of MACA. A, MACA converts gene expression matrix into cell-type score matrix based on cell marker database. MACA generates Label 1 by using max function and Label 2 by over-clustering all cells into small groups. MACA finally maps Label 2 to Label 1 via confusion matrix. B, Use of confusion matrix for cell-type annotation. How cluster label is assigned a cell type is shown in the panel on the right.

cluster. Then, MACA calculates frequency of each candidate cell-type in this pool and assigns the $\boldsymbol{i^{th}}$ cluster as the cell-type with the highest frequency if the frequency exceeds half the size of the cluster ($\boldsymbol{max_j(f_{i,j}) \geq 0.5}$) (Case 2a). Otherwise, the $\boldsymbol{i^{th}}$ cluster would be labeled as "unassigned" ($\boldsymbol{max_j(f_{i,j}) < 0.5}$) (Case 2b), which is the case that cells in this cluster do not have an agreement on which cell-types they belong to. For the choice of 0.5, we will show our examination in the next Results section. As we mentioned before, clustering-based cell-type identification largely depends on user's choice, for example the choices of clustering resolution and neighborhood size. Therefore, the outcome may vary among different users. To have a more reproducible outcome, we cluster cells with different clustering parameters to get multiple clustering assignments (Label 2s). Repeating the procedure of mapping Label 2 to Label 1 will enable us to get an ensemble annotation through voting, and this ensemble annotation is less influenced by a single clustering choice. Using ensemble approach also offers a naïve way of scoring MACA-based cell-type predictions. Users can set up a threshold to filter cells whose annotations are less consistent in outcomes of different clustering trials. In this study, we generated clusters using Louvain method with 3 different resolutions and 3 different numbers of neighborhood, which results in 9 different clustering labels (Label 2s). After mapping these 9 Label 2s to Label 1, we generated 9 cell-type annotations. Then, we used a voting approach to get the final annotations (the highest votes from the 9 annotations). Users can also increase the number of clustering trials to have a larger voting pool for annotation ensemble or decrease the number to save computation time. Back to converting gene expression matrix to cell-type score matrix, we collected 4 different scoring methods that

were proposed to do the conversion. These scoring methods are either named by authors, or we named them after the last name of the first author. PlinerScore was a part of Garnett that was designed to annotate cell-types through supervised classification(Pliner et al., 2019). The uniqueness of PlinerScore is the use of TF-IDF transformation to deal with specificity of a gene marker and a cutoff to deal with issue of free mRNA in single-cell RNA-seq data. AUCell comes from SENIC, which uses gene sets to quantify regulon activities of single-cell expression data (Aibar et al., 2017). In this study, AUCell quantifies the enrichment of every cell-type as an area under the recovery curve (AUC) across the ranking of all gene markers in a particular cell. This assessment is cell-wise and is different from PlinerScore that requires transformation of the whole dataset. Both CIM and DingScore simply use the total expression of all gene markers of a particular cell-type as the cell-type score (Ding et al., 2020; Efroni et al., 2015). CIM normalizes the total expression by multiplying a weight that is defined as the number of expressed gene markers divided by the number of all gene markers of this cell-type. DingScore, on the other hand, normalizes the total expression of one cell-type by dividing total expression of all genes. Since some cell-types have a longer list of marker genes than others, cell-types with more marker genes in the database would have larger cell-type scores. Normalization in CIM was considered to address this issue. However, PlinerScore and DingScore were not intentionally designed to cope with unbalanced marker lists. To deal with this issue, we did a similar processing to normalization in CIM, which is dividing the score of each cell type by the number of expressed markers in that cell type. However, AUCell is a completely different approach from the other 3 scoring methods, which does not simply sum up values

of marker genes for a given cell-type. So, we ran AUCell without extra processing for returned values.

In practice, we build MACA in the analysis pipeline of Scanpy, and MACA takes data in the format of "anndata" in Python(Wolf et al., 2018). Expression data are preprocessed through cell and gene filtering, and transformed by LogNormlization method, the common practice in single cell analysis. Then, the user provides marker information in the form of Python dictionary, and MACA transforms gene expression matrix to cell-type score matrix. Next, annotation by MACA can be summarized into 4 steps as shown in Figure 1: 1) Louvain clustering to generate Label 2; 2) Generating Label 1 via max function; 3) Mapping Label 2 to Label 1 through normalized confusion matrix; 4) Repeating step 1 to 3 to have ensembled annotation.

Result

The key component for optimal performance of MACA is constructing cell-type scores from the gene expression matrix. We investigated 4 scoring methods that have been proposed to transform gene expression matrix to cell-type score matrix (Aibar et al., 2017; Ding et al., 2020; Efroni et al., 2015; Pliner et al., 2019), and we tested these methods with 2 public marker databases (Franzén et al., 2019b; Zhang et al., 2019b) in 6 single cell studies comprised of 3000 to 20000 cells (Baron et al., 2016; Cui et al., 2019a; Tian et al., 2019; Vieira Braga et al., 2019; Wang et al., 2020c; Zheng et al., 2017), which include 3 benchmark datasets (Abdelaal et al., 2019). To evaluate these annotation outcomes, we used Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Both ARI and NMI are calculated by measuring similarity or agreement between our annotations and

authors' annotations. For the 3 benchmark datasets, authors' annotations would be the ground truth label, while authors' annotations in the other 3 datasets are at least created under careful investigation. Therefore, use of ARI and NMI, in this case, is to show how well we can reproduce authors' outcomes. We found annotations using PlinerScore with markers in PanglaoDB have the largest agreement with authors' annotations for all 6 datasets, in terms of both ARI and NMI (Table 1.1). Therefore, MACA uses PanglaoDB with PlinerScore as the main marker database and scoring method, respectively.

Next, we seek to compare MACA with other existing marker-based annotation tools. CellAssign and SCINA are two computational methods that have been proposed for automatic cell-type assignment (Zhang et al., 2019a; Zhang et al., 2019c). Both methods rely on statistical interference to compute the probabilities of cell types, which are time- and computation- intensive. Recently, Cell-ID was released for extraction of gene signature as well as cell-type annotation (Cortal et al., 2021). We also noticed scCATCH and SCSA, which are both cluster-based annotation tools (Cao et al., 2020c; Shao et al., 2020). Both scCATCH and SCSA require identifying differential marker genes for each cluster via a statistical test implemented in Seurat and then matching identified cluster markers to marker database (Butler et al., 2018). Here, we compared MACA with CellAssign, SCINA, Cell-ID, and scCATCH using these 6 single cell studies and cell markers in PanglaoDB. We tested MACA, CellAssign, SCINA, Cell-ID, and scCATCH on a workstation with 16-core CPU and 64GB memory. MACA can finish annotation within 1 minute (cells around 3,000) and less than 2 minutes for a relatively large dataset (cells up to 20,000 cells). On the datasets used and on our computational resources, scCATCH and Cell-ID took longer

Table 1.1. Performance of MACA, CellAssign, SCINA, Cell-ID, and scCATACH in 6 scRNA-seq datasets, measured by ARI and NMI. 8 different settings of MACA include using 4 cell-type scoring methods (PlinerScore, AUCell, CIM, and DingScore) with 2 marker databases (PanglaoDB and CellMarker).

ARI	PBMC (Zheng et al., 2017)	CellBench (Tian et al., 2019)	Pancreas (Baron et al., 2016)	Heart (Wang et al., 2020)	Heart (Cui et al., 2019)	Lung (Vieira et al., 2019)
PanglaoDB+PlinerScore	**0.95**	**0.92**	**0.90**	**0.71**	**0.61**	**0.45**
PanglaoDB+AUCell	0.04	0.00	0.78	0.39	0.47	0.29
PanglaoDB+CIM	0.28	**0.65**	**0.90**	0.27	0.30	0.33
PanglaoDB+DingScore	**0.83**	**0.74**	0.69	0.07	0.44	0.20
CellMarker+PlinerScore	0.38	0.43	0.27	**0.57**	0.13	0.21
CellMarker+AUCell	0.29	0.52	0.32	0.34	0.09	0.14
CellMarker+CIM	0.24	0.60	0.54	**0.56**	0.07	0.09
CellMarker+DingScore	0.22	0.55	0.38	0.37	0.19	NA
SCINA	0.46	0.63	**0.89**	0.13	**0.55**	0.31
CellAssign	NA	0.00	**0.89**	0.15	**0.53**	0.26
Cell-ID	0.50	0.17	0.57	0.10	0.49	**0.35**
scCATCH (best)	**0.62**	0.56	0.86	0.04	0.14	**0.60**
scCATCH (average)	0.57	0.40	0.66	0.04	0.05	**0.35**

NMI	PBMC (Zheng et al., 2017)	CellBench (Tian et al., 2019)	Pancreas (Baron et al., 2016)	Heart (Wang et al., 2020)	Heart (Cui et al., 2019)	Lung (Vieira et al., 2019)
PanglaoDB+PlinerScore	**0.89**	**0.92**	**0.88**	**0.59**	**0.62**	**0.59**
PanglaoDB+AUCell	0.09	0.00	0.79	**0.41**	0.50	0.31
PanglaoDB+CIM	0.51	**0.80**	**0.88**	0.30	0.44	0.40
PanglaoDB+DingScore	0.74	**0.85**	0.70	0.10	0.47	0.33
CellMarker+PlinerScore	0.44	0.64	0.57	**0.51**	0.32	0.42
CellMarker+AUCell	0.23	0.67	0.46	0.32	0.33	0.17
CellMarker+CIM	0.49	0.78	0.73	**0.41**	0.31	0.21
CellMarker+DingScore	0.43	0.73	0.60	0.34	0.33	0.08
SCINA	0.54	0.71	0.84	0.07	**0.54**	0.46
CellAssign	NA	0.06	**0.86**	0.08	0.51	0.49
Cell-ID	0.67	0.38	0.74	0.08	**0.55**	0.58
scCATCH (best)	**0.77**	0.70	0.84	0.05	0.30	**0.73**
scCATCH (average)	**0.75**	0.62	0.75	0.04	0.12	**0.63**

# of cell-types	PBMC (Zheng et al., 2017)	CellBench (Tian et al., 2019)	Pancreas (Baron et al., 2016)	Heart (Wang et al., 2020)	Heart (Cui et al., 2019)	Lung (Vieira et al., 2019)
MACA	8	6	11	8	7	13
SCINA	14	14	17	16	23	41
CellAssign	NA	9	17	18	24	31
Cell-ID	33	55	48	35	37	63
scCATCH (best)	9	5	10	3	3	16
Author's annotation	5	5	14	5	9	13

than MACA to compute annotations and ranks as the second and third fastest. In our hands, SCINA took around 20-minute time to finish annotation for a large dataset, and CellAssign took the longest time to complete cell-type assignment and failed to annotate data with > 20,000 cells due to lack of memory (Table 1.2). Because annotation by scCATCH needs clustering first and differential marker identification is highly affected by clustering outcome, the investigator will need to do a thorough investigation to make sure that clustering is not overdone or underestimated. In this study, we reported the highest and the averaged outcomes of scCATCH in each dataset. Comparing these results with manual annotations from the authors, we found 1) MACA labels cells had a higher consensus than CellAssign, SCINA, Cell-ID, and scCATCH, in terms of both ARI and NMI, and 2) MACA and scCATCH identify similar numbers of cell-types to author's annotations, while the other 3 methods, especially Cell-ID, report overall more different cell-types (Table 1.1). The low ARIs and NMIs of CellAssign and Cell-ID can be counted as results of 1) many "unassigned" cells and 2) exceeding numbers of different cell-types over the numbers reported by authors. It is important to note that other methods compared here were run on their default parameters.

In future, parameter tuning of those methods on a computer with higher memory should be carried out for a comprehensive benchmarking on many datasets. Finally, to better evaluate annotations, we used a machine learning approach to assess cell-type assignment. Training classifiers was recently proposed by Miao et al. to assist in finding a good clustering resolution (Miao et al., 2020), and we adopt this idea to evaluate our annotations. Basically, if the annotation is good enough, we can train a classifier to predict cell type

Table 1.2. Runtime of 5 annotation tools across 6 benchmark datasets. MACA was much shorter than other methods tested in this benchmark, running a workstation with 16-core CPU and 64GB memory.

Runtime (in min)	PBMC (Zheng et al., 2017)	CellBench (Tian et al., 2019)	Pancreas (Baron et al., 2016)	Heart (Wang et al., 2020)	Heart (Cui et al., 2019)	Lung (Vieira et al., 2019)
MACA	1.5	1	1	1	1	1
SCINA	18	5	16	8.5	8	19
CellAssign	NA	30	140	120	180	160
Cell-ID	15	2	5	5	2	10
scCATCH	7.5	NA	1.5	1	1	6.5

using gene-expression values with high accuracy. Conversely, if there are many wrong labels, it would be hard for a classifier to make the right decision. We performed 5-fold cross-validated training, where we split one dataset into 4-fold training set and 1-fold testing set and trained a SVM classifier on the training sets and applied the classifier to predict labels for the testing set. This procedure repeats 5 times to get a mean accuracy. Instead of treating authors' annotations as ground truth, this machine-learning evaluation provides an independent angle to judge annotation quality. Indeed, MACA achieves high concordance with authors' reported annotations and higher mean of accuracies than other methods (Table 1.3). Of note, high accuracy of SVM classifier is not equal to correctness of annotation. Meanwhile, ARI and NMI reports similarity between two annotations but cannot reflect the difference of annotation resolution. For example, MACA may return less cell-types than authors. Moreover, annotation resolution of MACA highly depends on the number of cell-types in the marker database, and it is likely that MACA cannot annotate some rare subtypes that do not show up in the marker database.

As we mentioned above, using ensemble approach also offers user an option to filter cells whose annotations are less consistent in outcomes of different clustering trials. However, it also causes loss of cells for downstream analysis, like cellular composition analysis. To find a good balance between having higher annotation quality and keeping most cells for downstream analysis, we tested threshold of voting from 1/9 to 9/9, where the numerator means the minimum number of votes required to keep the cell-annotation. With 1/9, all cells will be kept, with 2/9, cells with annotations with at least 2 votes will be kept, while only cells that have the same annotation across 9 clustering trials will be

Table 1.3. Mean accuracy of 5-fold SVM classifier. SVM classifiers were trained with labels from original reports, or generated by MACA, CellAssign, SCINA, Cell-ID, and scCATCH. Each data was split into 5 folds. Classifiers were trained on 4 folds, and they used the rest 1-fold to report accuracy. Results here came from the mean accuracy of 5-fold training. The highest accuracies obtained for that dataset is shown in bold. For most datasets, PanglaoDB+PlinerScore and authors' annotations achieve the highest accuracy in SVM classification.

	PBMC (Zheng et al., 2017)	CellBench (Tian et al., 2019)	Pancreas (Baron et al., 2016)	Heart (Wang et al., 2020)	Heart (Cui et al., 2019)	Lung (Vieira et al., 2019)
PanglaoDB+PlinerScore	**0.93**	0.99	**0.97**	**0.90**	**0.96**	**0.85**
PanglaoDB+AUCell	0.67	NA	0.89	0.85	0.89	0.57
PanglaoDB+CIM	0.77	**1.00**	**0.97**	0.83	0.91	0.73
PanglaoDB+DingScore	0.91	0.98	0.89	**0.92**	0.93	0.83
CellMarker+PlinerScore	**0.96**	0.98	0.93	**0.95**	0.90	0.83
CellMarker+AUCell	0.48	0.96	0.60	0.74	0.65	0.44
CellMarker+CIM	0.69	**1.00**	0.92	0.78	0.84	0.55
CellMarker+DingScore	0.76	0.99	0.90	0.62	**0.97**	NA
SCINA	0.80	0.74	0.91	0.43	0.90	0.78
CellAssign	NA	0.74	0.94	0.41	0.77	0.77
Cell-ID	0.82	0.36	0.80	0.53	0.78	0.70
scCATCH (best)	0.91	0.92	0.92	0.56	**0.94**	**0.89**
Authors' annotation	**0.97**	**1.00**	**0.98**	0.89	0.93	**0.90**

considered if threshold is set up as 9/9. This evaluation may provide a reference for user to choose a threshold that serves user's need. Of note, we kept all cells in other evaluations. Particularly, all cells were used in benchmark with other methods. Here, we suggest setting up the threshold as 7/9. Next, we expect to show that annotation by MACA is applicable for most single cell RNA-seq platforms. We re-annotated PBMC data from a new study by Ding et al. (Ding et al., 2020). This data consists of two biological samples from 9 platforms. We found that 1) both PBMC samples have the same major cell-types, and these 9 platforms can successfully profile them (Figure 1.2A), and 2) annotation by MACA shows that all platforms profile similar cellular components for these two PBMC samples, except CEL-Seq2 (Figure 1.2B). These results are largely consistent to the original report (Ding et al., 2020).

Finally, we applied MACA to a single-nuclei RNA-seq dataset from all 4 chambers of the human heart, comprised of ~290k nuclei (Tucker et al., 2020a). MACA could annotate each of the 4 chambers comprising of ~80K cells each in < 6 mins. Annotations by MACA have major agreement with author's reported annotations with an average ARI and NMI of 0.63 and 0.76, respectively (Table 1.4). However, we also found some disagreements exist in annotation of cells in from left and right atria. Therefore, we investigated disagreement between MACA's and author's annotations, and found the biggest difference stems from disagreement in assignments for neuronal cells and lymphocytes, which are both small-population cell types in this dataset (1702 neuronal cells and 1503 lymphocytes out of ~290k). We found neuronal cells weren't revealed and author-reported lymphocytes were reported as memory T cells in MACA's annotation.

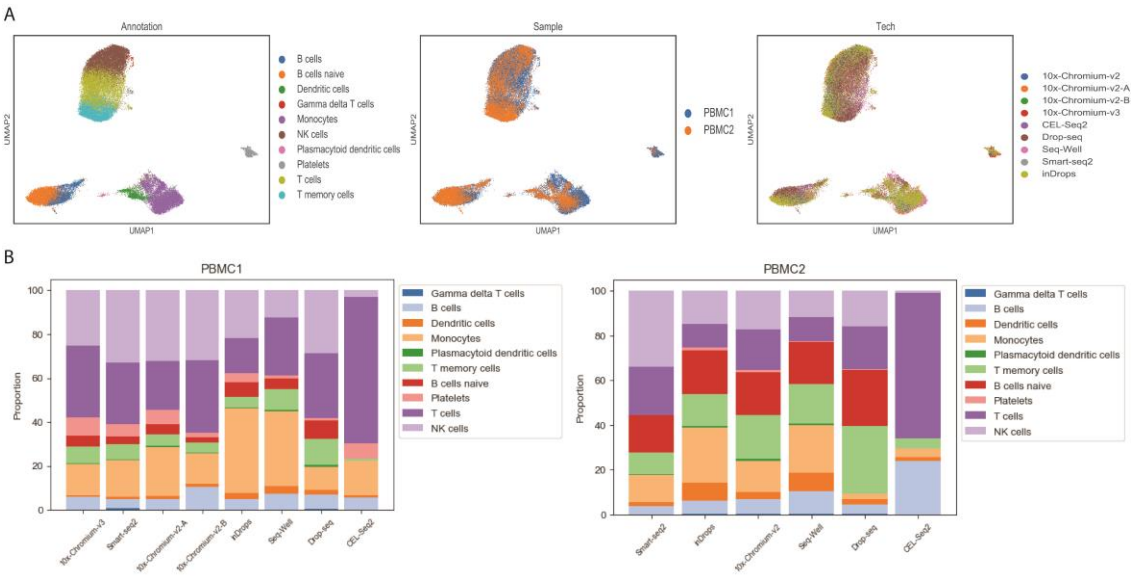


Figure 1.2. Integrated annotation of human PBMC and pancreas data across different single cell platforms. A, UMAP visualization of human PBMC data. Cells are colored according to annotation by MACA (left), source of sample (middle), and platform (right). UMAP dimension reduction is based on PlinerScore with PanglaoDB as marker database. B, cellular component analysis of human PBMC data. Proportion of each cell-type identified by MACA is calculated for each platform for two PBMC samples, separately.

Table 1.4. Performance of MACA in 4 human-chamber single-nuclei RNA-seq datasets, measured by ARI, NMI, and Accuracy of SVM classifier. RA: right atrium; LA: left atrium; RV: right ventricle; LV: left ventricle. The final MACA setting is using PanglaoDB as marker reference and PlinerScore as cell-type scoring method. The performance in human 4 chamber data (Tucker et al., 2020) is quantified by ARI and NMI against author's annotations (top 4 rows). Both MACA's and author's annotation were used to train SVM classifiers. Datasets were split into 5 folds. SVM classifiers were trained on 4-fold data and tested on the rest 1-fold. Means of accuracies were reported to show how reasonable MACA's and authors' annotations are (bottom 2 rows).

ARI	LV	RV	LA	RA
MACA	0.84	0.81	0.69	0.69

NMI	LV	RV	LA	RA
MACA	0.70	0.63	0.61	0.58

SVM_Accuracy	LV	RV	LA	RA
MACA	0.85	0.85	0.85	0.82
Authors' annotation	0.90	0.88	0.86	0.83

Conclusion

By default, MACA works with the list marker genes and cell-types present in PanglaoDB, but users can also input their own gene-lists. A major limitation of MACA is that it can only annotate cell-types that are pre-defined in the marker reference, but with more marker gene-sets becoming available with single-cell sequencing studies, we believe that MACA will be useful to annotate heterogeneous single-cell datasets. This points us two future directions to improve MACA. First, with more atlas studies that profile all sorts of biological systems, more refined markers for small cell populations can be defined, and MACA could reach finer annotation resolution by integrating markers from these new atlas studies. Second, weights of markers should be incorporated into the scoring method of MACA, for example marker specificity and expression strength. However, at the current stage, all markers have equal weights when they contribute to cell-type scores, and we believe that incorporating marker weights will be beneficial for accurate annotation. With a more refined marker database and cell-type scoring method, MACA would rapidly perform integrated annotation across multiple datasets, and this is very critical for downstream analyses like cellular component analysis across datasets under different conditions. In fact, we noticed that combining PlinerScore and PanglaoDB to generate new features has the advantages of correcting batch effects for integrated annotation across datasets. In the next chapter, we extended the use of MACA to standardization of cell-type annotation across datasets. Here, we conclude that MACA is a suitable tool for automatic cell-type annotation that can aid both experts and non-experts in rapid annotation of their single-cell datasets.

CHAPTER II
MARKER-BASED AND MODEL-FREE APPROACH FOR STANDARDIZATION AND INTEGRATION OF SINGLE-CELL TRANSCRIPTOMICS DATA

A version of this chapter is a manuscript by Yang Xu, Rafael Kramann , Rachel Patton McCord, and Sikander Hayat. This manuscript is currently under peer-review.

This chapter was slightly revised to be different from the original manuscript. Y.X. and S.H. planned and designed the study. Y.X. performed the computational analysis. Y.X. and S.H. analyzed and interpreted the data and wrote the manuscript. R.P.M. and R.K. edited the manuscript and advised on data interpretation. All authors read and approved the manuscript.

Abstract

Single-cell transcriptomics datasets from the same anatomical sites generated by different research labs are becoming mainstream. However, fast, and computationally inexpensive tools for standardization of cell-type annotation and data integration are still needed to increase research inclusivity. To standardize cell-type annotation and integrate single-cell transcriptomics datasets, we have built a fast, model-free integration method called **MASI** (**M**arker-**A**ssisted **S**tandardization and **I**ntegration). MASI can run integrative annotation on a personal laptop for approximately one million cells, providing a cheap computational alternative for the single-cell data analysis community. We demonstrate that MASI outperforms other methods based on speed, and its performance for the tasks of data integration and cell-type annotation is comparable or even superior to other existing methods. We apply MASI for integrative lineage analysis and show that it preserves the underlying biological signal in datasets tested. Finally, to harness knowledge from single-

cell atlases, we demonstrate three case studies that cover integration across research groups, biological conditions, and surveyed participants, respectively.

Introduction

Single-cell RNA-seq (scRNA-seq) technologies have rapidly evolved over the last decade. Numerous studies have demonstrated the utility of single-cell transcriptomics datasets in improving our understanding of cellular heterogeneity and molecular mechanisms at unprecedented resolution. Over the past years, many single-cell datasets have been made available from different research groups, using multiple single-cell platforms, and covering diverse biological conditions. Global collaborations, for example the Human Cell Atlas project, further make profiling millions of cells possible (Rozenblatt-Rosen et al., 2017). However, this trend of increasing data generation also introduces the challenge of data integration. Deep-learning-based approaches provide many solutions to integrate single-cell datasets (Kimmel and Kelley, 2021; Lopez et al., 2018; Lotfollahi et al., 2021; Xu et al., 2021a). Additionally, their availability to a wider research community is still limited due to the computational cost. Besides the need to reduce computational burden, we also face another challenge of standardizing cell-type annotation for data integration. Different research groups have their own practices for cell-type annotation. The same cellular system profiled by different research groups could have different cell-type annotations. For example, Litviňuková et al. defined 9 major cell types and 27 sub-types, while a similar atlas-level study by Tucker et al. defined 17 cell-types for the cardiovascular system (Litviňuková et al., 2020; Tucker et al., 2020b). Without the standardization of cell-type annotation, it is hard to establish agreement within the science community. This is also a

pressing issue for integrating COVID-19-related single-cell transcriptomics datasets that have been generated by researchers across the globe to understand the SARS-CoV-2 disease mechanism (Chan Zuckerberg Initiative Single-Cell et al., 2020; Chua et al., 2020).

To address these issues in integrative analysis of scRNA-seq data, we propose a fast, model-free method for standardization and integration of cell-type annotation. Our method relies on using putative cell-type markers from reference data to uniformly annotate query datasets, as putative cell-type markers are reliable cell-type indicators and should hold a constant truth across different studies. Because of its simplicity, our method can easily accommodate annotation for millions of cells using limited computational resources. Thus, we vision our tool could reach to a wider range of single-cell researchers who may not have advanced computational resources.

Result

Benchmark impacts of data processing on batch correction

In our previous study, we found that converting the gene expression matrix to cell-type score matrix through a scoring method PlinerScore (Pliner et al., 2019) based on cell-type markers in PanglaoDB (Franzén et al., 2019a) can be used for integrative cell-type annotation (Xu et al., 2021b). The results in our previous study suggested a simple data processing pipeline could address batch effects within scRNA-seq data for integrative analysis. Here, we first wanted to examine how different processing steps may have impacts on revealing cell-type separation and batch-mixing. For this, we selected 4 batch-involved datasets of 4 tissues and tested 10 different processing pipelines for revealing cell-type separation and batch-mixing (Figure 2.1). The #1 pipeline is the most basic data

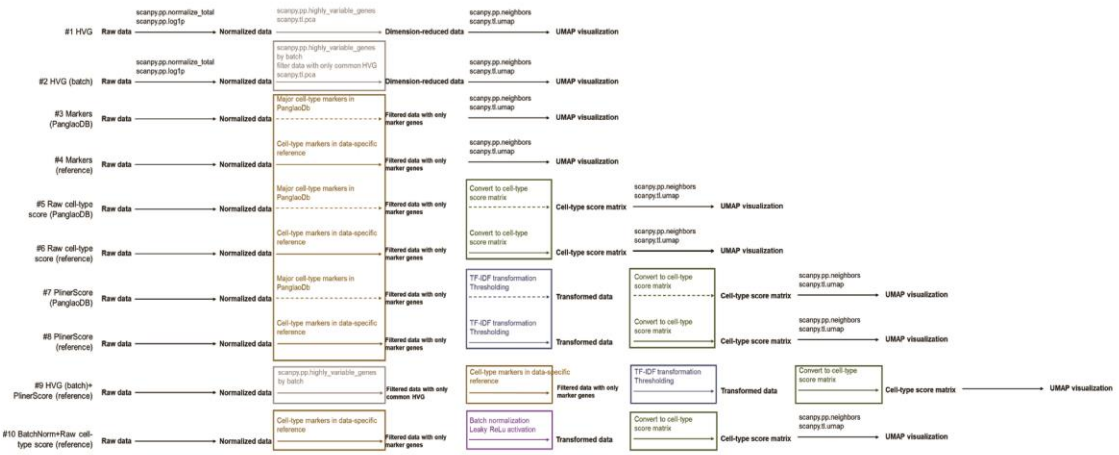


Figure 2.1. Illustration of 10 analysis pipelines for scRNA-seq data. Colored boxes highlight specific practices in a pipeline.

processing for scRNA-seq analysis, which does not take batch information into consideration for calculating the highly variable genes. The #2 pipeline differs from #1 in terms of identifying highly variable genes (HVG) by batch and only including shared HVGs in downstream processing. For #3 and #4 pipelines, we introduced cell-type markers that are obtained from either PanglaoDB or a specific reference data, and we only included those marker genes in the downstream analyses. For #5 and #6 pipelines, we further converted the gene expression matrix to a raw cell-type score matrix containing cell-types in PanglaoDB or in the specific reference data. In pipeline #7 and #8, we added a transformation process proposed by Pliner et al. (Pliner et al., 2019), before converting the gene expression matrix to a cell-type score matrix. #9 pipeline is a combination process of #2 and #8 pipelines. Deep-learning-based batch correction methods demonstrated a considerable success to integrative analysis of scRNA-seq data, and we noticed that the frequent practice across these methods is use of batch normalization layer and non-linear activation layer, which splits the whole dataset into multiple mini-batches, standardizes cells in each batch, and transforms the outcome with a non-linear activation function (Kimmel and Kelley, 2021; Lopez et al., 2018; Lotfollahi et al., 2021; Lotfollahi et al., 2019; Xu et al., 2022a). This batch normalization and non-linear activation process do not require weight training, and we included it into the #10 pipeline.

These 10 pipelines were evaluated in terms of how well these pipelines preserve cell-type structure while mixing batches (Figure 2.2). We defined cell-type silhouette score to quantify how the processing pipeline reveals cell-type structure and batch entropy mixing score to evaluate how well batches are mixed. Based on our benchmark, we observed that

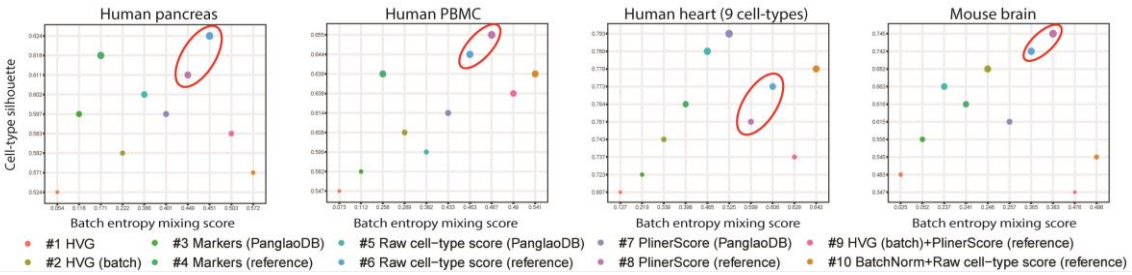


Figure 2.2. Benchmarking of impacts of 10 analysis pipelines on batch correction. Cell-type silhouette score (column) measures how well a pipeline perverse cell-type variation, and batch entropy mixing score (row) quantifies how well a pipeline mixed cells from different batches. Dots located in the top right should present good integration outcomes.

pipelines that use conversion using cell-type markers either obtained from PanglaoDB or from a specific reference data largely mixed different datasets better, and revealed cell-type structure (pipeline #5, #6, #7, and #8), while calling HVG by batch (pipeline #2) and using cell-type markers (pipeline #3 and #4) alone resulted in a lower batch entropy mixing score. We also noticed that the pipelines with PlinerScore (pipeline #7 and #8) had a slight improvement from the raw cell-type score pipelines (pipeline #5 and #6). Both pipeline #9 and #10 have a higher batch entropy mixing score, but a lower cell-type silhouette score. For a good balance of cell-type silhouette and batch entropy mixing, we selected #8 as the processing pipeline for mapping cell-type labels for a query dataset when a reference is available.

Workflow of MASI for integrative analysis

To annotate cell-types for a query dataset based on a fully annotated reference dataset, we propose a new workflow termed MASI. MASI 1) identifies cell-type marker genes from the reference dataset, 2) processes data with the pipeline #8, 3) annotates cell types via MACA (Xu et al., 2021b), and 4) performs other downstream integrative analyses (Figure 2.3A). Briefly, MACA is a marker-based cell-type annotation tool that converts a cell by gene matrix to cell by cell-type matrix, yielding a cell-type label for each identified cell cluster. The first step of the MASI workflow is to identify marker genes for each cell type via differential expression (DE) tests if author-verified markers are not available. To select the DE method that facilitates accurate cell-type annotation through MACA, we benchmarked 12 DE tests, including common DE tests implemented in Scanpy (Wolf et al., 2018) and Seurat (Stuart et al., 2019), and two newly proposed methods COSG (Dai et

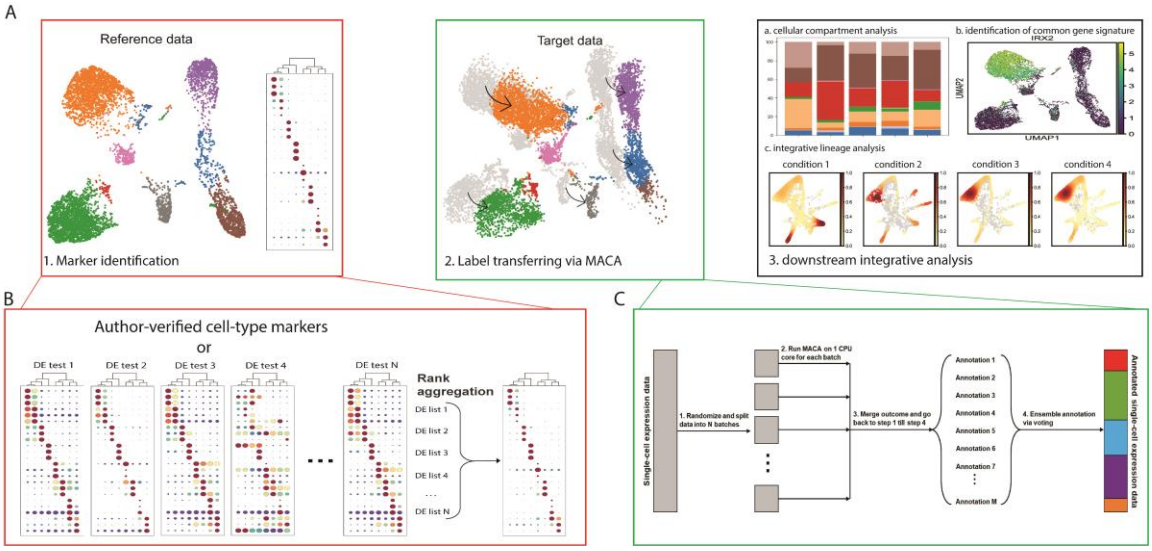


Figure 2.3. Integrative annotation pipeline through MASI. A, A workflow of integrative annotation through MASI, including marker identification from reference data, label transferring by MACA, and downstream integrative analyses. B, Ensemble approach to identify robust cell-type markers from reference data. N DE test outcomes are aggregated to get final ranked marker list. C, Parallel computation for fast annotation to accommodate large-scale scRNA-seq data.

al., 2022) and Cepo (Kim et al., 2021). For the 4 benchmark datasets, we found that marker genes obtained from these 12 DE tests have varying performances in terms of predicting cell types using MACA (Figure 2.4). This is consistent with results shown in other benchmark studies on DE tests (Mou et al., 2020; Soneson and Robinson, 2018; Squair et al., 2021), where none of single DE tests can faithfully identify reliable cell-type markers for all single-cell data. To account for influence by single DE test, we decided to construct ranked cell-type markers via an ensemble approach (Figure 2.3B and see Method in this chapter). Additionally, to accommodate large-scale scRNA-seq data, we reframed the MACA into a parallel manner by splitting data into multiple batches and distributing annotation onto multiple CPU cores (Figure 2.3C). This enables MACA to perform integrative analysis for large-scale scRNA-seq with limited computational resources.

Benchmarking cell-type annotation and data integration

We first benchmarked MASI and other selected methods using the 4 mixed-batch datasets that include a human pancreas data across 5 scRNA-seq platforms (Baron et al., 2016; Grün et al., 2016; Lawlor et al., 2017; Muraro et al., 2016; Segerstolpe et al., 2016), human hematopoietic data across 4 studies (Freytag et al., 2018; Oetjen et al., 2018; Sun et al., 2019), human heart atlas (Litviňuková et al., 2020), and mouse brain data across 4 studies (Rosenberg et al., 2018a; Saunders et al., 2018; Schaum et al., 2018; Zeisel et al., 2015). The human heart atlas data were collected from two institutes and covered single-cell, single-nuclei, and CD45+-enriched data. We selected linear and non-linear support vector machine (SVM) classifiers as supervised methods for benchmarking as a benchmarking study have previously demonstrated that SVM outperformed most sophisticated cell-type

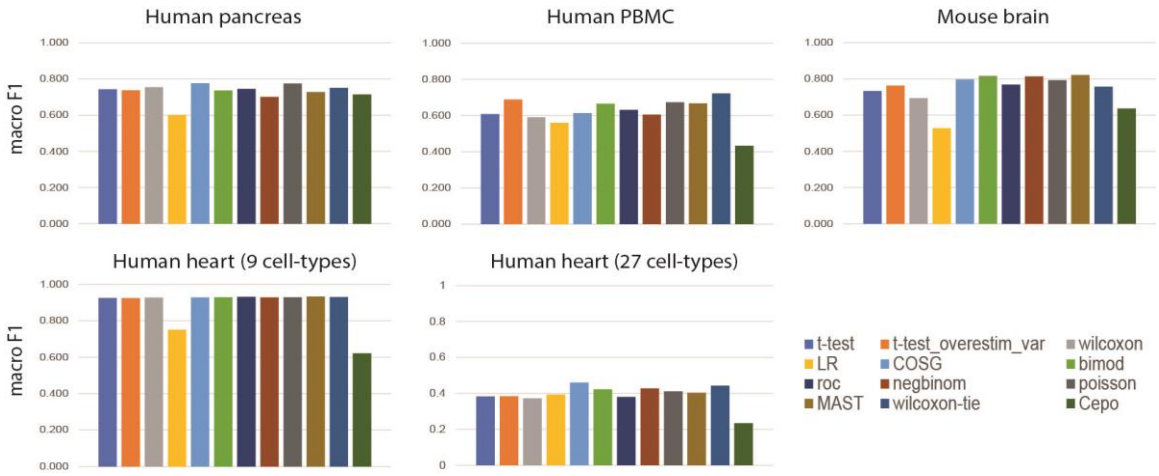


Figure 2.4. Benchmarking of impacts of 12 DE tests on MACA-based cell-type annotation. Cell-type markers are identified from a reference data using a specific DE test. Then, MACA annotates target data with markers identified by this specific DE test. Macro F1 score is reported to show how compatible the DE test is to MACA-based cell-type annotation.

classification methods (Abdelaal et al., 2019). scNym (Kimmel and Kelley, 2021) and scArches (Lotfollahi et al., 2021) are semi-supervised deep learning methods for cell-type annotation and data integration, and we also included these two methods in our benchmark. For a fair comparison, our benchmark study was performed on a local workstation with 64GB memory and Nvidia Quadro RTX 6000 as GPU support. Of note, both scNym and scArches use GPU to speed up computation, while MASI will not use GPU for computing. We first focused on how well mapping cell-type labels from reference data to query data is done by these methods. We used macro F1 and overall accuracy to quantify the performance of these methods in terms of how accurate annotation is for each cell type and how accurate annotation is for the overall dataset. We found that all methods have similar performance in terms of overall accuracy, but MASI has higher macro F1 scores across all datasets in our benchmark (Figure 2.5A). This suggests an advantage of MASI in annotating non-major cell types, considering most single cell data are class imbalanced. Next, we evaluated how well the representations learned by these methods reveal cell-type structures while mixing batches, using cell-type silhouette score and batch entropy mixing score. Here, MASI demonstrated a good balance between capturing cell-type variation and batch mixing (Figure 2.5B).

Dependence on choice of reference dataset

Given the high dependence on reference data for cell-type marker identification, MASI will not be able to annotate cell types in query data that have not been seen in reference data. However, it is still worth answering if a cell-type score matrix constructed using the reference data can preserve cell-type structure for query data, even though query data

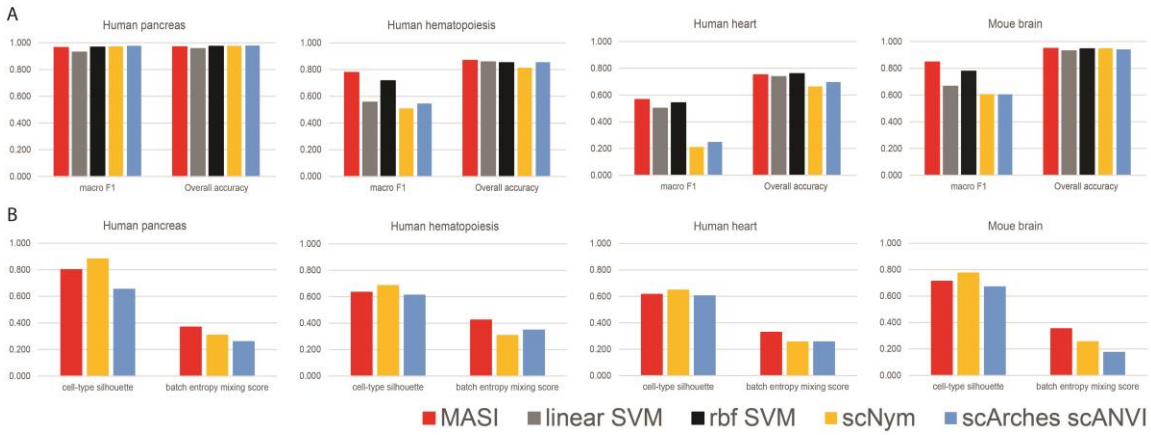


Figure 2.5. Batch correction and label transferring benchmarks. A, Comparison of label transferring for MASI, supervised, and semi-supervised methods. ACC: overall accuracy. Macro F1 is average of F1 scores per cell type. A higher score in both metrics suggests better cell-type prediction. B, Comparison of batch correction for MASI, scNym, and scArches scANVI. Cell-type silhouette score measures how well the integrated representation by these methods preserves cell-type variation, while batch entropy mixing score measures how well the same cell type from different batches is mixed.

contains unseen cell types. To understand the impact of choice of reference dataset on the cell-type annotation in the query dataset, we swapped reference data from Oetjen et al. to 10x Genomics data in the human hematopoietic benchmark dataset. The 10x Genomics data has only 12 cell types, while Oetjen et al. identified16 cell types in their original report (Oetjen et al., 2018). Thus, the query data would contain 4 extra unseen cell types. We performed marker gene identification and transformed the gene expression matrix to cell-type score matrix as above. We observed that the 12-dimension cell-type score matrix built upon the 10x Genomics dataset as reference can reveal cell-type structure for the Oetjen et al. data that have 16 cell types in total. However, as erythrocytes and erythroid progenitor cell-types are not present in the reference, MASI mislabeled them as CD14+ monocytes and HSPCs, respectively (Figure 2.6). We next asked if we could identify subtypes from MASI-reported cell types to match the author-reported annotation resolution. Here, we used SCCAF, a computational method that was previously proposed for the identification of putative cell types through a machine learning approach (Miao et al., 2020). The concept behind this machine learning is: if the clustering resolution reflects the number of true cell types within the data, a machine learning classifier can achieve a high accuracy with the clustering label. We applied SCCAF to identify potential subtypes for each major cell type identified by MASI. We further evaluated how these three approaches, MASI annotation, SCCAF identification, and SCCAF+MASI annotation respectively, revealed a similar annotation resolution by calculating ARI and NMI against the author's annotation. We found that SCCAF+MASI annotation matches the author's annotation resolution more than MASI annotation and SCCAF identification alone (Figure 2.6). To summarize cell-type

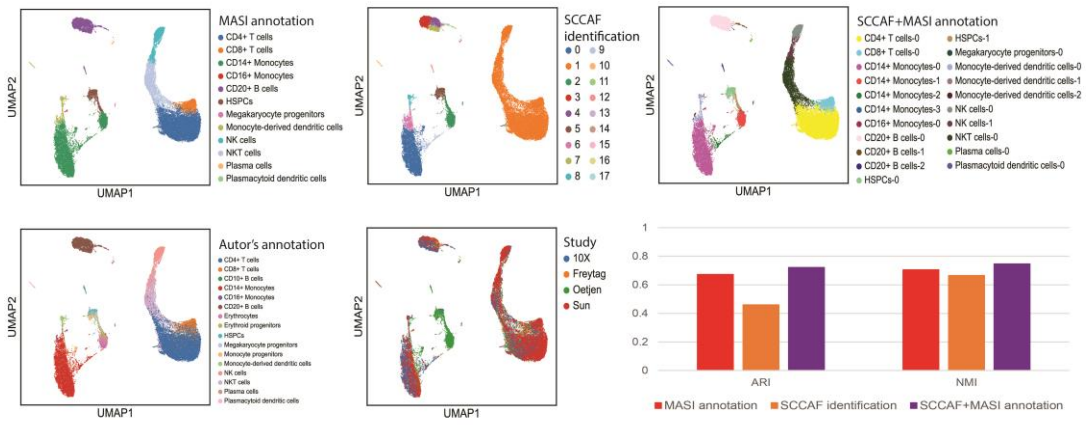


Figure 2.6. Comparison of annotation resolution for MASI, SCCAF, and combination of SCCAF and MASI. 10X data is used as reference for label transferring. Cluster identification through SCCAF is based on 12-dimension cell-type score matrix. SCCAF is applied to MASI-reported annotation to further identify subtypes. ARI and NMI are calculated by comparing method-reported annotation with author-reported annotation.

identification, we conclude that the choice of reference data is critical to the performance of MASI. Moreover, to unravel potential subtypes, users can combine SCCAF and MASI to reach a finer annotation.

Annotation of spatial transcriptomics data with MASI

Considering the capacity of MASI in integration of transcriptome data, we wonder if this also applies to sequencing-based spatial resolved transcriptome data. Thus, we examined this possibility in mapping cell type labels from scRNA-seq data to sequencing-based spatial transcriptomics data. We tested MASI on spatial hippocampus data profiled by Slide-seqV2, since Slide-seqV2 reaches a higher resolution of spatial profiling than 10X Visium (Stickels et al., 2021). Integrating Slide-seqV2 with scRNA-seq further suggests a potential application of MASI in spatial transcriptomic analysis (Figure 2.7A). MASI was able to assign cell type labels to the mouse hippocampus Slide-seqV2 data (Figure 2.7B). Spatial expression patterns of marker genes for 5 distinct cell types also match with their cell locations in space (Figure 2.7C).

Integrative temporal analysis using cell-type score matrix

An advantage of using cell-type scores as features is that it condenses biological information from high-dimension gene feature space into a lower dimension cell-type feature space. Meanwhile, we showed above that converting gene feature space to cell-type feature space is useful for mixing batches coming from different studies. Thus, we could apply our approach to a continuous system and study lineage differentiation. For this, we selected three datasets for integrative lineage analysis: 1) human peripheral blood

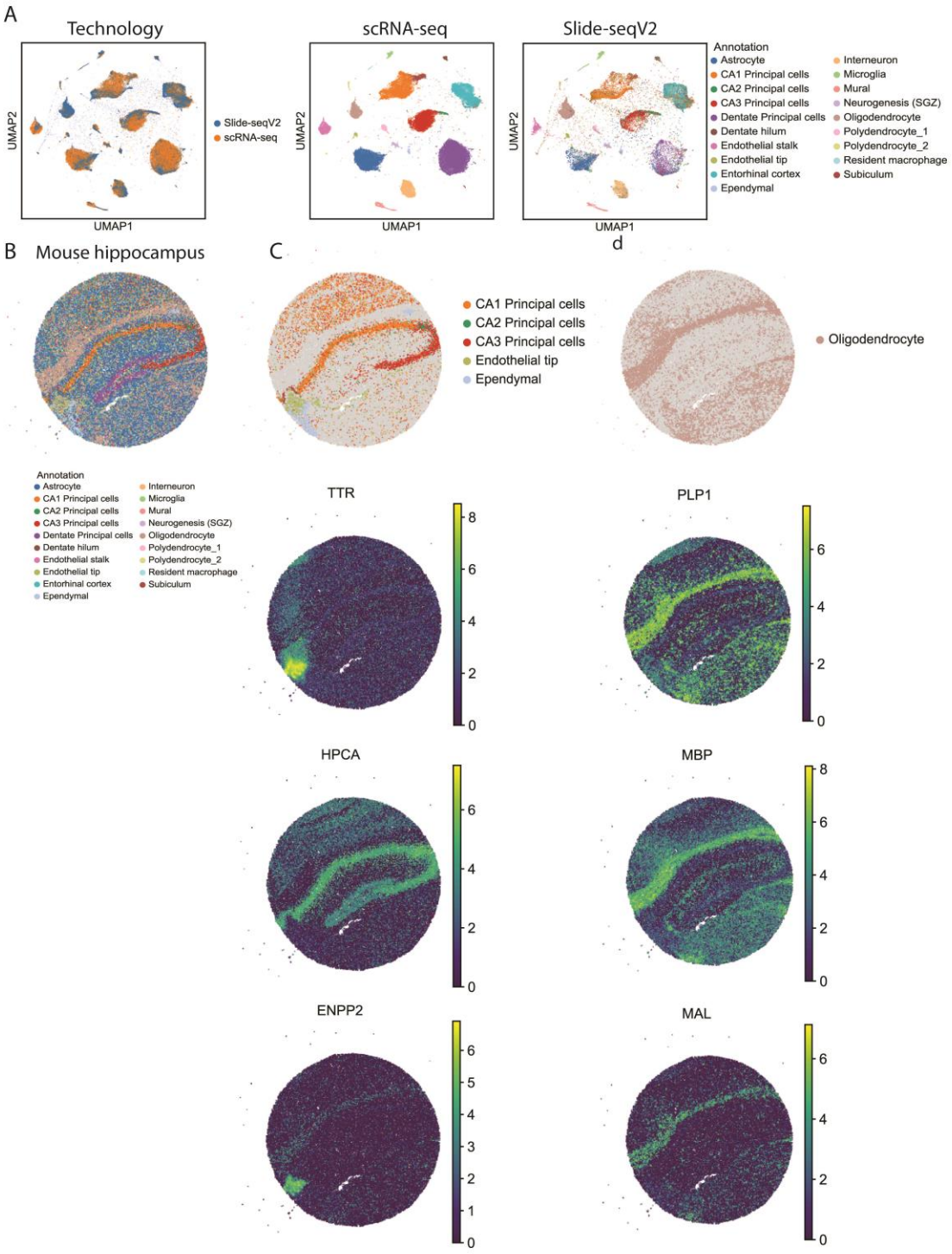


Figure 2.7. Integration of scRNA-seq and Slide-seqV2 by MASI. scRNA-seq is used as reference and Slide-seqV2 data is annotated according to cell type identified in the reference. Markers for principal cells, endothelial tip and oligodendrocyte are selected for visualization, shown below cell-type annotation.

mononuclear cell (PBMC) data of patients with Kawasaki disease obtained before and after IVIG (intravenous immunoglobulin) treatment (Wang et al., 2021b), and 2) zebrafish embryo from two studies that cover 13 major developmental stages (Farrell Jeffrey et al., 2018; Wagner Daniel et al., 2018).

Human hematopoiesis studies were in multi-condition design, and we were able to obtain cell-type markers from an externally annotated 10X Genomics PBMC data. We constructed an integrative lineage map with cell-type score matrices and visualized population density and cell-type score (Figure 2.8A). Using cell-type scores to interpret data, we were able to identify lineage changes, which is consistent with the original report. For example, we observed decreased B1 B-cell and CD16+ monocyte lineages as well as restored plasma cell and CD4+ T native lineages after IVIG treatment for acute Kawasaki disease patients (Figure 2.8A).

Our integrative analysis for developing zebrafish embryos consists of data from two independent data sources that cover different time points of post fertilization. Wanger et al. collected cells from 7 stages including 4, 6, 8, 10, 14, 18, and 24 hpf (hours post fertilization), while Farrell et al. designed 12 finer stages ranging from 3 to 12 hpf (Farrell Jeffrey et al., 2018; Wagner Daniel et al., 2018). We were unable to find an external marker gene reference for the two developing zebrafish datasets. Given they were in a time-series design, we reasoned that the end-point data should contain all mature cell types. Therefore, we intrinsically selected the end-point data that has 30 cell types as reference to identify cell-type markers. Next, we transformed the combined gene expression matrix into a 30 cell-type score matrix and built an integrative lineage map of the developing zebrafish

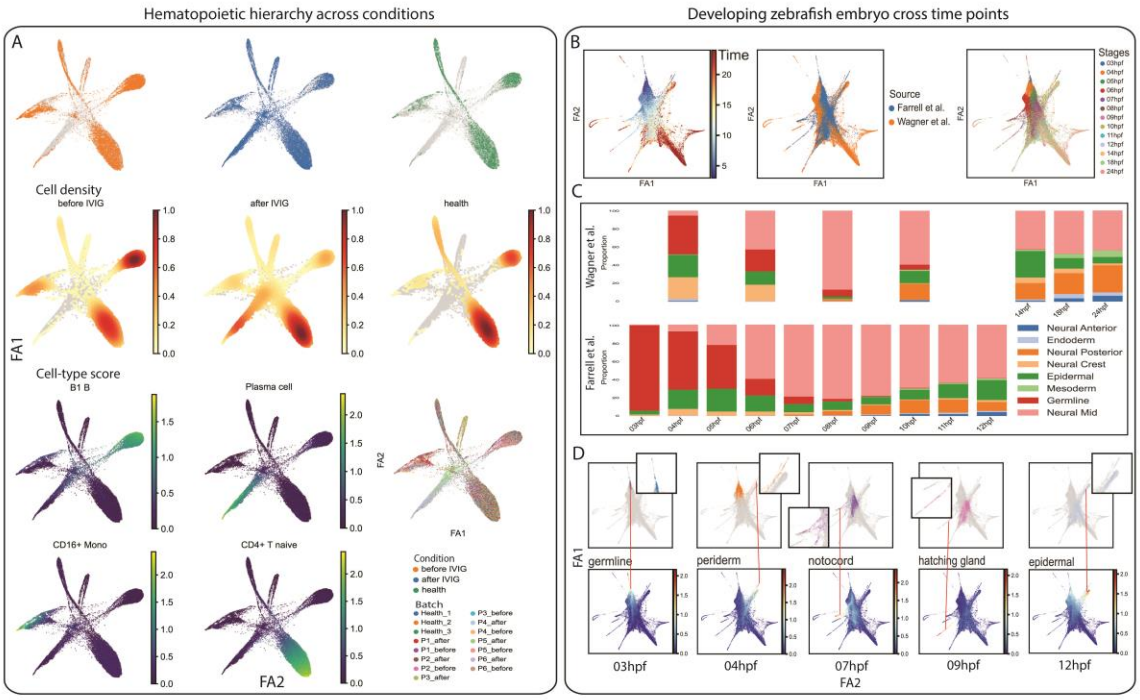


Figure 2.8. Integrative lineage analysis using cell-type score matrix. A, Integrative lineage analysis for multi-condition human hematopoiesis study. Cell density, cell-type score, and batch id for human hematopoiesis samples under different conditions are visualized separately through the first two ForceAtlas2. B, integrative lineage analysis for two developing zebrafish embryo data. Cells are visualized through UMAP and are colored according to developmental time (left), study id (middle), and developmental stages (right). C, Components of 8 major lineages along the developmental stages in zebrafish embryo. All lineages sum up to 1 in one stage, and data from the two studies are visualized separately. D, Identification of lineage origin time. Visual investigation is conducted by matching emergence of a cell-type with the earliest developmental stage in the data.

embryo. Because of the design differences, we manually summarized all developmental stages in 13 major stages (Figure 2.8B). In total, 2 independent studies cover 30 cell types along the 13 developmental stages. Instead of assigning cells to these 30 cell types, we annotated them as 8 major lineage types using MASI. The choice for these 8 major cell-types was based on the lineages observed in 24 hpf. Markers for these 8 major lineage types were also identified from data at the 24 hpf time point. We then visualized how lineage components change along the developmental timeline (Figure 2.8C). First, we found that the two studies are largely consistent. Second, we observed a decline of germline and lineage diversification along these developmental stages (Figure 2.8C). We further investigated the original time point of different cell lineages based on our integrated lineage map and found that the development of germline can be retrieved back at least at the 3 hpf (Figure 2.8D). In Wagner et al., the earliest time point at which germline cells were observed is 4 hpf. However, in Farrell et al., the authors report that the germ layer appears before 4 hpf and that many other lineages do not separate until 4 hpf. This is consistent with our finding from the integrated lineage map, where we show that germline cells are observed at 3 hpf and are the major cell lineage component until that time point (Figure 2.8C and D). The notochord defines the longitudinal axis of the embryo and determines the orientation of the vertebral column, and our analysis suggests the notochord emerges at around 7 hpf, while both Farrell et al. and Wagner et al. showed emergence of the notochord takes place between 6hpf and 8hpf. We also observed that epidermal lineage appears at 3 hpf (Figure 2.8C and D), consistent with Farrell et al. who observed this epidermal lineage at 3.3 hpf. Additionally, we observe that non-neural ectoderm separates

from epidermal cells at 12 hpf in our analysis, as seen in Farrell et al. Taken together, these three analyses for temporal datasets using MASI shows a simple and intuitive approach for integrative lineage analysis.

Case 1: Using human heart atlas for integration of single-cell human heart across studies

Tucker et al. (Tucker et al., 2020b) and Litviňuková et al. (Litviňuková et al., 2020) provide two atlas level resources for human heart data at single-cell resolution. In addition, other human heart datasets are also available (Cui et al., 2019b; Wang et al., 2020c). However, these studies did not use the same cell-type naming style and reported annotation at different resolutions. The human heart atlas identified 27 subtypes while Tucker et al., (17 subtypes), Wang et al. (5 cell types), and Cui et al. (9 cell types) reported different numbers of cell-types in their dataset (Cui et al., 2019b; Litviňuková et al., 2020; Tucker et al., 2020b; Wang et al., 2020c). We think uniform annotation of cell-type labels and batch-mixing of these datasets can yield insights into common themes and inter-human variability across these datasets. Since the human heart atlas data revealed the greatest number of subtypes, we chose human heart atlas data as reference. Because cell-type naming and annotation resolution vary among these studies, we changed to use ARI and NMI for evaluation. We found all methods compared here have similar performance for mapping cell-type labels to Tucker et al., but MASI shows better outcome than the other 4 methods in both Wang et al. and Cui et al. (Figure 2.9A). Moreover, both Wang et al. and Cui et al. have distinct difference in the number of cells profiled and both have a lower annotation resolution than Litviňuková et al. and Tucker et al. Relying on a greater resolution of

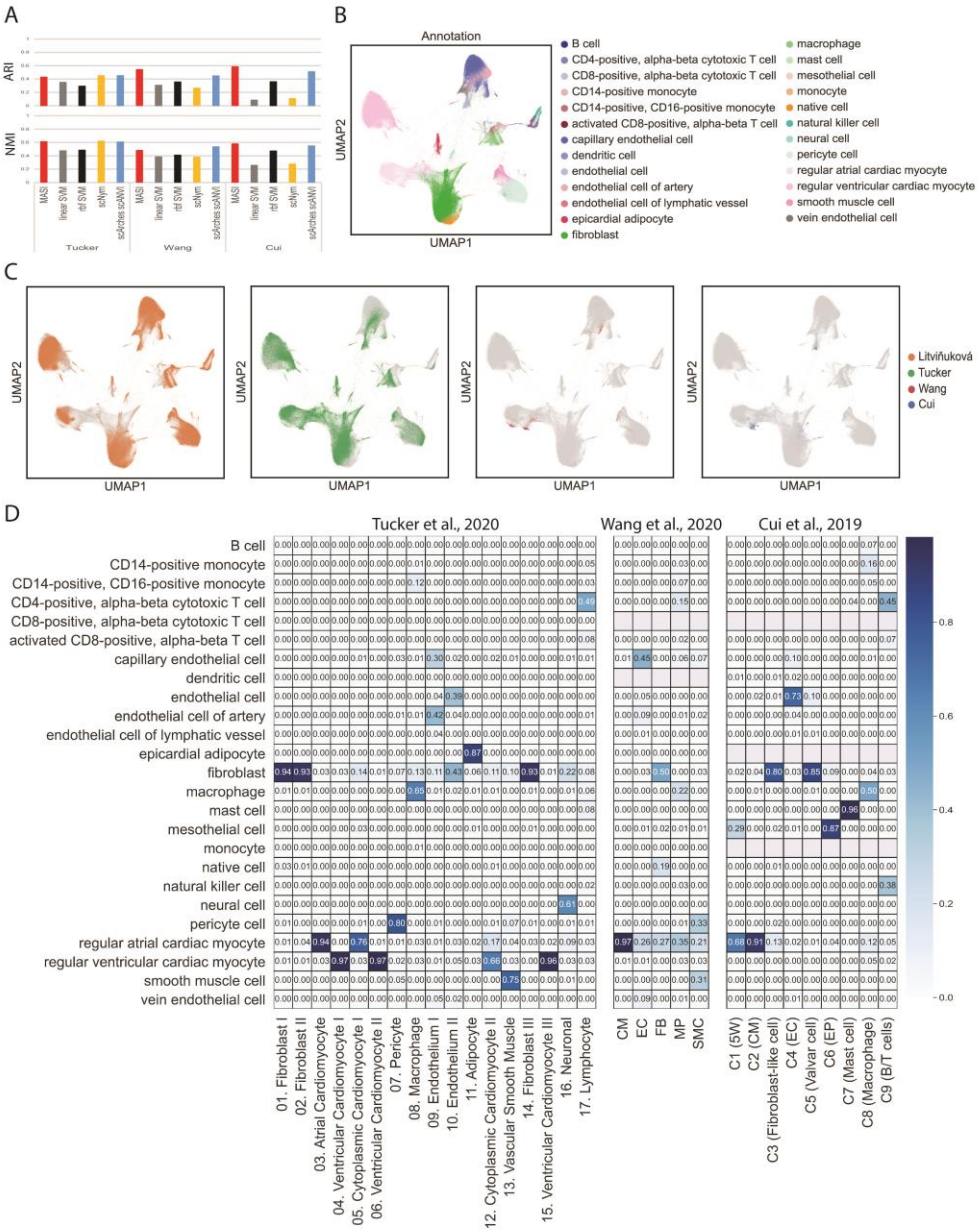


Figure 2.9. Transferring human heart atlas for integration of single-cell human heart across research groups. A, Comparison of label transferring for MASI, supervised, and semi-supervised methods. ARI and NMI are calculated by comparing method-reported annotation with author-reported annotation in a study-wise manner. B, Visualization of the integrative annotation by MASI. Cells are colored according to MASI-reported cell-type annotation. C, Visualization of integration by MASI. Cells are colored according to study id. D, confusion matrix of MASI-reported annotation against author-reported annotation. Confusion matrix is normalized to have column sum as 1. Row names use the naming style of human heart atlas, and column names remain the original naming styles of Tucker et al, Wang et al, and Cui et al data.

Litviňuková et al. data, we were able to level up annotation for the other 3 studies (Figure 2.9B). We visualized integration via MASI and observed no distinct batch differences (Figure 2.9C). We noticed MASI annotated several cells in Tucker et al. as fibroblast while the author-reported annotation for these cells includes cardiomyocyte, endothelium, and neural cells (Figure 2.9D). With MASI, we identified pericyte in Wang et al. and natural killer cell in Cui et al., which were not reported by authors (Figure 2.9D).

Case 2: Using human kidney atlas for integration of single-cell human kidney across multiple conditions

The first human kidney atlas profiled 27 distinct cell types in mature kidney (Stewart Benjamin et al., 2019). This atlas study provides a good reference to understand cellular irregularities in kidney diseases. So far, independent single-cell studies have been conducted to reveal mechanisms in different kidney diseases (Arazi et al., 2019; Kuppe et al., 2021; Wilson et al., 2019; Wu et al., 2018a). An approach that can provide an integrative view for multiple kidney diseases may further add an insight into how cellular irregularities vary among different kidney diseases. We used kidney atlas data as reference and mapped cell-type labels to human kidney data that were collected under different conditions, including CKD (chronic kidney disease) and DKD (diabetic kidney disease). Benchmarking in this task showed MASI has better agreement with author-reported annotations with consistency (NMI values of 0.49, 0.648, and 0.728, respectively) (Figure 2.10A). Overall mapping, cell type standardization and batch-mixing results are shown in Figure 2.10B and C. Next, we focused on the human DKD data, which came with its control set. Population density map suggested decrease of proximal tubule (Figure 2.10D) and

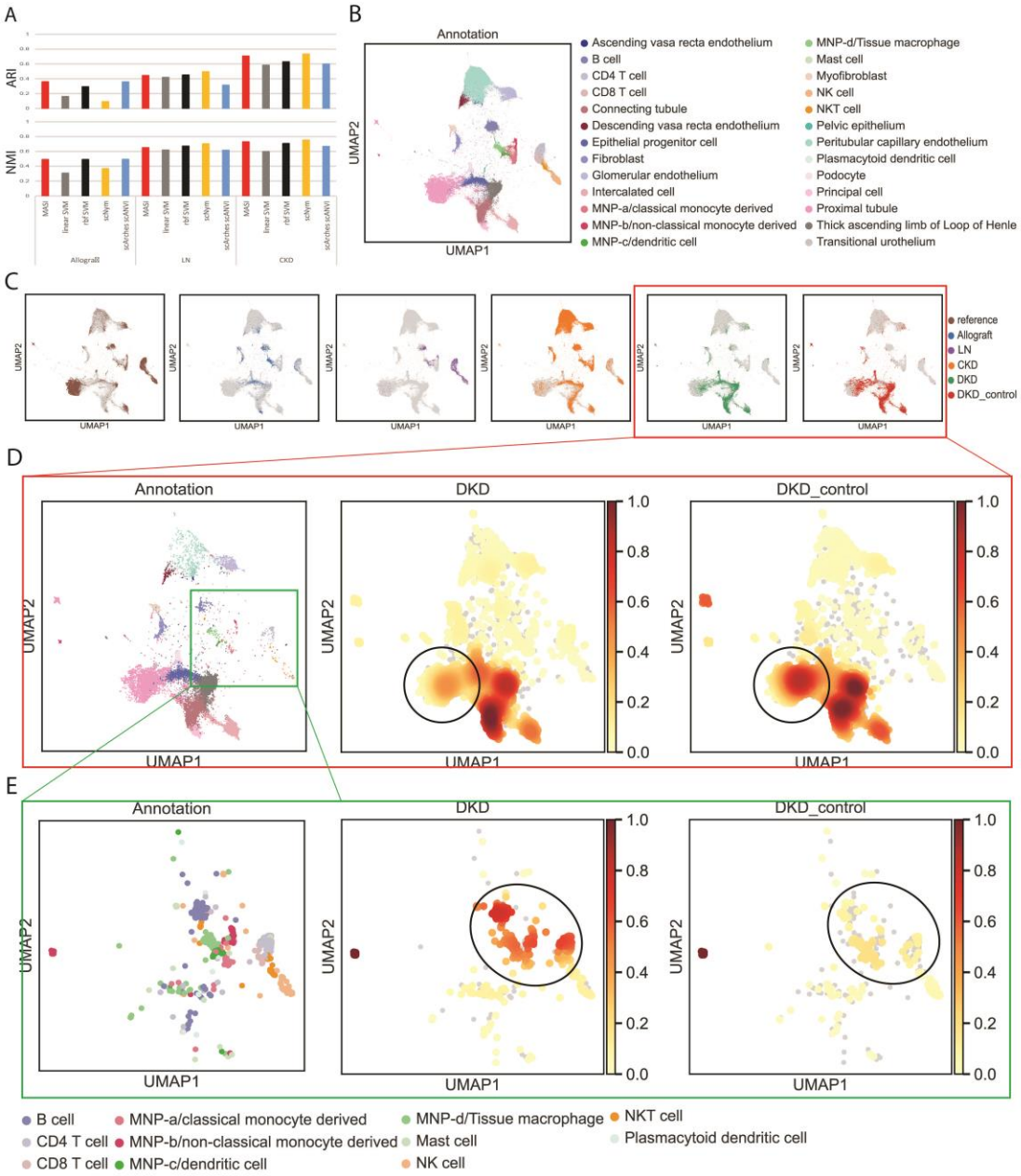


Figure 2.10. Transferring human kidney atlas for integration of single-cell human kidney across conditions. A, Comparison of label transferring for MASI, supervised, and semi-supervised methods. ARI and NMI are calculated by comparing method-reported annotation with author-reported annotation in a study-wise manner. B, Visualization of the integrative annotation by MASI. Cells are colored according to MASI-reported cell-type annotation. C, Visualization of integration by MASI. Cells are colored according to study id. d, Population densities in DKD and control samples. Cell type annotation is shown on the left panel. Cell type population densities of DKD and control samples are presented separately to highlight difference of cell type populations.

increase of immune cells (Figure 2.10E), consistent with an increase of immune response identified in DKD patients.

Case 3: transferring human lung atlas for integration of single-cell COVID19 data across participants

Our third MASI application is transferring knowledge learned from human lung atlas to understand the global COVID19 pandemic at cellular level among healthy and COVID19 participants. The human lung atlas data served as reference data with 59 identified subtypes (Travaglini et al., 2020). Using this annotation, we aimed to annotate 80 COVID19 samples collected from nasal swabs (58 participants and 32818 cells) and airways across different individuals (22 participants and 143168 cells) (Chan Zuckerberg Initiative Single-Cell et al., 2020; Chua et al., 2020). These COVID19 data included both negative (21 participants) and positive samples (59 participants) from multiple research institutes. Due to cell-type annotation and resolution differences, we cannot directly compare cellular differences between healthy and COVID19 participants. We used MASI to annotate the COVID19 data to match the annotation resolution of human lung atlas. Again, we benchmarked MASI with two SVM classifiers, scNym, and scArches, using ARI and NMI as evaluation metrics. We found MASI has greater agreement with author-reported annotations for all COVID19 data than the other 4 methods (Figure 2.11A). Since cell-type annotations for all participants were leveled up to the same resolution, we were able to directly compare the cellular differences between healthy and COVID19 participants. We observed distinct cellular components between healthy and COVID19 groups, and the distinct cellular component is consistent across participants within the same group (Figure 2.11B). Then,

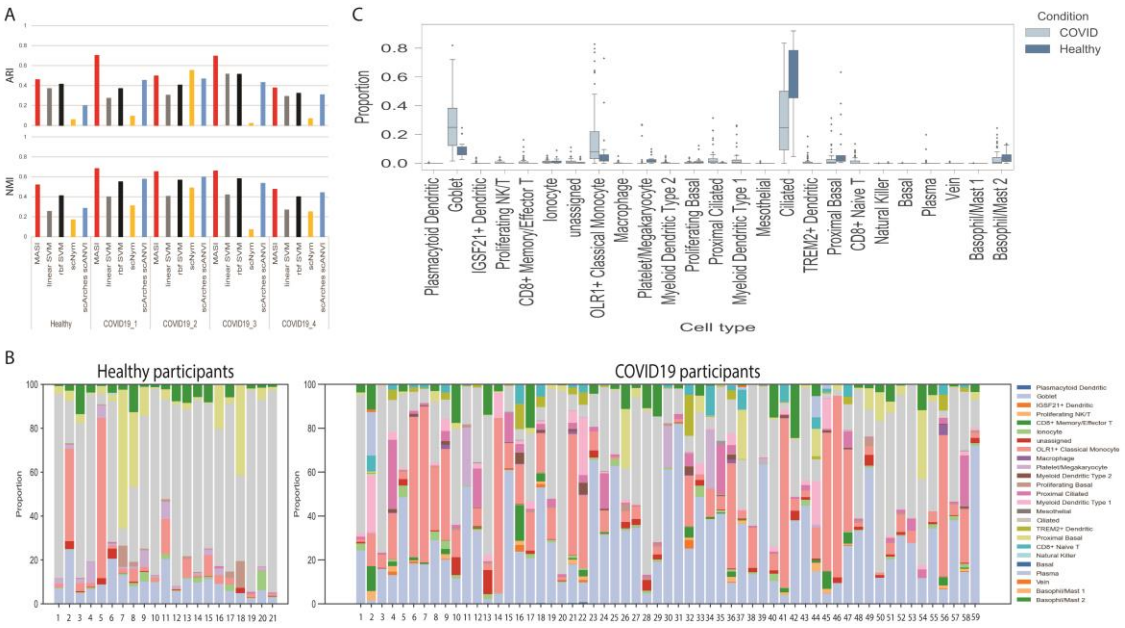


Figure 2.11. Transferring human lung atlas for integration of single-cell COVID19 data across individuals. A, Comparison of label transferring for MASI, supervised, and semi-supervised methods. ARI and NMI are calculated by comparing method-reported annotation with author-reported annotation in a study-wise manner. B, Cellular components of healthy and COVID19 participants. Each column represents one individual. C, Cellular component comparison of healthy and COVID19 participants.

we quantified the changes of cellular composition for all cell types and found an increase in the proportion of Goblet cells and a decrease in ciliated cell proportions in the COVID19 group (Figure 2.11C). This discovery may explain other investigations of SARS-CoV-2 virus targeting ciliated cells via ACE2 (Ahn et al., 2021; Lee et al., 2020).

Discussion

Here, we present MASI, a new tool to quickly and accurately annotate single-cell datasets based on marker genes obtained from a reference dataset. We show that MASI can also be used for batch-mixing and serve as a data integration method for single-cell transcriptomics data. We benchmarked MASI with supervised and semi-supervised methods, and our results show that performance of MASI is comparable or even superior to supervised/semi-supervised methods based on the benchmarking datasets used. We also showed that cell-type scores can be used as features for integrative lineage analysis and demonstrated its intuitive interpretability. Finally, we showed the utility of MASI in three different case studies of data integration covering different research groups, biological conditions, and surveyed participants. Like other supervised and semi-supervised methods replying on reference data, accurate annotation via MASI is also dependent on the quality of reference data. Thus, the choice and resolution of the reference are critical to downstream analysis. If query data has unseen cell types not in reference, MASI in combination with SCCAF can be used to identify subtypes within major cell-types. Additionally, we showed that MASI can also be applied for cell-type prediction in spatial transcriptomics datasets using comparable single-cell transcriptomics datasets as reference.

There are many well-established integration methods available to address batch effects in scRNA-seq datasets, for example Seurat, Harmony, and LIGER (Korsunsky et al., 2019; Liu et al., 2020; Stuart et al., 2019). Additionally, some deep learning-based methods such as HDMC and CarDEC are also available (Lakkis et al., 2021; Wang et al., 2021a). In next chapter, I will also show another deep learning approach to address batch effects. While in this chapter, we rigorously tested cell-type score-based integration via MASI across various single-cell platforms, cytoplasm/nuclei, research groups, conditions, and individuals. Our analyses suggest marker-based feature engineering can be useful for reference-based cell-type annotation, batch-mixing, and data integration. We also demonstrate that integration via MASI preserves biological information for lineage analysis with 3 different examples.

Overall, MASI is easy to set up and requires limited computation resources to run. It can be used for reference-based cell-type annotation and batch-mixing, which could facilitate quick hypothesis-driven exploration of diverse datasets obtained from different labs. Moreover, the democratization of single-cell transcriptomics data (larger cellular output with lower cost) could empower researchers even with limited computational resources to investigate millions of single cells among diverse biological systems.

Methods

Data preprocessing

Raw gene expression count data were ‘LogNormalized’, which divides the total count in that cell and multiplies it by a scale factor of 10 000 (in all our analyses), followed by log-transformation to get the normalized expression matrix. For implementing MASI, we

skipped the step of calling highly variable genes, because only the identified marker genes were used for integrative annotation. For training scNym and scArches, we used top 5000 highly variable genes by batch, which were calculated using function "pp.highly_variable_genes" in Scanpy (Wolf et al., 2018).

Marker rank aggregation

We considered two ensemble marker ranking schemes. In the first scheme, the top 20 marker genes from each DE test were compiled together. For the second scheme, only statistically significant marker genes based on the p-values corrected for multiple hypothesis correction were considered. In the first scheme, we searched the consensus ranking via robust rank aggregation (Kolde et al., 2012). In the second scheme, rank aggregation was done through Lancaster combination (Li et al., 2021).

Weighing markers

When data to be annotated contains distinct cell types and cell types do not share marker genes, we reasoned that weighing markers would not influence the final annotation by MASI. However, this can be beneficial to distinguish cell subtypes that share common markers, for example subtype T cells. We used a simple weighing strategy that returned good label transferring. Given N markers for cell type A, the 1st marker in this ranked list will contribute 100% of its expression to the cell-type score of A, while the Nth marker only contributes 50% of its expression. For the ith marker in the rest, we form this discount calculation $\boldsymbol{A = 1 - \left(\frac{i}{N}\right) * (\frac{1}{2})}$ to get their weights in cell-type A. Beside the weighing

strategy above, other weighing strategies, including Rank Order Centroid and Ratio method, can also be considered for customization.

Converting gene expression matrix to cell-type score matrix

Cell-type score for a given cell-type A with N expressed markers is calculated by summing up the expression of all N markers with consideration of weighing markers as above. This is defined as the raw cell-type score. From this, the PlinerScore is calculated by adding a TF-IDF transformation and suppressing expression values of a marker gene to zeros if they are below the X-percentile of expression values across all cells before the raw cell-type score conversion. The default value for PlinerScore threshold is 0.25 as the percentile threshold (Pliner et al., 2019).

Classification by linear and non-linear SVM

Both linear and non-linear SVM classifiers can be impacted by feature selection. As a benchmark reported, linear and non-linear SVM can have varying prediction accuracies for scRNA-seq data, when different feature selection processes were applied (Ma et al., 2021). Nevertheless, using more discriminative features should improve the accuracy of these two supervised models. Instead of using highly variable genes and PCA-reduced features, we used the same cell-type markers that were used for MASI to train both linear and non-linear SVM classifiers.

Transfer learning through MASI

Once cell-type markers are identified, Mapping cell-type labels to query data is performed using MACA (Xu et al., 2021b). Briefly, for each cell, MACA generates two labels - a per-

cell cell-type Label 1 and group-based clustering Label 2. Then, MACA maps clustering Label 2 to cell-type Label 1 to get the overall cell-type annotation. In the previous, we used different clustering parameters to generate multiple Label 2s, for the purpose of reproducibility. In this study, we also ran Louvain community detection with a range of clustering parameters to get multiple clustering Label 2s. These include clustering resolution 3, 5, 7 with 5, 10, 15 as neighborhood sizes to over-cluster cells. With multiple clustering Label 2s, we were able to map them to Label 1 and get a more reproducible ensembled cell-type annotation. To accommodate for large-scale scRNA-seq data, we split the whole data into N batches and ran MACA with one batch per CPU core.

Transfer learning using scNym and scArches

Both scNym and scArches are deep-learning-based transfer learning methods. Therefore, an optimal outcome for a specific data might require customized parameter tuning. However, for benchmarking, we used default pipelines of both methods for all data involved in this study. Respective tutorials can be found at their host GitHub.

2D visualization using UMAP

To visualize integrations by these three methods, we used the same parameter setting for all datasets. We set up metrics “cosine” to define distance, cells within 0.1 were considered as neighbors, and minimum 15 cells form a community.

Integrative lineage analysis

We used ForceAtlas2 with PAGA (partition-based graph abstraction) initialization to layout integrative lineage maps with cell-type scores instead of any other hidden space

features, like PCA (Principal Component Analysis) representation or representation from neural network model (Jacomy et al., 2014; Wolf et al., 2019). To initialize PAGA, we performed Louvain community detection to assign cells as multiple meta cells. We used resolution 5 for Louvain community detection to get enough meta cells. Once cells are laid out on the ForceAtlas2 space, we directly visualize lineage paths with cell-type scores, without clustering cells into cell types.

Evaluation metrics

Overall accuracy: Acc=(Total number of correction predictions)/(Total number of cells).

Macro F1: $\boldsymbol{F1} = \frac{\mathbf{precision} * \mathbf{recall}}{(\mathbf{precision\ recall})} * \mathbf{2}$. F1 was calculated for each cell type, then macro F1 is the average of F1 scores for all cell types. Because this metric doesn't consider class weights for imbalanced data, a higher macro F1 could suggest correction predictions for both dominant and non-dominant cell types.

Cell-type silhouette score: We first used function "sklearn.metrics.silhouette_score" in scikit-learn Python package to calculate a typical silhouette score. The author-reported cell type label served as the ground truth. This calculation uses the hidden space returned by integration methods with cell-type label. Both scNym and scArches learned a 10-dimension hidden space representation by default. The lower representation by MASI depends on the number of unique cell type labeled available in the reference dataset. Next, we rescaled the score from 0 to 1 by (1+S)/2 to defined as cell-type silhouette score. The higher the score is, the better cell-type variation is captured.

Batch entropy mixing score(Haghverdi et al., 2018): $\boldsymbol{E} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} \boldsymbol{x_i log}(\boldsymbol{x_i})$. In this study, $\boldsymbol{x_i}$ is the proportion of cells from batch $\boldsymbol{i}$ in a region of the first two UMAPs, and $\sum_{\boldsymbol{i=1}}^{\boldsymbol{c}} \boldsymbol{x_i} = \mathbf{1}$.

This score should quantify how well mixed cells from different batches are in a region. The same as Cell-type silhouette score, calculation of Batch entropy mixing score is based on the hidden space returned by integration methods with batch information as label. The higher the score is, the better mixing.

Adjusted rand index (ARI): The rand index (RI) measures a similarity or agreement between two clustering labels. The ARI then is defined through ARI=(RI-expected RI)/(max(RI)-expected RI). In this study, we used ARI to measure the agreement between cell-type annotation reported by a transfer learning method and the author-reported cell-type annotation.

Normalized mutual information (NMI): Like ARI, NMI also qualifies the agreement between two clustering labels. It is defined as $\boldsymbol{NMI} = \frac{\boldsymbol{I}(\boldsymbol{P}, \boldsymbol{T})}{\sqrt{\boldsymbol{H}(\boldsymbol{P})\boldsymbol{H}(\boldsymbol{T})}}$. $\boldsymbol{P}$ and $\boldsymbol{T}$ are empirical categorical distributions for the predicted and real clustering, $\boldsymbol{I}$ is the mutual entropy, and $\boldsymbol{H}$ is the Shannon entropy.

CHAPTER III
MUTUAL INFORMATION LEARNING FOR INTEGRATION OF SINGLE-CELL OMICS DATA

A version of this chapter is a manuscript by Yang Xu, Priyojit Das, and Rachel Patton McCord. This manuscript was published in *Bioinformatics*.

This chapter was revised to be different from the original manuscript. Y.X. conceived and developed the method with guidance from R.P.M. and produced all the figures. P.D. computationally processed raw sequencing data for input to SMILE. Y.X. and R.P.M. wrote the manuscript with input from P.D.

Abstract

Deep learning approaches have empowered single-cell omics data analysis in many ways and generated new insights from complex cellular systems. As there is an increasing need for single cell omics data to be integrated across sources, types, and features of data, the challenges of integrating single-cell omics data are rising. Here, we present an unsupervised deep learning algorithm that learns discriminative representations for single-cell data via maximizing mutual information, SMILE (**S**ingle-cell **M**utual **I**nformation **Le**arning). Using a unique cell-pairing design, SMILE successfully integrates multi-source single-cell transcriptome data, removing batch effects and projecting similar cell types, even from different tissues, into the shared space. SMILE can also integrate data from two or more modalities, such as joint profiling technologies using single-cell ATAC-seq, RNA-seq, DNA methylation, Hi-C, and ChIP data. When paired cells are known, SMILE can integrate data with unmatched feature, such as genes for RNA-seq and genome wide peaks for ATAC-seq. Integrated representations learned from joint profiling technologies can then be used as a framework for comparing independent single source data.

Introduction

Deep-learning-based single-cell analysis has gained great attention in recent years and has been used in a range of tasks, including accurate cell-type annotation (Ma and Pellegrini, 2020), expression imputation (Arisdakessian et al., 2019), and doublet identification (Bernstein et al., 2020). In these tasks, deep learning showed some striking advantages. For example, in cell-type annotation, the automatic and accurate annotation using a deep classification model saves researchers from manual cell-type annotation (Kimmel and Kelley, 2021; Lopez et al., 2018; Ma and Pellegrini, 2020). Another application of deep learning is data imputation and denoising. Though there has been a dramatic improvement in scRNA-seq technology, the problem of measurement sparsity remains as a grand challenge in single-cell transcriptome data (Lähnemann et al., 2020). Due to the difficulty of modeling technical zero values and biological zeros, deep learning has become a more appealing alternative for this task. An autoencoder (AE) is a common artificial neural network that is used to learn representations for data in an unsupervised manner. Both DeepImpute (Arisdakessian et al., 2019) and DCA (Eraslan et al., 2019) adopt a variant of AE model to impute gene expression and de-noise scRNA-seq data. These approaches and many others are revealing the power of deep learning applied to single-cell omics datasets.

Data integration is a rising challenge in single-cell analysis, as increasing numbers of single-cell omics datasets become available, and the types of omics data become more diverse. Consequently, data integration becomes a key research domain for understanding a complex cellular system from different angles (Argelaguet et al., 2021; Forcato et al., 2021; Longo et al., 2021; Stuart et al., 2019). In single-cell transcriptomes, batch effects

are usually a prominent variation when comparing data from multiple sources and removing batch effects is a critical step for revealing biologically relevant variation. Besides integrating single-cell transcriptome data, integration of multimodal single-cell data is even becoming more important as technological breakthroughs make it possible to capture multiple data types from the same single cell. For example, sci-CAR and SHARE-seq can simultaneously profile chromatin accessibility and gene expression for thousands of single cells (Cao et al., 2018; Ma et al., 2020b). scMethyl-HiC and sn-m3C-seq can profile DNA methylation and 3D chromatin structure at the same time at single-cell resolution (Lee et al., 2019; Li et al., 2019). However, these data types do not naturally share the same feature space: transcriptomes are described using genes as features, while chromatin accessibility is reported across all intergenic spaces. Therefore, integration of multimodal data is more challenging because of this feature discrepancy. Furthermore, a new technology named Paired-Tag now achieves joint profiling of gene expression and 5 different histone marks for thousands of single nuclei (Zhu et al., 2021). This new technology brings the further challenge of integrating more than 2 modalities.

Currently, there are 3 major approaches of integration for single-cell data integration: horizontal, vertical, and diagonal approaches, respectively. Argelaguet et al. outlined published methods in each category (Argelaguet et al., 2021). Horizontal approaches rely on feature anchors for integration. Methods in this category can address batch effects within multi-source single-cell transcriptome data or multimodal single-cell data integration, if shared features exist. Since independent single-cell assays over these years have generated most single-cell omics data, integrative analysis by horizontal approaches

is critical for a full use of these independent studies. On the other hand, vertical approaches will need cell anchors to learn the integration and have their unique niche when shared features do not exist across different data type, or matching features is counterintuitive. For example, in joint Methyl/Hi-C data, Hi-C data quantitates 2D interaction features across the genome while methyl measures the methylation level of genomic regions in 1D (Lee et al., 2019; Li et al., 2019). Either matching the 1D methylation features to 2D interaction features or vice versus is not practical. Meanwhile, horizontal approaches may not necessarily learn a shared space even though different data types can have shared features. In such case, vertical approaches are good alternatives for data integration. As for diagonal approaches, computational studies are showing a greater challenge, and there has not been a method that demonstrates a general use for most single-cell data integration. Though we expect to have an ideal diagonal method to solve most integration problems, horizontal and vertical methods are still the mainstream in data integration so far.

To address challenges above in a single method, we designed a deep learning model, SMILE, that learns a discriminative representation for data integration in an unsupervised manner. In our approach, we restructured cells into pairs, and we aimed to maximize the similarity between the paired cells in the latent space. Because of this cell-pairing design, SMILE extends naturally into integration of multimodal single-cell data, where data from two sources (RNA-seq/ATAC-seq or Methyl/Hi-C) exist for each cell and thus form a natural pair. We demonstrated that SMILE can effectively project RNA-seq/ATAC-seq data, as well as Methyl/Hi-C data, into the shared space and achieve data integration. We demonstrate how our representation allows us to identify genes and regions of accessibility

that are critical for the mutual definition of distinctive cell types by these different data types. We also show how an integrated representation created using jointly profiled data can then be used to project and interpret single source data. Finally, we present a combinatorial use of SMILE models to integrate single-cell RNA-seq, H3K4me1, H3K9me3, H3K27me3, and H3K27ac data generated by Paired-Tag. In summary, SMILE performs as well or better than other methods designed for data integration while also having increased flexibility in terms of data input types.

Method

Architecture of SMILE, p(paired)SMILE and mp(modified paired)SMILE

We proposed three different variants of SMILE models to serve different uses of single-cell data integration (Figure 3.1). All three variants of SMILE have encoders as the main components for feature extraction. In SMILE, there is only one encoder. This encoder consists of three fully connected layers that have 1000, 512 and 128 nodes, respectively. Each fully connected layer is coupled with a BatchNorm layer to normalize output which is further activated by ReLu function. Different from SMILE, pSMILE and mpSMILE have another encoder that has the same structure as the one in SMILE but takes an input with different dimension. The use of two independent encoders (Encoder A and B) in pSMILE and mpSMILE is to handle inputs from two modalities with different features, for example RNA-seq and ATAC-seq. Therefore, pSMILE and mpSMILE do not require extra feature engineering to match features for inputs from two different data sources. Modified from pSMILE, mpSMILE has a duplicated Encoder A, which shares the same weights. Using duplicated encoders take advantage of the discriminative power of RNA-seq or

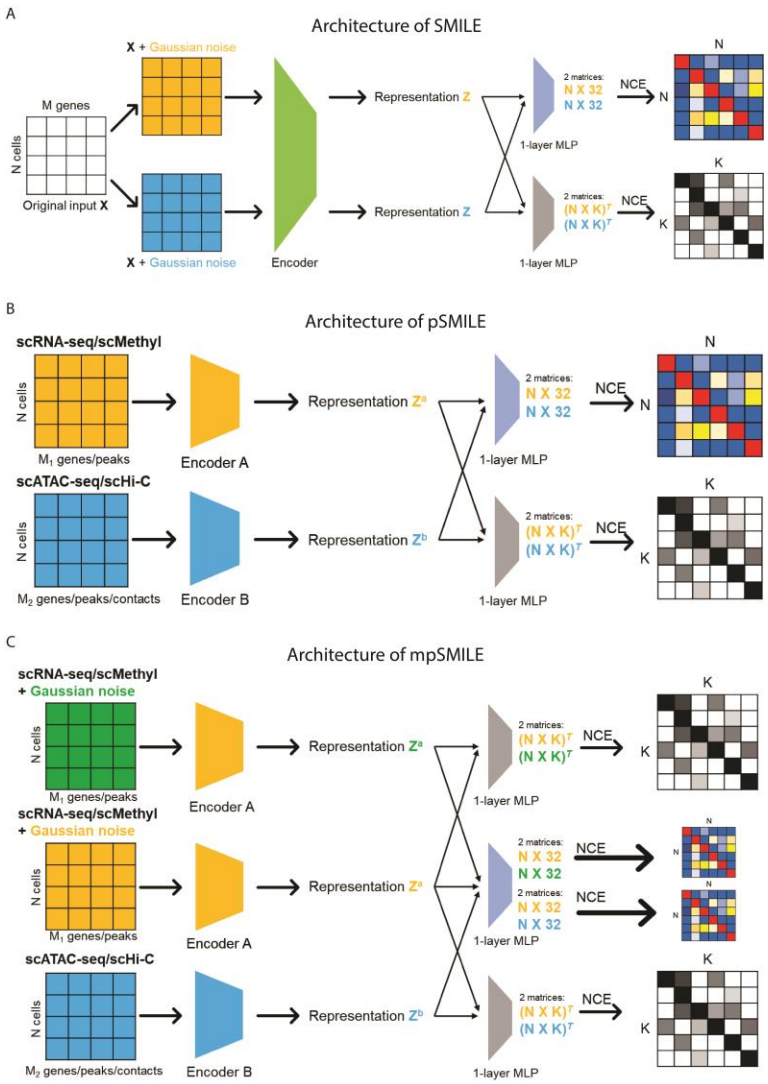


Figure 3.1. Architectures of 3 SMILE variants. A, Architecture of SMILE. Original input *X* represents a gene expression matrix where each row represents a cell, and each column stands for a gene. Random Gaussian noise is added to differentiate the input *X* into two *X*s, which are the same except for the added noise. Encoder (green) projects *X* into 128-dimension representation *z*. Two independent fully connected layers (blue and grey) are stacked onto the encoder to further reduce *z* into 32-dimension output and *K* pseudo cell-type probabilities, respectively. B, Architecture of pSMILE. scRNA-seq/scMethyl would be forwarded through Encoder A to produce representation z^a, and scATAC-seq/scHi-C would be forwarded through Encoder B to produce representation z^b. Two one-layer MLPs in pSMILE are the same as those in SMILE. C, Architecture of mpSMILE. scRNA-seq or scMethyl data are forwarded through Encoder A to produce representation z^a, and scATAC-seq or scHi-C are forwarded through Encoder B to produce representation z^b. mpSMILE has duplicated Encoder As, and cells in scRNA-seq/scMethyl part would be duplicated by adding gaussian noises and become self-pairs.

Methyl to learn a more discriminative representation. This is because we observed a compromise between the modality with more discriminative power and the other with less, and several other integration methods also show that giving more weight to RNA-seq data is critical for learning discriminative representation (Jain et al., 2021; Lin et al., 2021; Peng et al., 2021; Stuart et al., 2019). We stack two independent fully connected layers to the encoder(s), which further reduce the output from encoder(s) to a 32-dimension vector with ReLu activation and K pseudo cell-type probabilities with SoftMax activation. In this study, we set K to 25 for all datasets, based on the observation that most single-cell data would contain no more than 25 major cell-types unless it is an atlas-scale data. Meanwhile, though SMILE is fully unsupervised, it can be easily turned into a semi-supervised method with no extra modification of the model architecture, because of the incorporation of K pseudo cell-type probabilities. When cell-type labels are available for a proportion of data, K can be set as the number of known cell-types and the user can add a classification loss to reframe SMILE into semi-supervised learning.

Next, we provide a detailed explanation about the architecture of SMILE (Figure 3.1A). The main component is a multi-layer perceptron (MLP) used as an encoder that projects cells from the original feature space X to representation z. To achieve this goal, SMILE relies on maximizing mutual information between X and z. Mutual information measures the dependency of z on X. If we maximize the dependency, we can end up using low-dimension z to represent the high-dimension X. Contrastive learning is one approach to maximize mutual information, and it usually requires pairing one sample with a positive or negative sample for representation learning (Amid and Warmuth, 2019). Then, the goal is

to maximize similarity between the positive pair and dissimilarity between the negative pair in the representation z. Due to the lack of labels, pairing samples is a challenging task. However, treating a sample itself as its positive sample and any other cells as negative samples can be a shortcut for reframing the data into pairs, and Chen et al. demonstrated that such a simple framework can effectively learn visual representation for images (Chen et al., 2020a). In single-cell transcriptome data, we adopt the same framework to pair each cell to itself. To prevent each pair from being completely the same, we add gaussian noises to expression values of each cell. Then, we maximize mutual information by forcing each cell to be like its noise-added pair and to be distinct from all other cells. To implement this in the neural network, we used noise-contrastive estimation (NCE) as the core loss function to guide the neural network to learn (See **Loss function** in this chapter) (Wu et al., 2018b). We did not directly apply NCE on representation z, but further reduced z to a 32-dimension output and K pseudo cell-type probabilities, by stacking two independent one-layer MLPs onto the encoder. A one-layer MLP generating a 32-dimension vector will produce rectified linear unit (ReLU) activated output, and the other will produce probabilities of pseudo cell-types with SoftMax activation. Finally, NCE was applied on the 32-dimension output and pseudo probabilities, independently. These two one-layer MLPs produce two independent outputs which both contribute to the representation learning of the encoder (Li et al., 2020). Once trained, the encoder serves as a feature extractor that projects data from the original space X to a low-dimension representation z.

To apply SMILE to joint profiling data, we modified it into new architectures, pSMILE (Figure 3.1B) and mpSMILE (Figure 3.1C). These new architectures contain two separate

encoders (Encoder A and Encoder B). Encoder A projects RNA-seq or methylation data into representation z^a, while Encoder B handles projection of ATAC-seq or Hi-C into representation z^b. We aim to learn z^a and z^b that will be confined in the shared latent space, so we also apply the same one-layer MLPs to each, which further reduce them into 32-dimension output and probabilities of K pseudo cell types. This is the same as we did in the basic SMILE model. Using the RNA/ATAC- or Methyl/Hi-C-joint data to train p/mpSMILE will be the same as using the self-paired data, except that we introduced two separate encoders to handle inputs from two modalities. Differentiating from pSMILE, mpSMILE has two duplicated Encoder As that share the same weights. In mpSMILE, besides pairing cells from RNA-sea/Methyl and their corresponding cells from ATAC-sea/Hi-C, we will also do self-pairing for cells in RNA-seq/Methyl data as we did in SMILE, by introducing gaussian noise. As for the difference between pSMILE and mpSMILE, we provided experimental outcomes in **Results** in this chapter.

Cell pairing

SMILE takes paired cells as inputs. When using SMILE for integration of multi-source single-cell transcriptome data, we treat each cell itself as positive pair. To prevent the two cells in each pair from being completely the same, we add gaussian noise to differentiate them. Other noise-addition approaches should be applicable here. For example, randomly masking some expression values has been shown to an alternative way to learn discriminative representation for single-cell RNA-seq data (Ciortan and Defrance, 2021). Here, we choose gaussian noise, and the learning process will maximize the true similarities between the pair while minimizing the effect of gaussian noise. When using

pSMILE and mpSMILE for integration of multimodal single-cell data, we pair a cell from RNA-seq/Methyl with its counterpart from ATAC-seq/Hi-C. Since joint profiling quantifies two aspects of one single cell, we know that one cell in RNA-seq/Methyl has a corresponding cell in ATAC-seq/Hi-C. When RNA-seq and ATAC-seq come from two separate studies, the user may need to pair cells of the same cell type manually. Here, we suggest using "FindTransferAnchors" function in Seurat to generate cell pairs. Then, user can use these paired cells to train p/mpSMILE.

Loss function

Noise-contrastive estimation (NCE). The core concept of making cells resemble themselves resides in the use of NCE as the main loss function. In training, we divide a whole dataset into multiple batches, and each batch has N cells. For multi-source single-cell transcriptome data, we differentiate each cell into two by adding random gaussian noise. Therefore, there are 2N cells in one batch. In each batch, each cell has itself as positive sample and the rest of 2(N-1) cells as its negative samples. For joint profiling data and in an N-pair batch, one of N cells in RNA-seq/Methyl has its corresponding cell among N in ATAC-seq/Hi-C as the positive sample and the rest of 2(N-1) cells summed from both RNA-seq/Methyl and ATAC-seq/Hi-C as negative samples. Let $sim\,(u,v) = u^T v/\|u\|\|v\|$ denote the dot product between L_2 normalized u and v. Then, NCE for a positive pair of examples (i,j) can be defined as (*Eq. 1*).

$Eq.\,1: loss_{i,j} = -\log \frac{\exp\,(sim(z_i,z_j)/\tau)}{\sum_{k=1\,(k\neq i)}^{2N} \exp\,(sim(z_i,z_k)/\tau)}$, where τ denotes a temperature parameter.

Mean squared error (MSE). We use MSE as additional loss function in p/mpSMILE to push the representation of ATAC-seq/Hi-C to be closer to the representation of RNA-seq/Methyl in the latent space (*Eq. 2*). Of note, MSE alone is unable to drive the model to learn a discriminative representation.

$Eq.2: loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(z^a - z^b)$, where z^a and z^b are representations through Encoder A and B.

Data integration through SMILE/pSMILE/mpSMILE

We used "StandardScaler" in sklearn to scale all input data before training SMILE, except that the dimension-reduced Hi-C data has been scaled. We trained SMILE with batch size as 512 for all multi-source single-cell transcriptome data in this study, and the SMILE model can converge within 5 epochs in all cases, indicated by the total loss. In benchmark of integration of 4 joint profiling RNA-seq and ATAC-seq data, we use all cell pairs for training, and we trained p/mpSMILE for 20 epochs with batch size as 512. For sn-m3C-seq data, we trained p/mpSMILE on whole data for 10 epochs with batch size as 512. For all experiments in this study, we used learning rate as 0.01 with 0.0005 weight decay. There are also 3 key parameters in all three SMILE variants, temperature τ_N for the 32-dimension output, temperature τ_K for the K pseudo cell-type probabilities, and variance of gaussian noise. In default setting, we fixed τ_N as 0.1, τ_K as 1, and gaussian variance as 1 for all integration by SMILE. For multimodal single-cell data integration by p/mpSMILE, we fixed gaussian variance as 1 across different datasets. Differently, we fixed τ_N as 0.05 and τ_K as 0.15. A study published after our initial preprint provides detailed support that a fundamentally similar approach can learn discriminative representation for single-cell data

(Ciortan and Defrance, 2021). However, they did not extend their method to data integration, which is our focus in this study.

Evaluation of data integration

To evaluate batch-effect correction, we use ARI and silhouette score as the evaluation metrics. For ARI, we perform Leiden clustering to re-cluster cells using the batch-removed representation. Then, we compare the new clustering label with the cell-type label reported by the authors. For fair comparison with LIGER, Harmony, and Seurat, we use multiple resolutions to get the clustering label that has the highest ARI against author-reported cell-type label, and report that ARI value as the performance of LIGER, Harmony, and Seurat in that data. For silhouette score, we defined batch silhouette and cell-type silhouette. Batch silhouette measures how well different batches align. We use batch information as labels to calculate a typical silhouette score *S* and then report its absolute value *abs(S)* as batch silhouette. Cell-type silhouette indicates how disguisable representation of cells from one cell-type are from other cell-types. Here, we use author-reported cell-type labels to calculate the typical silhouette score S and then transform it through *(1+S)/2*. Therefore, both batch and cell-type silhouette scores will range from 0 to 1. Batch silhouette scores closer to 0 indicate good batch correction, while cell-type silhouette scores closer to 1 show good cell-type separation.

To evaluate integration of multimodal single-cell data, we also use the same silhouette scores as above, but we renamed batch silhouette as modal silhouette because we use modality information for calculation. Again, modal silhouette closer to 0 represents better mixing of RNA-seq and ATAC-seq data, while cell-type silhouette closer to 1 suggests

cells are separated by their cell-type no matter which modality they belong to. Meanwhile, we know each cell pair because of joint profiling. So, we also measure Euclidean distance of paired cells in the 2D UMAP space, before and after training, to show if paired cells become closer in the integrated representation.

Evaluation of label transferring

Once we have an integrated representation for multiple datasets, we can transfer labels from a known data to other unknown data. To evaluate label transferring, we use macro F1 score. In integration of multi-source single-cell transcriptome data, we select one source as the training set to train a Support Vector Machine (SVM) classifier, and test the accuracy measured through macro F1 score in other sources. In joint profiling data, we reason a good integration should allow mutual label transferring, either from RNA-seq/Methyl to ATAC-seq/Hi-C or back from ATAC-seq/Hi-C to RNA-seq/Methyl. So, we report a macro F1 score for both transferring directions. We use the author-reported cell-type label as ground-truth to train a SVM classifier on RNA-seq or Methyl and test it in ATAC-seq or Hi-C or vice versa.

Processing of RNA-seq, ATAC-seq, Methyl, Hi-C, and histone marker data

For RNA-seq data, raw gene expression count data was normalized through a 'LogNormalize' method, which normalizes the raw count for each cell by its total count, then multiplies this by a scale factor (10,000 in our analysis), and log-transforms the result. Then, we use "highly_variable_genes" function in Scanpy to find most variable genes as input for SMILE (Wolf et al., 2018). For ATAC-seq data at peak level, we perform TF-

IDF transformation and select top 90-percentile peaks as the input for SMILE. For ATAC-seq data at gene level, we first use "CreateGeneActivityMatrix" function in Seurat to sum up all peaks that fall within a gene body and its 2,000bp upstream, and we use this new quantification to represent gene activity (Stuart et al., 2019). Then, we apply LogNormalize to gene activity matrix and find most variable genes as input for SMILE. For CG methylation data, we first calculate CG methylation level for all non-overlapping autosomal 100 kb bins across entire human genome. Then, we apply LogNormalize to the binned CG methylation data. For Hi-C data, we use scHiCluster with default setting to generate an imputed Hi-C matrix at 1MB resolution for each cell (Zhou et al., 2019). Due to the size of Hi-C matrix, we are unable to concatenate all chromosomes to get a genome-wide Hi-C matrix. Therefore, we use a dimension-reduced Hi-C data of whole genome, which is implemented in scHiCluster. For histone mark data, we perform TF-IDF to transform the raw peak data and select top 95-percentile peaks as the input for SMILE.

Downstream analysis

We performed wilcoxon test to identify key differential genes/peaks and their ranking in mouse skin RNA-seq, ATAC-seq gene activity, and ATAC-seq peak data, using author-reported cell-type label. We only selected top 15 genes in RNA-seq, top 150 genes in ATAC-seq gene activity, and top 3000 peaks in ATAC-seq peak of each cluster as key differential genes. For testing which features contribute to the representation learned by SMILE, key genes/peaks for each cluster were sequentially assigned values of zero and then the dataset was fed back through the encoder to determine the effect on the representation. We also suppressed activity value of each gene to zero and forwarded the

data with one gene suppressed through the encoder. Then, we measure how much disruption one gene has on the integrated representation, and we selected the top 5% genes that have the most disruption in one particular cell type. Finally, we forwarded data, in which these top 5% screened genes were suppressed to 0s, through the encoder again to measure the collective disruption on the integrated representation. To screen candidate peaks, we first assigned all peaks into 50 topics via negative matrix factorization, and used the same approach to identify which topic contributes most to a particular cell type. Our screening approach is also conceptually similar to the motif screening used to probe a deep learning representation in study by Fudenberg et al. (Fudenberg et al., 2020).

Results

SMILE accommodates many single cell data types

Before we demonstrate applications of SMILE in data integration, we first show that SMILE can handle most types of single-cell omics data separately. We tested SMILE on RNA-seq data from human pancreas, ATAC-seq data from Mouse ATAC Atlas, and Hi-C data from mouse embryo cells(Baron et al., 2016; Collombet et al., 2020; Cusanovich et al., 2018). SMILE can effectively learn a discriminative representation for single-cell human pancreas RNA-seq (Figure 3.2A). Meanwhile, SMILE distinguishes tissue types within the Mouse ATAC Atlas, and it also recovers major cell types in the brain tissue (Figure 3.2A). In the task of clustering single-cell Hi-C data, SMILE has a slight advantage of distinguishing different cell stages compared to PCA (the baseline) (Figure 3.2B). However, we want to point out that, for a single source data, SMILE does not show substantial difference from a standard PCA approach. Since PCA finds most variations and

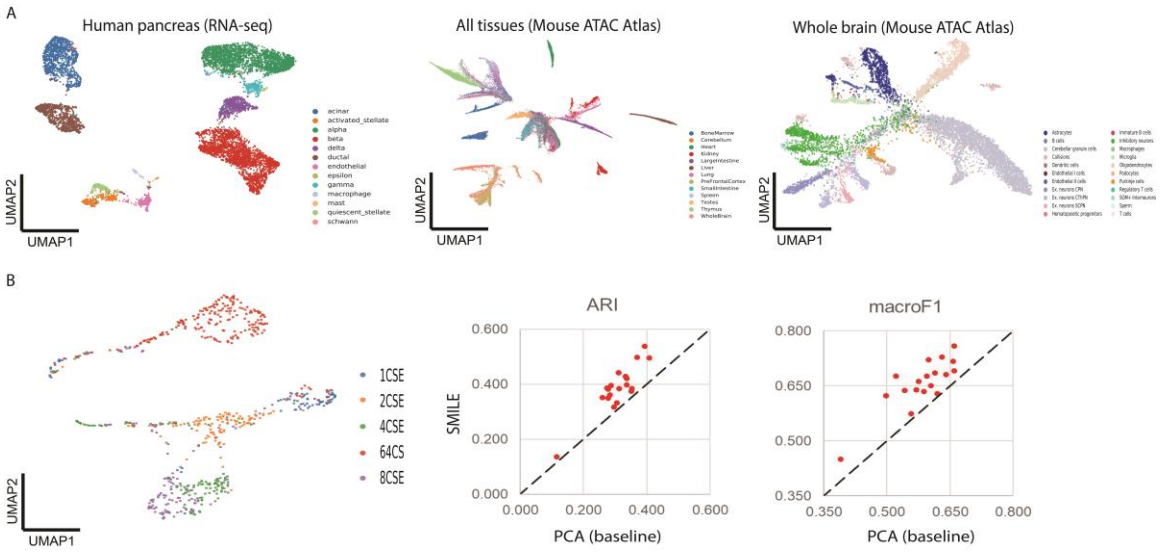


Figure 3.2. Application of SMILE in single-source scRNA-seq, scATAC-seq and scHi-C. A, UMAP visualization of SMILE representation of Human pancreas and Mouse ATAC Atlas. Left panel is the visualization of single-cell human pancreas RNA-seq data. Middle panel is the visualization of whole dataset and cells are colored by tissue types. Right panel is the visualization of a subset of cells from mouse brain. Cells are colored by cell-types reported by the author. B, UMAP visualization of SMILE representation of mouse embryo single cell Hi-C data (left), and comparison of SMILE and PCA (baseline) for each different chromosome separately (right). Cells are colored by developmental stages. Calculation of ARI and macro F1 is based on the developmental stage of a cell as the ground truth.

it is unlikely that unwanted variations are confounded within single-source data, there would be no obvious advantage of using SMILE to find biological variations. Instead, we turn to data integration as the primary application of SMILE is for single-cell data integration.

SMILE eliminates batch effects in single-cell transcriptome data from multiple sources

It is now common to find multiple single-cell transcriptomics datasets for the same tissue or biological system generated by different techniques or research groups. A standard clustering analysis often fails to identify cell types, but instead only detects differences between experimental batches. In contrast, SMILE directly learns a representation that is not confounded by batch effect and can be combined with common clustering methods for cell type identification. We tested batch-effect correction in human pancreas data, human peripheral blood mononuclear cell (PBMC) data, and human heart data (Baron et al., 2016; Grün et al., 2016; Lawlor et al., 2017; Litviňuková et al., 2020; Muraro et al., 2016; Segerstolpe et al., 2016; Tucker et al., 2020b; Zheng et al., 2017). To benchmark the performance of SMILE in removing batch effects, we compared SMILE with LIGER, Harmony, and Seurat. These 3 methods have been reported as the 3 top methods in a benchmarking study of batch-effect correction (Butler et al., 2018; Korsunsky et al., 2019; Liu et al., 2020; Tran et al., 2020). We found that SMILE has comparable performance to Harmony across these 3 systems in terms of batch and cell-type silhouette scores (Figure 3.3A). Meanwhile, the integrated representations learned by SMILE and Harmony are friendly for de novo clustering and label transferring via classification, as measured

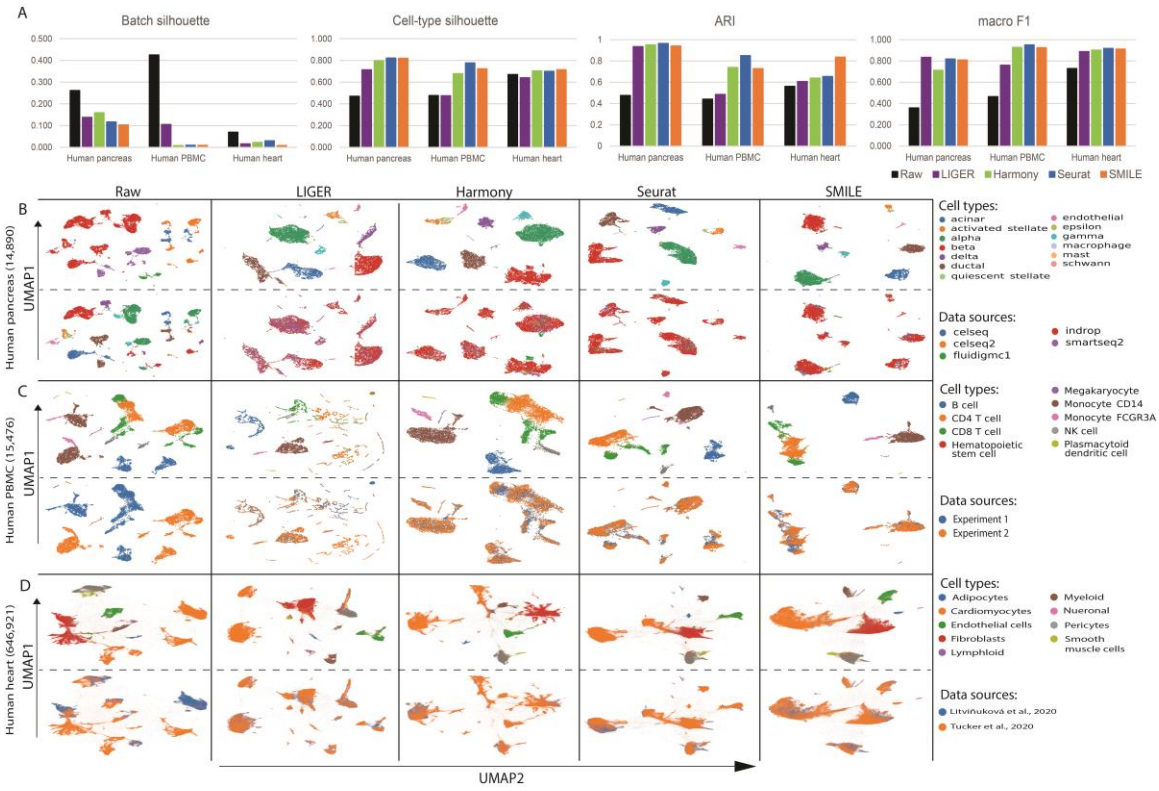


Figure 3.3. Integration of multi-source single-cell transcriptome data using SMILE. A, Evaluation of batch-effects correction. Batch and cell-type silhouette scores measure how well different batches are mixed while distinct cell-types are separated apart. Batch silhouette closer to 0 indicates a good mixing of different batches, while cell-type silhouette closer to 1 represents that distinct cell-types are separated well in the integrated representation. ARI shows how well the learned representation can recover cell-types. ARI closer to 1 indicates that the clustering labels better match original cell-type labels in that study. label transferring is measured through macro F1 score. SVM classifiers are trained with single source data, and then macro F1s are calculated by assigning cell types to the rest of the data sources using that classifier. B-D, UMAP visualization of integrated representations of B) human pancreas data, C) human PBMC data, and D) human heart data, using raw data, or representations learned by LIGER, Harmony, Seurat, and SMILE. Cells in upper rows are colored according to their sources or batch ID, and cells in lower rows are colored by putative cell-types reported in original studies.

through ARI and macro F1 scores (Figure 3.3A). Both methods removed batch effects and recovered cell types identified in original reports (Figure 3.3B-D). In our benchmark study, LIGER is ranked as the 4th place in all 4 metrics. We also noticed LIGER returns the worst representation in human PBMC data (Figure 3.3C). On the other hand, Seurat ranks as the best method overall. However, a substantial disadvantage of Seurat is its inefficient computation design for large datasets.

Joint clustering through mpSMILE improves upon previous methods and reveals key biological variables

Moving from integration of multi-source single-cell transcriptome data, we next tested the performance of pSMILE on a simulated joint single-cell transcriptome dataset and two joint profiling datasets generated by SNARE-seq and sci-CAR to demonstrate the applicability of SMILE in multimodal single-cell data integration (Cao et al., 2018; Chen et al., 2019). The results with simulated joint data, produced by splitting a single scRNA-seq dataset into two subsets with separate genes indicates that pSMILE can integrate data from two entirely different feature spaces (Figure 3.4). It is often observed that RNA-seq data has a greater cell type discriminative power than ATAC-seq, and other integration methods give more weights to RNA-seq in representation learning (Jain et al., 2021; Lin et al., 2021; Peng et al., 2021; Stuart et al., 2019). Therefore, we introduced mpSMILE to take advantage of the discriminative power of RNA-seq.

In this part, we benchmarked methods that fall into all 3 integration categories (horizontal, vertical, and diagonal). We selected UnionCom to represent the diagonal approach (Cao et al., 2020b), LIGER and Harmony for horizontal (Korsunsky et al., 2019;

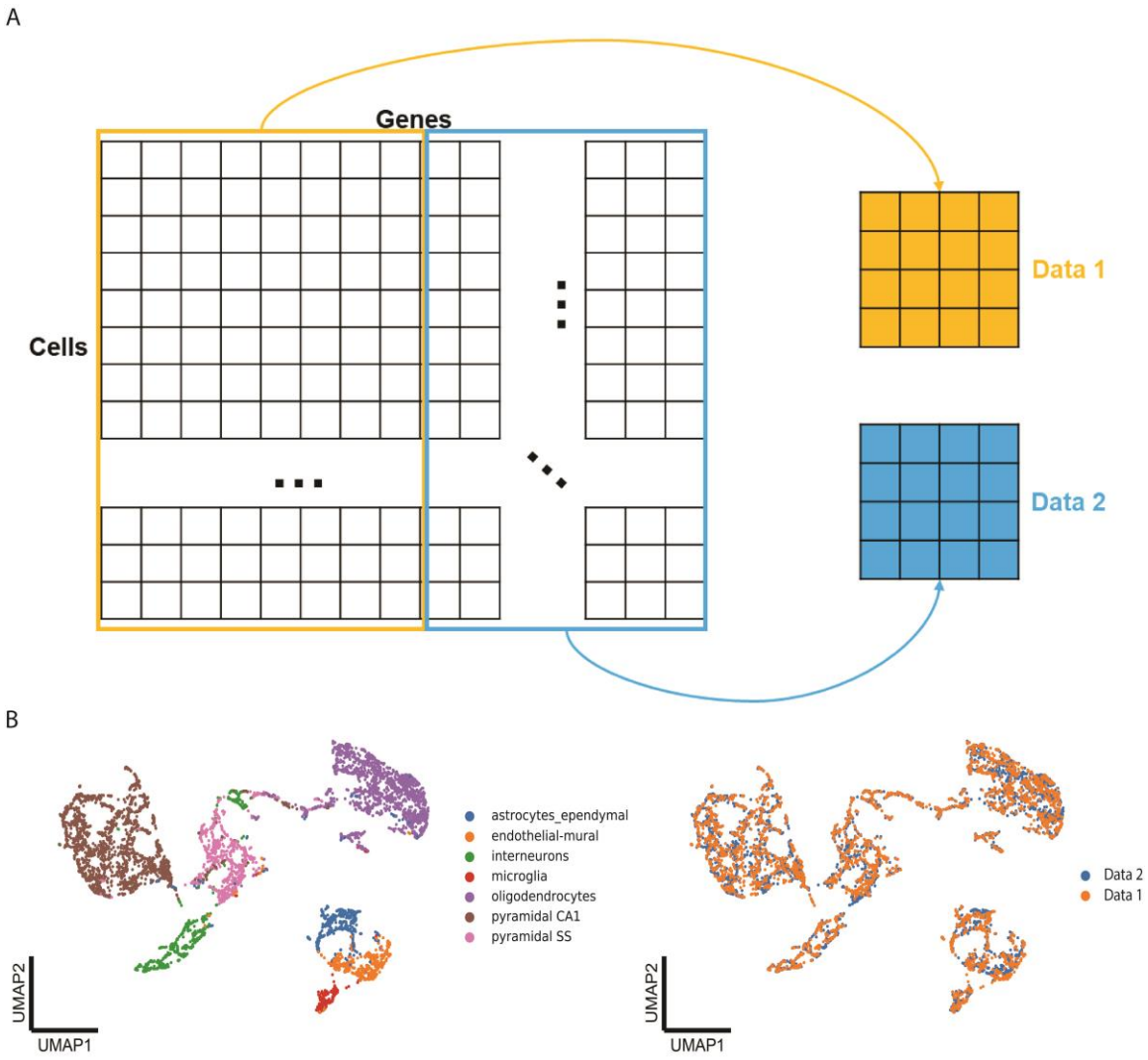


Figure 3.4. Integration of synthetic multimodal single-cell data through pSMILE. A, Construction of synthetic multimodal single-cell data. The synthetic multimodal data is based on a real single-cell RNA-seq data from mouse cortex. Data 1 and data 2 have the same cells, and each cell in data 1 is paired with its corresponding cell in data 2. Data 1 and data 2 are generated from the original data through splitting genes into two halves. Therefore, data1 and data 2 do not share any common features. B, UMAP visualization of integrated representation of mouse cortex by pSMILE. Cells are colored by cell-types (left) and data types (right).

Liu et al., 2020), and MCIA and Seurat for vertical (Meng et al., 2014; Stuart et al., 2019). Our mpSMILE stands with MCIA and Seurat in the category of vertical approach. Though Seurat falls into the category of vertical approach, there is one difference of Seurat from MCIA and SMILE. When projecting joint profiling data into the same space, SMILE accomplishes a similar purpose as Seurat, but with more flexibility of input. Seurat implements canonical correlation analysis (CCA) to project both RNA-seq and ATAC-seq data into the same low-dimension space (Stuart et al., 2019). The use of CCA requires the two datasets to share the same features. As shown above, SMILE can't work with datasets where the two data types involve entirely different features (e.g., genes vs. genomic bins). To make the data work for all methods compare SMILE with Seurat_v3, we re-quantified the ATAC-seq into gene activities, and we further included mouse brain and mouse skin datasets generated by SHARE-seq (Ma et al., 2020b). Since cell pairs are known in these datasets, we used all pairs to train all 3 vertical methods both Seurat_v3 and mpSMILE. We ran all methods with default settings, and we visualized integration results by these methods through UMAP with the same settings. We found that all vertical methods outperform the two horizontal approaches, LIGER, and Harmony. This suggests that there is a more prominent unknown discrepancy between two modalities, even though two modalities have been processed to have shared features. In our hands, UnionCom, the diagonal approach, failed integration tasks for all 4 datasets. Currently, there does not appear to exist a truly successful diagonal method to integrate complex multimodal single-cell data without knowing either feature or cell anchors. All three vertical approaches, MCIA, Seurat, and mpSMILE were able to project ATAC-seq and RNA-seq data into the

shared space while discriminating cell types for the mixed cell-lines data, but MCIA shows less power of cell-type discrimination than the other two (Figure 3.5A). Of note, Seurat showed poor performance on the mouse kidney data by sci-CAR and the mouse brain data by SHARE-seq, failing to project the two data sources into the shared space and distinguish cell types (Figure 3.5A an C). For the mouse skin data by SHARE-seq, MCIA, Seurat and mpSMILE have comparable performance (Figure 3.5D). In summary, we conclude that horizontal methods may not necessarily address modality discrepancy, even though the feature discrepancy is solved via feature engineering. This emphasizes the benefit of joint profiling data to create an integrated space through vertical methods onto which other single source data can be projected for comparison and cell type annotation. Compared to Seurat, MCIA and mpSMILE show a higher performance with mutual label transferring. Surprisingly, Seurat has good label transferring from RNA-seq to ATAC-seq, but this quality of label transfer is not reversible, as shown with lower macro F1 scores from ATAC-seq to RNA-seq. Comparing MCIA and mpSMILE, we observed that SMILE has better multimodal integration in terms of modal and cell-type silhouette scores (Figure 3.5A). In terms of using pSMILE or mpSMILE, we would always recommend mpSMILE for discriminative representation learning. However, if user highlights equal contribution from both modalities, pSMILE should an alternative.

To evaluate which biological factors drive the co-embedding we observe, we set candidate genes from the mouse skin to zero and passed this altered data through the mpSMILE encoder to evaluate whether the co-embedding would be disrupted. Indeed, when we remove key differential genes, clusters are greatly disrupted in the co-embedding

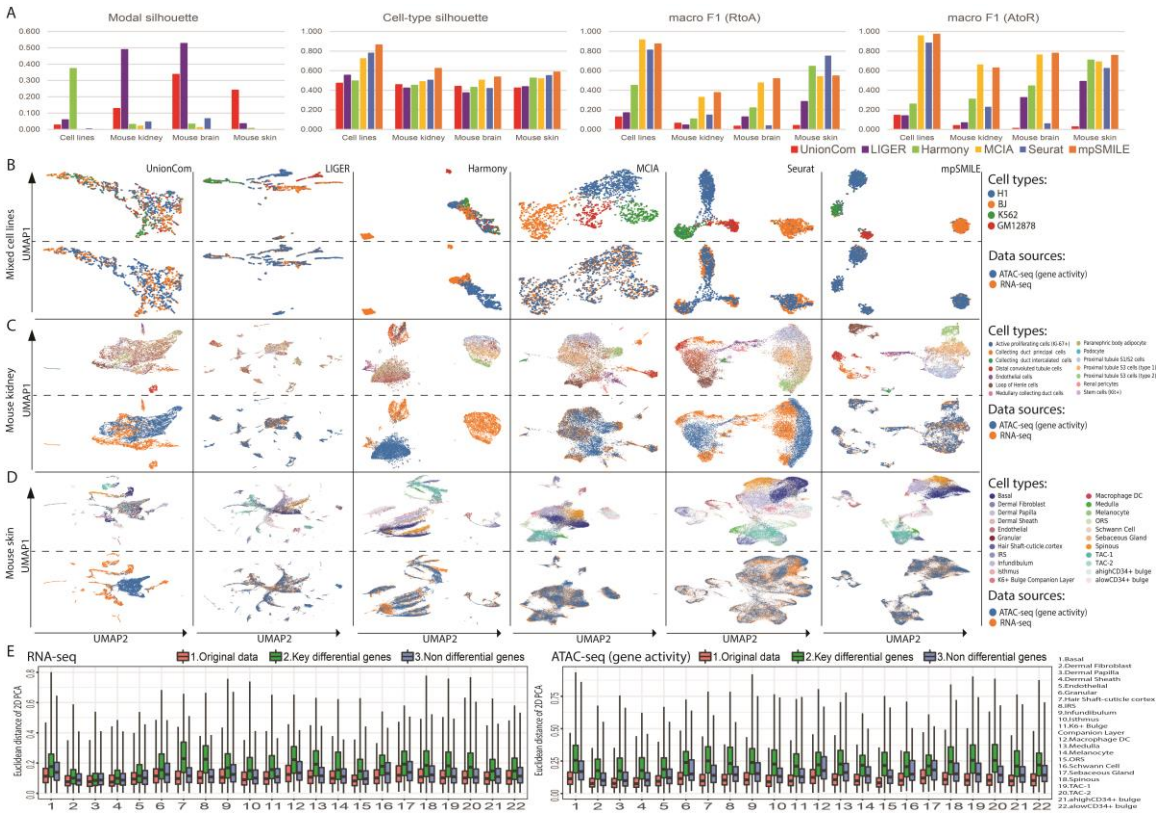


Figure 3.5. Integration of scRNA-seq and scATAC-seq through mpSMILE. A, Evaluation of integration by modal silhouette, cell-type silhouette, macro F1 (RtoA), and macro F1 (AtoR). For macro F1 and cell-type silhouette score, 1 indicates the best performance, and higher is better. RtoA represents label transferring from RNA-seq to ATAC-seq, and AtoR represents from ATAC-seq to RNA-seq. For modal silhouette score, 0 is the best, indicating that both modalities align up. B-D, UMAP visualization of integrated representation of (B) mixed cell-lines, (C) mouse kidney, (D mouse skin data. From left to right, methods used for integration are UnionCom, LIGER, Harmony, MCIA, Seurat, and mpSMILE. Cells are colored by cell-type in the upper panels and colored by data types in the lower panels. E, boxplot of Euclidean distances between paired cells. Salmon box: original data was forwarded through trained mpSMILE and Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D PCA. Green box or blue box: either key differential genes or non-key genes were suppressed to zeros, then the suppressed data was forwarded through trained mpSMILE, and Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D PCA. Upper panel is suppression of key gene expression, and lower panel is suppression of key gene activity.

(Figure 3.5E). Next, since this type of evaluation is rapid and does not require re-training, we asked if we could use a screening approach to identify gene or peak sets that contribute to the co-embedding for each cell-type. We focused on 8 previously identified cell-types (Ma et al., 2020b). Then, we suppressed the gene activity value of each gene to zero one at a time and test how this suppression affects the co-embedding. We then selected the top 5% genes that can disrupt the co-embedding. We observed that the disruption by top 5% screened genes is larger than the disruption by random genes but lower than separately identified key differential genes (Figure 3.6).

Though p/mpSMILE was designed to do joint clustering for joint profiling data, it can be combined with pair-identification tools to achieve integration for non-joint-profiling data. Seurat implements "FindTransferAnchors" function, which can identify quality pairs in bimodal datasets. Here, we combined Seurat and SMILE to achieve integration for non-joint-profiling human hematopoiesis and mouse kidney data(Granja et al., 2019; Miao et al., 2021). Empowered by Seurat, SMILE did decent integration for both non-joint-profiling datasets (Figure 3.7A). Since RNA-seq and ATAC-seq were annotated separately by authors, we can fairly compare the performance of Seurat and SMILE, even though SMILE relies on Seurat for anchor identification. Consistent to our benchmarking in joint profiling data, SMILE demonstrated better modality mixing, in terms of modality silhouette. We found SMILE favorably compares to Seurat in terms of cell-type silhouette and macro F1 from RNA-seq to ATAC-seq. Surprisingly, SMILE significantly outperforms Seurat in label transferring from ATAC-seq to RNA-seq, indicating SMILE learns an integration that better preserves mutual information between two modalities

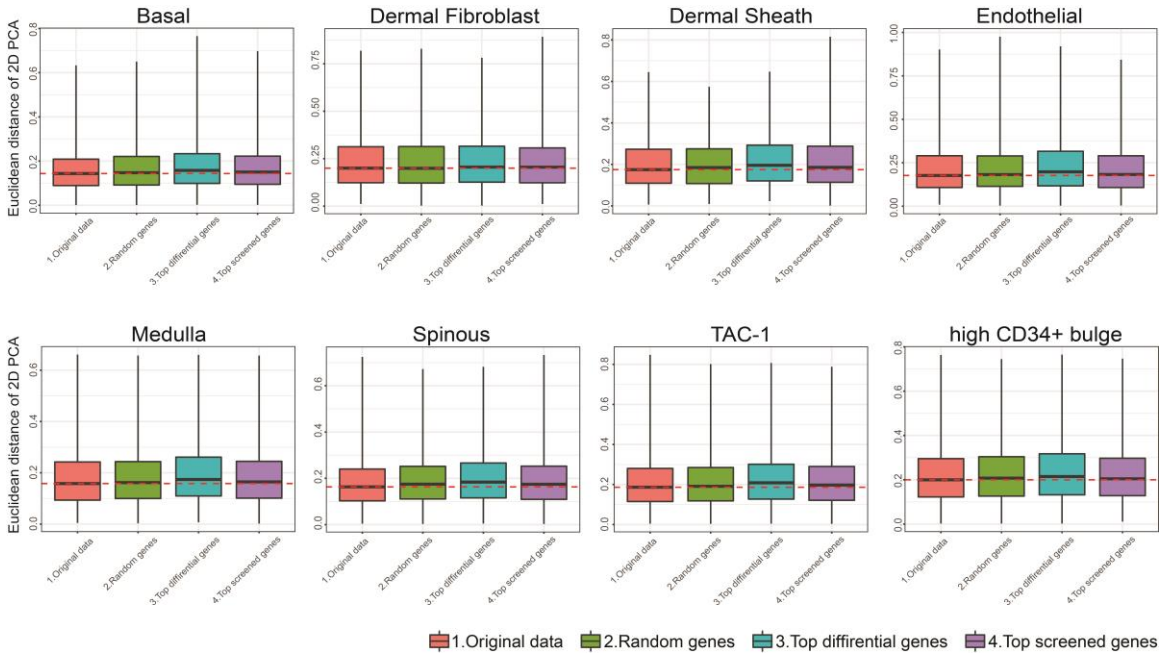


Figure 3.6. Explanation of co-embedding of RNA-seq and ATAC-seq by cluster-specific differential genes and top 95-percentile genes identified through screening. Boxplot of Euclidean distances of paired cells in 2D PCA. Salmon box: original data was forwarded through trained mpSMILE and Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D PCA. boxes of other colors: a random gene set that contains the same number of genes as key differential genes, key differential genes that are specific to each cell type, and top 95-percentile genes identified through screening. These gene sets were suppressed to zeros, and the suppressed data were forwarded through trained mpSMILE. Euclidean distances between cells in RNA-seq and their corresponding cells in ATAC-seq were measured in the integrated 2D PCA.

(Figure 3.7B). Of note, after pairing is accomplished, SMILE can allow the user to input mismatched features for the two modalities.

Application of p/mpSMILE in joint profiling DNA methylation and chromosome structure data

We next evaluated the applicability of SMILE to the integration of joint profiling DNA methylation and chromosome structure data of mESC and NMuMG cells, and the human prefrontal cortex (PFC) (Lee et al., 2019; Li et al., 2019). Unlike integration of RNA-seq and ATAC-seq, it is difficult to match DNA methylation features to chromosome structure features since chromosome contacts are represented in a two-dimensional space. Therefore, using CCA for integration of Methyl and Hi-C as in Seurat would be a challenging task. Indeed, any horizontal integration method requiring matched features will not fit this task. Thus, SMILE has the unique advantage of not requiring feature matching. We applied pSMILE in both mESC and NMuMG data and human PFC data. pSMILE can distinguish mESC and NMuMG cells but only revealed 5 major cell types in human PFC (Figure 3.8A and B). Then, we applied mpSMILE by using Methyl data in place of RNA-seq and Hi-C in place of ATAC-seq, because Methyl data recovers more distinct cell types in Lee et al. (Lee et al., 2019). However, mpSMILE did not reveal more cell types than pSMILE (Figure 3.8B). Because we used all 100kb bins of CG methylation as input for SMILE, it is possible that SMILE cannot fully unlock the discriminative power of methylation data. Thus, we further projected Hi-C cells onto the tSNE space of CG methylation from the original study, but in a SMILE manner. The tSNE space of CG methylation preserves the distinct structure for each cell type identified by the author, and it should save us from learning

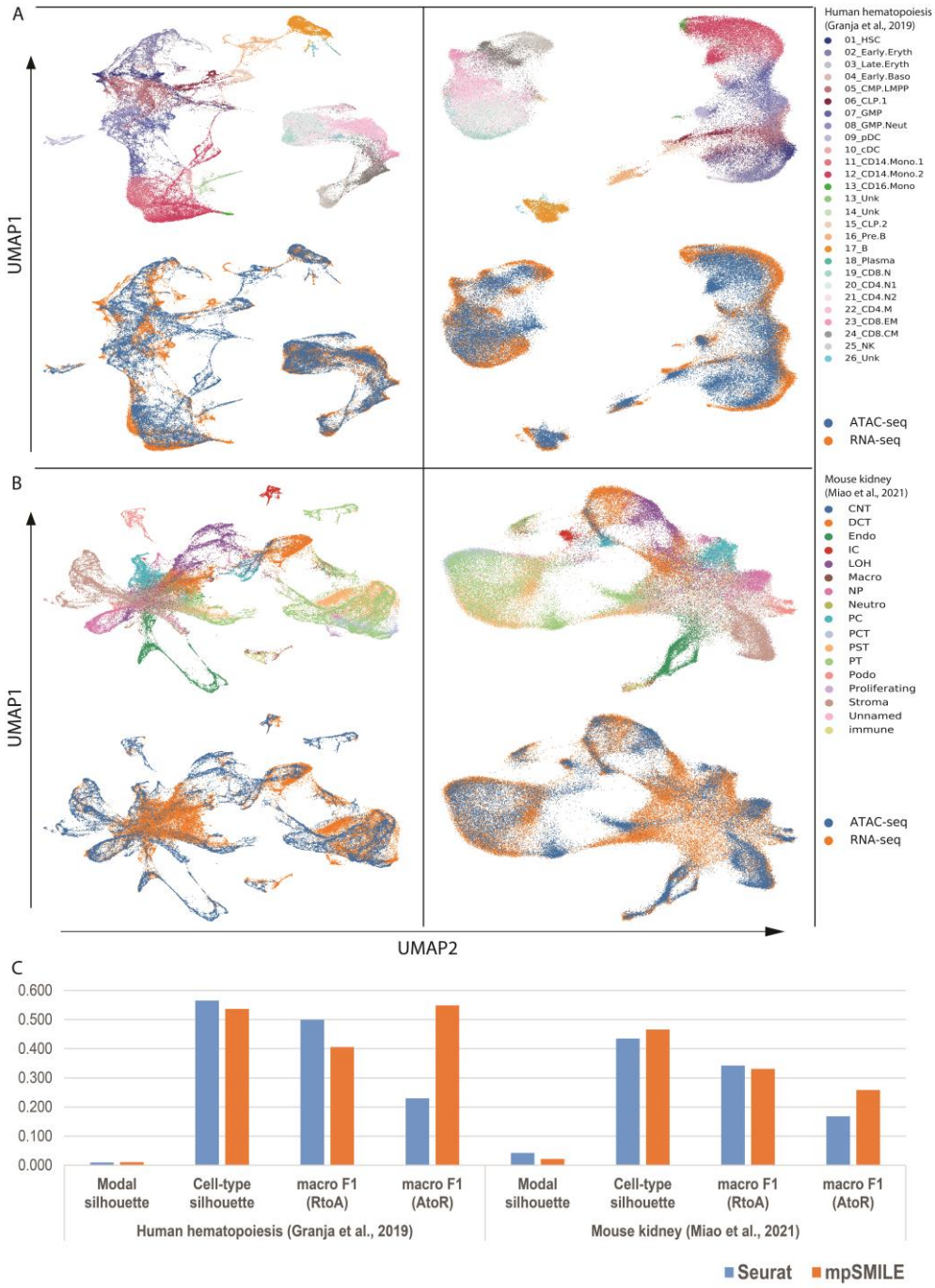


Figure 3.7. Integration of multimodal human hematopoiesis and mouse kidney data through Seurat and mpSMILE. (A and B) UMAP visualization of integrated representation of human hematopoiesis (A) and mouse kidney (B) data, by Seurat and mpSMILE. Cells are colored by author-reported cell types (left panel) and colored by modality types (right panel). (C) Comparison of Seurat and mpSMILE. Calculations of modality silhouette, cell-type silhouette and two macro F1 scores are based on author-reported cell types as the ground truth.

discriminative representation for Methyl data. However, neuron sub-types did not align well with their methylation counterparts in this projection task, though other cell types did (Figure 3.8C). These results suggest that the Hi-C data on human PFC has less discriminative power to distinguish certain neuron sub-types than DNA methylation data, as the original report showed. This is also consistent to a new study in mouse forebrain, where chromatin conformation data has less ability to reveal neuron sub-types than expression data (Tan et al., 2021).

Combining SMILE and pSMILE for integration of more than 2 data modalities

The recently published Paired-Tag technology can jointly profile one histone mark and gene expression from the same nucleus (Zhu et al., 2021). The unique design of this study paired RNA-seq data with 5 different histone marks, and it provides us demonstration data to show how we can combine SMILE and pSMILE to achieve integration of more than 2 modalities. With these modifications, SMILE can integrate RNA-seq, H3K4me1, H3K9me3, H3K27me3, and H3K27ac (Figure 3.9A). In the first step, we used SMILE to integrate RNA-seq data from 6 batches, as we did previously for multi-source transcriptome data. Then, we replaced Encoder A in pSMILE with the trained encoder in SMILE with frozen weights. Therefore, RNA-seq data would be only forwarded through the Encoder A to generate representation za and no gradients will be sent back during training. Since SMILE had already learned discriminative representation for RNA-seq data, training pSMILE in the second step was aimed to project histone mark data into the representation of RNA-seq. Because these histone mark data are not paired, we trained 4 pSMILE models with 4 paired RNA-seq/Histone marker data. Finally, we can project all

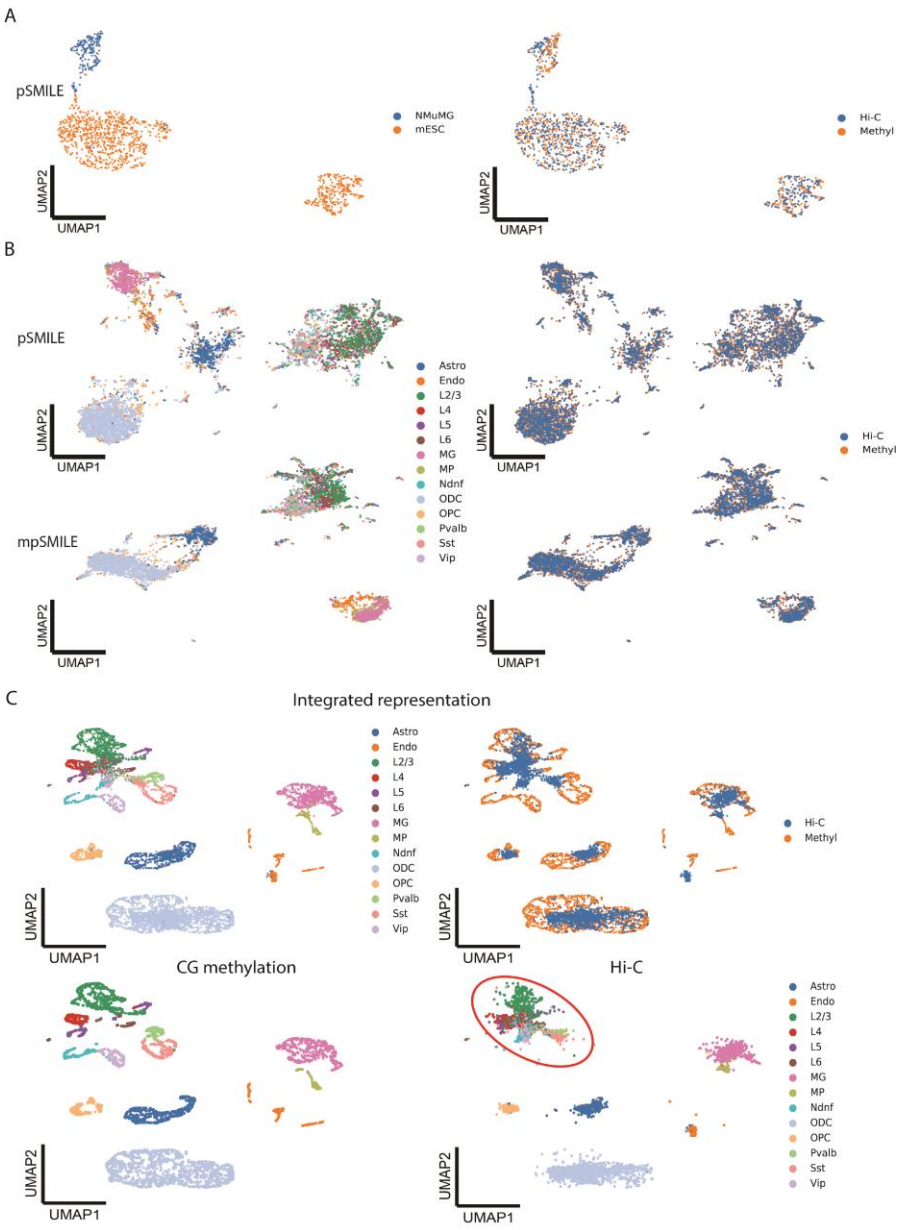


Figure 3.8. Integration of scMethyl and scHi-C through p/mpSMILE. A and B, UMAP visualization of integrated representation of A) mESC and NMuMG cells and B) human PFC data, by p/mpSMILE. Cells are colored by cell-types reported by the author (left panel), or data types (right panel). C, Projection of Hi-C onto tSNE space of CG methylation using SMILE. We used the tSNE space of CG methylation as input for Encoder A instead of 100kb bins of CG methylation. Training SMILE in this case becomes training Encoder B to project Hi-C data into the tSNE space of CG methylation, though we visualized the integrated representation through UMAP. Top row: Hi-C and methylation data on the same graph. Second row: Same representation as above, but with CG methylation (left) and Hi-C (right) shown separately. Red circle highlights region of indistinct neuronal cell types.

nuclei from the 5 modalities into the same UMAP space for visualization. Indeed, this approach preserved distinct properties of cell types while mixing data types (Figure 3.9B). As we did above with the ATAC-seq and RNA-seq joint profile, this learned encoder could be used to screen for which modification peaks are most important for cell type discrimination.

Discussion

Contrastive learning has been extensively shown to learn good representation for different data types (Amid and Warmuth, 2019; Chen et al., 2020a). A simultaneous study further extended contrastive learning to single-cell RNA-seq analysis (Ciortan and Defrance, 2021). However, all these previous studies focus on learning good data representation and have not shown a potential use of contrastive learning in data integration. Here, we designed SMILE, a contrastive-learning-based integration method, and introduced the new use of contrastive learning. We presented three variants of SMILE models that perform single-cell omics data integration in different cases. Through our benchmarking, we demonstrated that our SMILE approach effectively accomplished both batch-correction for multi-source transcriptome data and multimodal single-cell data integration with comparable or even better outcomes than existing tools. Encoders learned by SMILE can be used to determine what biological factors underlie the derived joint clustering and to transfer cell type labels to future related experiments. We further applied our SMILE to the joint Methyl and Hi-C data, and we showed that SMILE can save users from engineering shared features if cell anchors exist for training. Finally, we demonstrated how to combine or modify our SMILE models to address the integration of more than 2 modalities. For the

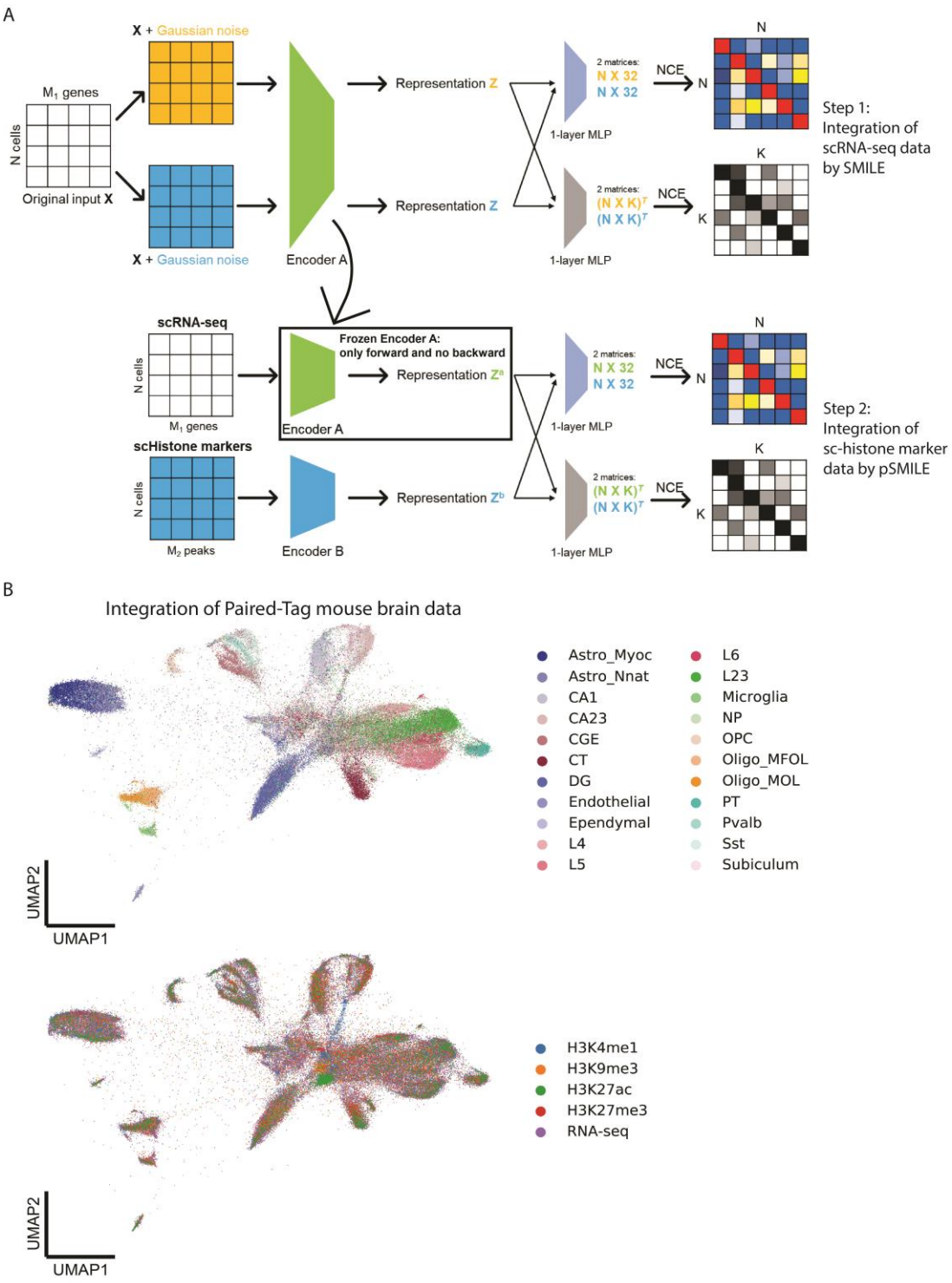


Figure 3.9. Integration of Paired-Tag mouse brain data through SMILE. A, Procedure of combining SMILE and pSMILE for integration of Paired-Tag data. B, UMAP visualization of integrated representation of mouse brain data. Cells are colored by cell-types (upper panel) and data types (lower panel).

joint-profiling data and the Paired-Tag data, the learned encoders could be used to project other single source datasets (e.g., ATAC-seq, ChIP-seq, or Hi-C without paired RNA-seq or Methylation data) into the shared space for cell-type classification or cellular composition analysis across different conditions.

The basic SMILE model demonstrates its ability to remove batch effects, without specific batch-effect modeling. Simply adding random gaussian noises to each gene expression, SMILE could learn to preserve batch-invariant cell-type structure form multi-source data. It could be possible that batch effects exist in the form of gaussian noise. Revisiting we convert gene expression matrix to cell-type score matrix in Chapter II, we may have possible explanation why this conversion addresses batch effects so well without batch modeling neither: Summing all marker genes of one cell-type is countering off random gaussian noises in each marker gene. Another explanation would be avoiding encoding batch-effects into the latent space. Unlike PCA that encodes all possible variation into latent space, MACA and MASI converted gene expression matrix to cell-type score matrix, and each dimension has a biological meaning. SMILE also avoided learning latent representation by capturing major variations.

Integrating single-cell data is still a grand challenge in the community. Among 3 categories of integration approaches, our method falls into the category of horizontal approach for integration of multi-source single-cell transcriptome data and the category of vertical approach for multimodal single-cell data integration. The distinct difference between horizontal and vertical approaches is either using features or cells as anchors. When data are generated through separated single-cell assays, cell anchors are not available

and horizontal approaches can rely on shared features to bring data from different sources to the shared space. However, anchoring cells may be necessary for data integration when engineering shared features is not straightforward. As we demonstrated in joint Methyl/Hi-C data, Hi-C data quantitates 2D interaction features across the genome while methyl measures the methylation level of genomic regions in 1D. Either matching the 1D methylation features to 2D interaction features or vice versus is difficult. Increasing numbers of joint profiling technologies are coming out and will provide more cell-anchored references, and we could combine SMILE with these joint profiling technologies to achieve multimodal integration that brings gene expression, epigenetic modification, chromatin structure and even imaging-based phenotypes to the shared space. With the ability to integrate data without shared features, SMILE has its niche in such scenarios.

In these joint profiling datasets, where the pairing between datasets is already known, it may be less obvious why an integrated representation is needed. Indeed, some uses of such data uses just one datatype to classify cell types and then examines the properties of the other data within those established categories (Lee et al., 2019). We show here that by learning a joint representation where each datatype is separately projected into the space, we can evaluate what biological features are most important for allowing the two datatypes to be embedded in the same space. Further, with this type of joint representation, we can then project a new single source data (i.e., ATAC-seq or Hi-C) from a different experiment onto this joint space. Thus, we can use the power of both paired datatypes to create the representation space and call cell types, and then we can take a new dataset that has only

one type of data and compare it and annotate cell types for the new single data based on the joint space.

Overall, our SMILE approach shows the ability to integrate single-cell omics data as a comprehensive tool. One limitation of SMILE for multimodal integration is that cell pairs must be known. Therefore, training SMILE involves creating self-pairs across a single modality or using the natural pairs in joint profiling data. We combined Seurat and SMILE to show how SMILE can also be used for non-joint profiling data. A benchmark study on computational cell-anchor identification methods would provide insight about anchoring cells with higher accuracy, rather than relying on joint profiling assays.

Ideally, we would like to perform integration without knowing either feature or cell anchors, and developing useful diagonal methods is needed for single-cell community. However, diagonal integration faces extreme computational and theoretical challenges. So far, none of the integration methods in this category have been extensively tested across multiple datasets. In our hands, the diagonal approach, UnionCom, did not achieve any ideal integration of multimodal single-cell data. Therefore, we argue that horizontal or vertical integration still play a critical role in revealing underlying mechanisms for multimodal data. In the end, we leave an open question to the field to discuss if either anchoring feature or cell is necessary to learn the integrated representation for multimodal single-cell data.

CHAPTER IV
INTEGRATING SINGLE-CELL CHROMATIN ACCESSIBILITY AND GENE EXPRESSION DATA VIA CYCLE-CONSISTENT ADVERSARIAL NETWORK

A version of this chapter is a manuscript by Yang Xu, Edmon Begoli, and Rachel Patton McCord. This manuscript is currently under peer-review.

This chapter was revised to be different from the original manuscript. Y.X. conceived and developed the method with guidance from R.P.M. and produced all the figures. E.B. provided critical suggestions on building adversarial network. Y.X. and R.P.M. wrote the manuscript.

Abstract

The boom in single-cell technologies has brought a surge of high dimensional data that come from different sources and represent cellular systems from different views. With advances in these single-cell technologies, integrating single-cell data across modalities arises as a new computational challenge. Here, we present a novel adversarial approach, sciCAN, to integrate single-cell chromatin accessibility and gene expression data in an unsupervised manner. We benchmarked sciCAN with 5 existing methods in 5 scATAC-seq/scRNA-seq datasets, and we demonstrated that our method dealt with data integration with consistent performance across datasets and better balance of mutual transferring between modalities than the other 5 existing methods. We further applied sciCAN to 10X Multiome data and confirmed that the integrated representation preserves biological relationships within the hematopoietic hierarchy. Finally, we investigated CRISPR-perturbed single-cell K562 ATAC-seq and RNA-seq data to identify cells with related responses to different perturbations in these different modalities.

Introduction

Within the last decade, single-cell technologies have advanced our understanding in a broad range of biological systems. Single-cell RNA-seq and single-cell ATAC-seq, along with other single-cell assays, have revealed distinct cellular heterogeneity at a comprehensive level, from genomic variations to epigenomic modifications and transcriptomic regulation (Carter and Zhao, 2021; Kelsey et al., 2017; Macaulay et al., 2017; Stuart and Satija, 2019; Wagner and Klein, 2020). Analyses based on single-cell data have also provided reliable databases for biomedical research and valuable references for medical discovery. As the number of single-cell omics datasets grows, there is increasing demand for fast and accurate computation. Consequently, deep learning has become a trending topic in single-cell data analysis. Much recent research has focused on developing reliable and fast deep learning tools to accommodate the scaling demand, such as cell-type annotation (Ma and Pellegrini, 2020), doublet identification (Bernstein et al., 2020), data de-noising (Arisdakessian et al., 2019), and batch correction (Lopez et al., 2018).

Among all applications of deep learning in single-cell analysis, data integration remains one of the grand and rising challenges in the community (Efremova and Teichmann, 2020; Ma et al., 2020a). Many different single-cell RNA-seq platforms were simultaneously and rapidly developed, leading to an initial focus on methods to integrate datasets from different platforms. Batch effects are usually the most prominent variation when datasets from different sources are collected for integrative analysis but often are not biologically relevant. Single-cell databases confounded by batch effects are not applicable for general use. Therefore, removing batch effects is a critical step for revealing true biological

variation and necessary for building batch-invariant and applicable databases. So far, multiple methods have been proposed to address this problem (Butler et al., 2018; Hie et al., 2019; Korsunsky et al., 2019; Liu et al., 2020; Lopez et al., 2018; Polański et al., 2020). Among these integration methods, deep generative models were also extensively tested in single-cell analysis and demonstrated their efficacy of learning discriminative representation from the original high dimensional space. The most common generative models are Variational Autoencoder (VAE). Variants of VAE models, which differ in their sampling approaches, have been proposed to learn representations for single-cell gene expression data (Bahrami et al., 2020; Dincer et al., 2020; Lopez et al., 2018; Lotfollahi et al., 2019; Wang et al., 2019). The core component of VAE is the use of reconstruction loss, which encodes a sample in a representation that is drawn from a certain distribution, for example, a Gaussian distribution. The use of reconstruction loss also has an advantage of mapping noisy data to high-quality data, which further extends the ability of generative model to de-noise data or impute gene expression. Instead of using VAE to learn representation for single-cell RNA-seq data, two research groups simultaneously modified VAE to address batch effects using an adversarial approach (Bahrami et al., 2020; Dincer et al., 2020). Two methods, named scGAN and AD-AE, respectively, used generative adversarial network (GAN) as the main framework for learning the latent space that is not entangled with batch effects. Starting from a VAE model, both scGAN (Bahrami et al., 2020) and AD-AE (Dincer et al., 2020) introduced adversarial domain loss into the generative model and transferred the learning from reconstruction of data to diminishing of non-biological variation. This approach turned out to be effective in removing batch

effects within single-cell gene expression data. Previous work has only focused on the use of adversarial learning in single-cell RNA-seq data.

Considering the success of deep generative models in batch-effect correction, we extended its use to single-cell data integration across different modalities. In this study, we focus on modality differences and developed an improved adversarial domain adaption approach to address multimodal data integration for chromatin accessibility (ATAC-seq) and gene expression (RNA-seq) data. Our method differs from both scGAN and AD-AE in that it uses a cycle-consistent adversarial network to learn the joint representation for both chromatin accessibility and gene expression data (Zhu et al., 2017). We term our method sciCAN (single-cell chromatin accessibility and gene expression data integration via Cycle-consistent Adversarial Network), which removes modality differences while keeping true biological variation. We previously developed a deep learning method, SMILE, to perform integration of multimodal single-cell data (Xu et al., 2022a). SMILE requires cell anchors for integration. This limits the use of SMILE in cases where corresponding cells are known across modalities. Different from our previous work, sciCAN doesn't require cell anchors and thus, it can be applied to most non-joint profiled single-cell data. We first benchmarked our method with 5 existing methods across 5 ATAC-seq/RNA-seq datasets, and we demonstrated that our method deals with data integration with a better ability to transfer cell type labels in both directions between modalities than the other 5 methods. To demonstrate the method's utility in integrative analyses, we applied sciCAN to joint-profiled peripheral blood mononuclear cells (PBMC) data by 10X Multiome platform and we confirmed that the hematopoietic hierarchy is

conserved at both chromatin accessibility and gene expression levels. Finally, we investigated CRSIPR-perturbed single-cell K562 ATAC-seq and RNA-seq data, and we identified that some cells in both modalities share common biological responses, even though the two modalities were profiled with different gene perturbations. Combining the results above, we expect our work will fill the gap to allow generative models to be used in integrative analysis of multimodal single-cell data.

Results

Overview of sciCAN and potential applications

We first show the model architecture of sciCAN, which contains two major components, representation learning and modality alignment (Figure 4.1A). Encoder *E* serves as a feature extractor that projects both high dimensional chromatin accessibility and gene expression data into the joint low dimension space. For representation learning, we use noise contrastive estimation (NCE) as the single loss function to guide *E* to learn the discriminative representation that can preserve the intrinsic data structure for both modalities. For modality alignment, we use two separate discriminator networks for two distinct uses. The first discriminator network D_{rna} is attached to *E* and is trained with adversarial domain adaptation loss. D_{rna} aims to distinguish which source the latent space *z* extracted by *E* comes from, while *E* is pushed to learn the joint distribution so that D_{rna} is less able to distinguish the modality source of latent space *z*. The second discriminator network D_{atac} follows a generator network *G* that generates chromatin accessibility data based latent space *z* from gene expression data. Adversarial training here will push *G* to find a connection between chromatin accessibility and gene expression data. Since the

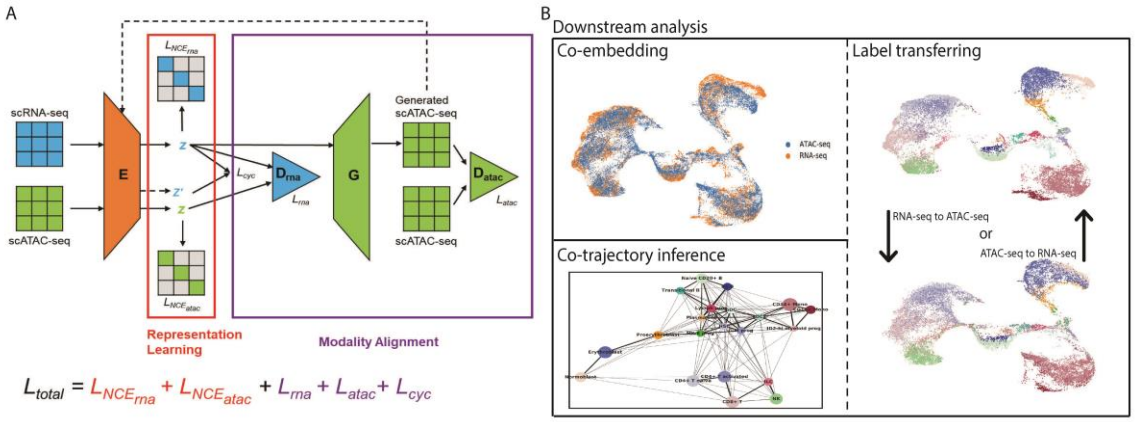


Figure 4.1. Overview of sciCAN and potential applications. A, sciCAN model architecture. sciCAN contains two major components, representation learning and modality alignment. The representation learning part of the model is highlighted in the red box, and the modality alignment part in the purple box. Inputs of scATAC-seq and scRNA-seq have been preprocessed to have the same feature dimensions, so they can share one single encoder E. The final total loss (L) is the sum of loss of representation learning in red and loss of modality alignment in purple. Of note, calculation of NCE is independent for scATAC-seq and scRNA-seq data. B, downstream integrative analyses can include but are not limited to co-embedding, co-trajectory, and label transferring.

generated chromatin accessibility data is based on the latent space *z* of real gene expression data, the new latent space *z′* of generated data should align with its corresponding *z* of real gene expression data. Therefore, we add cycle-consistent loss as demonstrated in cycleGAN method to facilitate finding the connection between two modalities (Zhu et al., 2017). In practice, we build *E* with fully connected layers, which are followed by a batch normalization layer with Rectified Linear Unit (ReLU) activation. D_{rna} takes the 128-dimension *z* as input and forwards it through a three-layer multi-layer perceptron (MLP) to produce 1-dimension sigmoid activated output that predicts if the input *z* comes from single-cell RNA-seq data. Differently, D_{atac} takes output from *G* and forwards the input through a three-layer MLP to produce 1-dimension sigmoid activated output that predicts if input is generated by *G*. *G* is a decoder structure, which has two-layer MLP to restore dimension-reduced *z* to the original dimension of input data. Instead of calculating NCE directly on *z*, we further reduced *z* to 32-dimension output with linear transformation and 25-dimension SoftMax activated output, through two separated one-layer MLPs. This practice is the same as our previous study, in which we demonstrated an effective approach to learn discriminative representation for single-cell data (Xu et al., 2021c). Once model training is done, we use encoder *E* to project both modalities into the joint representation for downstream analyses (Figure 4.1B).

Benchmark of sciCAN with existing integration methods

To demonstrate the competency of sciCAN in the task of data integration, we first selected 3 top integration methods for comparison that have been extensively tested in integrating single-cell RNA-seq data (Tran et al., 2020), including LIGER (Liu et al., 2020), Harmony

(Korsunsky et al., 2019), and Seurat(Stuart et al., 2019). Besides integration specialized methods, we noticed availability of streamline analysis tools for single-cell ATAC-seq data, including ArchR (Granja et al., 2021), MAESTRO (Wang et al., 2020b), and Cicero (Pliner et al., 2018). These streamline analysis tools either built in capacity of integrating ATAC-seq and RNA-seq (ArchR and MAESTRO) or not (Cicero). Both ArchR and MAESTRO used Seurat as infrastructure to integrate ATAC-seq and RNA-seq data, while ArchR did modification to differentiate from Seurat. Thus, we included ArchR in our benchmark test. As sciCAN shares the same architecture as SMILE to learn representation for single-cell data and both methods are proposed for data integration, we also included SMILE. However, SMILE requires cell anchors across modalities to learn the joint representation, but benchmark datasets do not all include this information. Therefore, we used Seurat to identify cell anchors and SMILE would reply on Seurat-identified cell anchors to integrate ATAC-seq and RNA-seq data. For the benchmark purpose, we collected 5 datasets that consist of distinct cellular systems. They are mixed cell lines (Chen et al., 2019), human hematopoiesis (Granja et al., 2019), human lung (Wang et al., 2020a), mouse skin (Ma et al., 2020b), and mouse kidney (Miao et al., 2021), respectively. RNA-seq and ATAC-seq modalities may have different numbers of cells and even different numbers of cell types, except where both modalities were jointly profiled.

We introduced two variants of silhouette score to measure modality mixing and cell-type preserving, respectively. The first metric, modality silhouette, evaluates how well two modalities align, and it directly reports whether discrepancy between chromatin accessibility and gene expression data is removed (maximum alignment gives a score of

0). Across 5 datasets, Harmony, Seurat, and sciCAN integrated chromatin accessibility and gene expression data well, giving a smaller modality silhouette value. Among all methods, LIGER ranked the last in modality mixing, with the worst modality silhouette values in 3 datasets (Figure 4.2A). Though all 6 methods diminish the modality difference between chromatin accessibility and gene expression, it did not necessarily indicate that they learned to present distinctness of each cell-type. This led to the use of cell-type silhouette, which quantifies how well the joint representation reflects the data structure by distinguishing cell-types (in this case, a value of 1 is ideal). Here, we used the author-reported labels as the ground truth. All other 5 methods, except sciCAN, reported the last-ranked cell-type silhouette in the 5 datasets at least once (Figure 4.2A). Though ArchR performs integration upon infrastructure of Seurat, we observed noticeable difference between ArchR and Seurat. Different from Seurat that maps connections between RNA-seq and ATAC-seq data as whole, ArchR only does the "subspace" mapping (Granja et al., 2021), and this "subspace" mapping is highly influenced by a good estimation on correspondence between RNA-seq "subspace" and ATAC-seq "subspace". Considering good balance between modality mixing and cell-type preserving, sciCAN shows the most consistence of integration across the 5 datasets among all methods.

Next, we focused on label transferring. Here, our goal is that the user could rely on the integrated space to predict cell-type labels for data from a single modality, given availability of cell-type labels from the other modality. We found Seurat has overall the best performance for label transferring from RNA-seq to ATAC-seq (Figure 4.2B). This may relate to the design of Seurat. Different from the other 3 methods, Seurat inherently

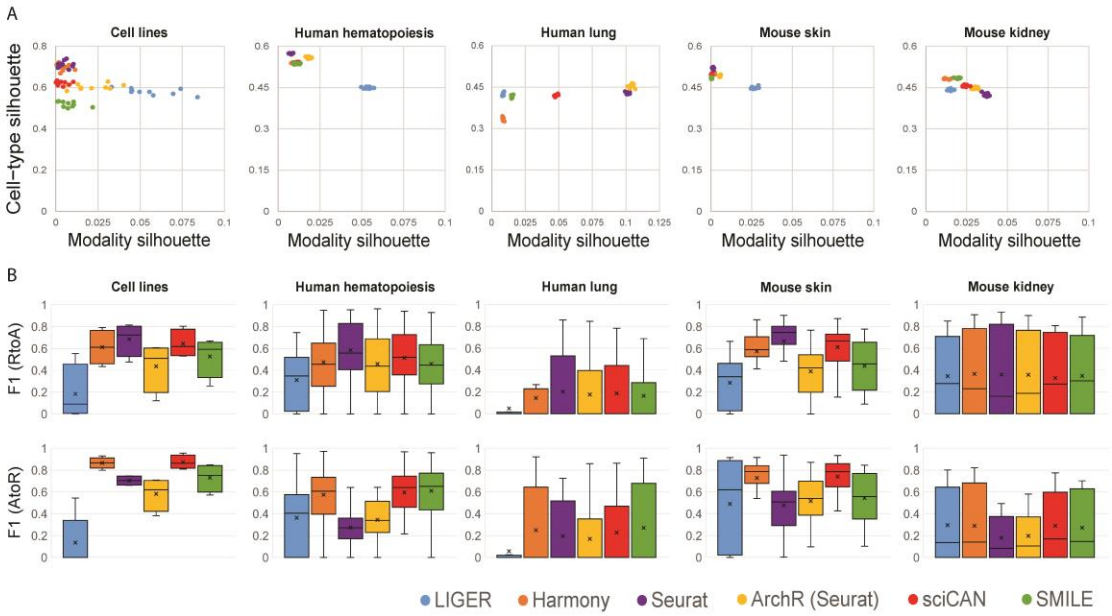


Figure 4.2. Benchmarking of sciCAN against other 5 existing integration methods. A, Integration evaluation by modality and cell-type silhouette scores across 5 datasets. x axis corresponds to modality silhouette score while y axis to cell-type silhouette score. Ideal integration should be in the top left corner of each dot plot. To generate the dot plot, we randomly subsample 20% cell population to calculate both modality and cell-type silhouette scores for each method and each dataset. B, Integration evaluation by F1 scores across 5 datasets. upper panel corresponds to label transferring from RNA-seq to ATAC-seq (RtoA) while lower panel indicates label transferring from ATAC-seq to RNA-seq (AtoR). A Boxplots was plotted based on F1 scores for all cell type in that dataset. The median value was marked with a horizontal line within the box, and the "X" mark represents macro F1 score, which is the average of F1 scores for all cell types.

uses gene expression data as reference data and projects chromatin accessibility data to the gene expression space. Contrarily, sciCAN has overall the best performance of label transferring from ATAC-seq to RNA-seq (Figure 4.2B). Among all methods, LIGER shows the worst performance regarding label transferring (Figure 4.2B).

The default architecture of sciCAN shown in Fig. 1 has RNA-seq data playing the central role, primarily because RNA-seq data usually shows greater discriminative power than ATAC-seq in terms of cell-type identification (Jain et al., 2021; Lin et al., 2021; Peng et al., 2021; Stuart et al., 2019; Xu et al., 2021c). We wondered if this setup is critical to good integration by sciCAN. Thus, we switched the roles of RNA-seq and ATAC-seq data in the model training. Indeed, the ATAC-centered sciCAN model is consistently less accurate than RNA-centered sciCAN, suggesting discriminative representation learning benefits from taking advantage of the cell-type discriminative power of RNA-seq (Figure 4.3). Combining the results above, we conclude that the RNA-centered sciCAN shows consistently good integration performance across different cellular systems.

Integration learned by sciCAN preserves hematopoietic hierarchy

The hematopoietic hierarchy has been extensively studied through single-cell analysis. Independent studies using scRNA-seq or scATAC-seq alone also confirmed that the cellular hierarchy of the hematopoietic system is observed at both chromatin accessibility and gene expression levels (Buenrostro et al., 2018; Corces et al., 2016; Han et al., 2018; Rodriguez-Fraticelli et al., 2018; Velten et al., 2017). Thus, hematopoietic data can be a good example for us to verify whether the integration learned by sciCAN is biologically meaningful. Instead of using scRNA-seq and scATAC-seq data that were profiled

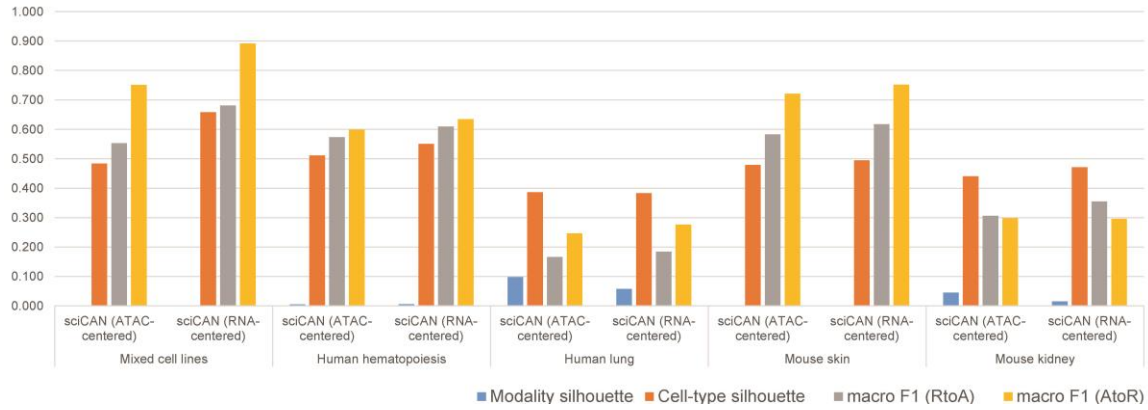


Figure 4.3. Comparison of RNA-centered and ATAC-centered integration by sciCAN. Performances of RNA-centered and ATAC-centered sciCAN were evaluated by modality- and cell-type silhouette scores, and RtoA and AtoR macro F1 scores.

separately, we utilized a jointly-profiled human PBMC dataset obtained through the 10X Multiome platform, which enables us to evaluate the integration with ground truth. Blinding ourselves to cell pairing information, our first task is co-embedding RNA-seq and ATAC-seq and performing co-trajectory analysis to evaluate whether the joint representation learned by sciCAN preserves the hematopoietic hierarchy at both chromatin accessibility and gene expression levels. Indeed, PAGA, a trajectory inference tool for single-cell data, constructed a hematopoietic stem cell (HSC)-centered trajectory with the 128-dimension joint representation learned by sciCAN (Wolf et al., 2019). We also confirmed that progenitor cells surround the HSCs and branch towards their differentiated cells, and their lineage commitments at both chromatin accessibility and gene expression levels can be explained by the same gene signatures (Figure 4.4). Given that the integrated representation learned by sciCAN preserved the hematopoietic hierarchy, we next asked if we could infer transcriptional dynamics between chromatin accessibility and gene expression across the trajectory from progenitor to differentiated cells. To do so, we borrowed and transformed the concept of RNA velocity into activity-expression velocity. In the original RNA velocity concept, positive velocity is inferred when an increase in unspliced transcripts is followed by up-regulation in spliced transcripts (Bergen et al., 2021; La Manno et al., 2018). This idea was further extended to velocity analysis of nuclear mRNA vs cytoplasmic mRNA (Xia et al., 2019), and of more compact vs less compact chromatin regions (Tedesco et al., 2021). Here, we reframed this analysis into activity-expression velocity. We found that the trajectories of the resulting velocity calculation follow the expected hematopoietic differentiation (from stem and progenitor to

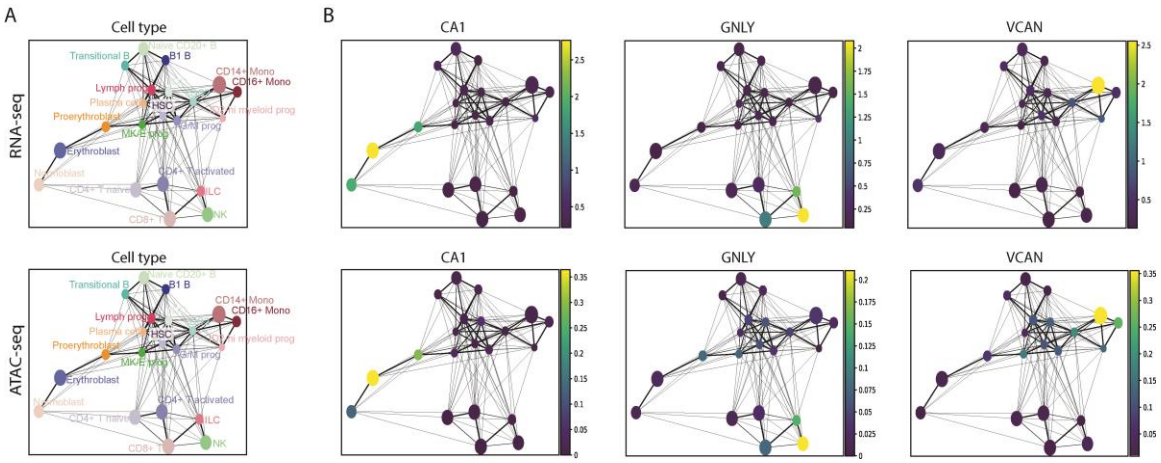


Figure 4.4. Integration learned by sciCAN preserves hematopoietic hierarchy. A, Co-trajectory analysis via PAGA using joint representation learned by sciCAN. Each dot is the sum of all cells annotated as the same cell type. Trajectory is visualized using RNA-seq (upper panel) and ATAC-seq (lower panel), separately. B, Enrichments of signature genes for 3 different lineages using both RNA-seq (top) and ATAC-seq (bottom) data. Color bar indicates gene expression (top) or gene activity level (bottom), respectively.

differentiated type) when we calculate positive velocity as an increase in gene expression first, followed by an increase in gene activity (accessibility). This directionality suggests that in this system gene expression may be activated first, followed by a chromatin state encoding of this expression pattern as the new cell type is established. Given the joint representation, we predicted gene expression based on gene activity. Then, we used the true activity matrix and the predicted expression matrix to compute the activity-expression velocity with scVelo (Bergen et al., 2020). Taking advantage of the ground truth from the cell pairing information, we also performed the same analysis using the true activity matrix and true expression matrix. We found that velocity computed with the predicted expression data resembles and correlates well with the velocity computed with true expression data, in accordance with the correlation between predicted and actual expression (Figure 4.5). Consistent with co-trajectory analysis, velocity with predicted expression data revealed that MK/E progenitor cells move towards erythroblasts while G/M progenitor cells move towards monocytes (Figure 4.5). Combining the results above, we concluded that sciCAN preserves meaningful biological information within the learned joint representation.

sciCAN identifies common responses after CRISPR perturbation

Combining single-cell sequencing with CRISPR enables a systematic examination of cellular response to genetic perturbation. Dixit et al. first introduced Perturb-seq to identify single-cell cellular response at the expression level after CRISPR perturbation (Dixit et al., 2016). Then, Perturb-ATAC was introduced to profile single-cell chromatin accessibility after CRISPR perturbation (Rubin et al., 2019). Nevertheless, a CRISPR-coupled joint-profiling single-cell assay has not been introduced. Therefore, multiple modality data

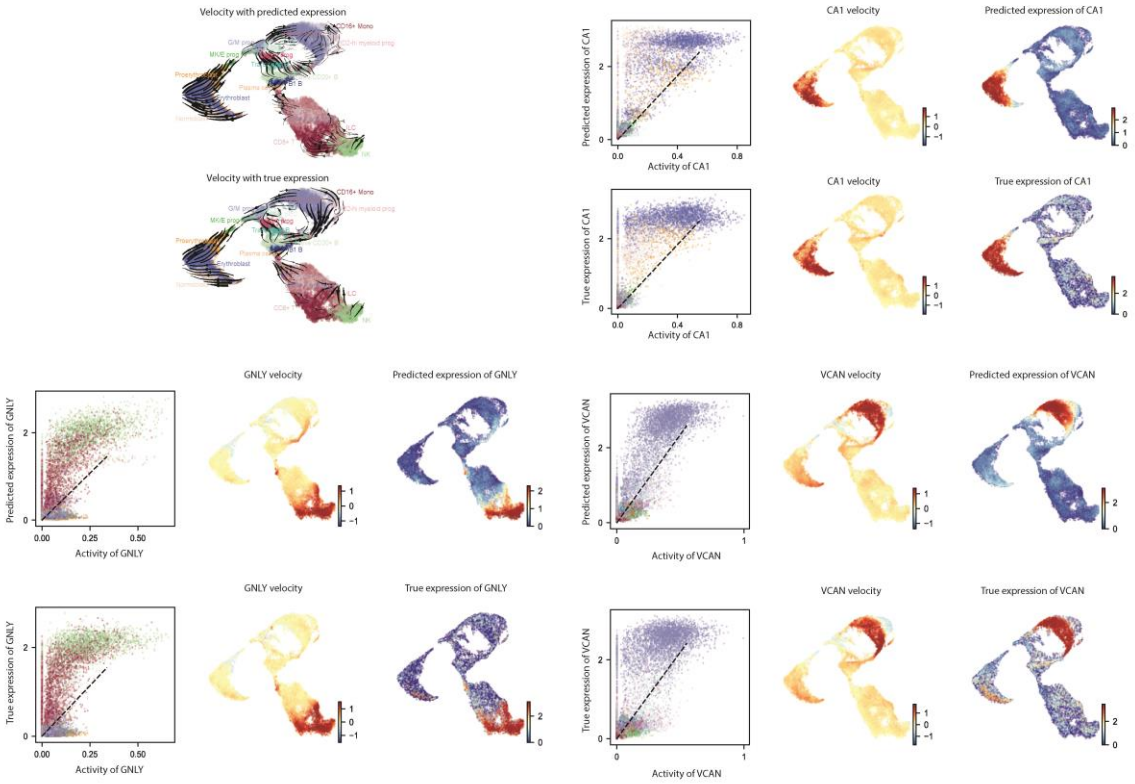


Figure 4.5. Activity-expression velocity of the hematopoietic hierarchy. Velocity was calculated using predicted expression data (upper panel) or true expression data (lower panel). Activity-expression velocity of signature gene CA1, GNLY, or VCAN with either predicted expression data (upper panel) or true expression data (lower panel). Left: CA1, GNLY, or VCAN expression (predicted from ATAC-seq or measured by RNA-seq) is plotted vs. gene activity (accessibility) for each cell. Cell type indicated by color that corresponds to labels in previous panels. Dotted line indicates an estimated 'steady-state' ratio. Area above the dotted line suggests positive velocity, in which opening up of gene accessibility leads to up-regulation of its expression. Middle: the calculated velocity of CA1, GNLY, or VCAN superimposed onto the integrated representation across the hematopoietic hierarchy. Right: the expression of CA1 predicted by ATAC-seq vs. the true expression of CA1, GNLY, or VCAN superimposed onto the integrated representation.

integration is needed to determine how single cell responses to genetic perturbation compare at the transcriptomic and chromatin accessibility levels. As the final demonstration about potential application of sciCAN, we performed computational integration via sciCAN to create a joint view of cellular response after CRISPR perturbation. We selected single-cell K562 RNA-seq data by Perturb-seq and single-cell K562 ATAC-seq data through Spear-ATAC (Dixit et al., 2016; Pierce et al., 2021). Notably these two studies used quite different sgRNA sets, sharing only 3 targets (*sgELF1*, *sgYY1*, and *sgGABPA*), so the integration cannot simply group like targets, but instead will be challenged to find similar biological responses to different gene perturbations. First, sciCAN enabled us to co-embed and co-cluster RNA-seq and ATAC-seq data, and we identified 3 distinct clusters (Figure 4.6A). Next, we asked if the co-clustering makes sense in terms of gene signatures that lead to these clusters. Though the two studies used different sgRNA sets, we found gene activities of these 3 clusters have strong correlation to the gene expression profiles of the corresponding clusters in RNA-seq (Figure 4.6B). Further, cells within each cluster shared gene signatures in both expression and accessibility (Figure 4.6C). This suggests that cells may have similar response to different CRISPR-perturbations. Next, we ranked sgRNA targets for each cluster in both RNA-seq and ATAC-seq data. We found the 3 shared targets are in the top ranking in cluster 1 in RNA-seq but not ATAC-seq (Figure 4.6D). We reason those cellular responses to perturbation at the chromatin accessibility level may be more variable than the responses at the gene expression level. Indeed, none of the ATAC-seq cell clusters have strongly dominant sgRNA targets as seen in the RNA-seq data. Therefore, we separated out cells that were

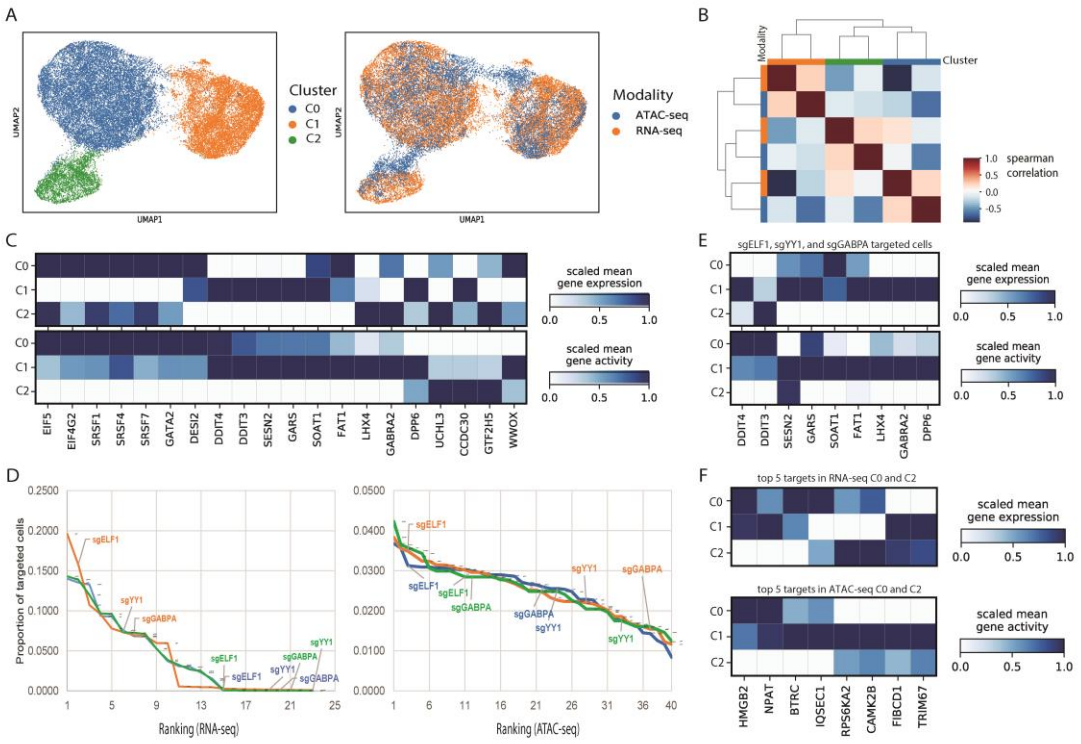


Figure 4.6. sciCAN identifies common response after CRISPR perturbation. A, Visualization of single-cell CRISPR-perturbed K562 RNA-seq and ATAC-seq data via UMAP. Cells are colored by identified cell clusters (left) and modality source (right). B, Spearman correlation between RNA-seq and ATAC-seq profiles of cells in different clusters in both modalities. Gene expression or gene activity matrix was averaged by cell clusters. C, Shared gene signatures of the 3 cell clusters in both modalities. Differential gene activities or expression were identified through 'wilxocon' test in Scanpy package. D, Ranking of sgRNA representation in each cluster (blue = C0, orange = C1, green = C2) in both RNA-seq (left) and ATAC-seq (right) data. Genes perturbed in both experiments are highlighted. E, Gene signatures of cells targeted by sgELE1, sgYY1, and sgGABPA in cell cluster 1. F, Genes whose activity patterns distinguish cells in cluster 0 and cluster 2 among cells in these clusters perturbed by the same gRNAs.

targeted by the common targets *sgELF1*, *sgYY1*, and *sgGABPA* for a closer examination. We found that cells targeted by *sgELF1*, *sgYY1*, and *sgGABPA* in cluster 1 in both RNA-seq and ATAC-seq do have a distinct gene expression and activity signature compared to cluster 0 and 2, even though these cells were perturbed by the same sgRNAs (Figure 4.6E). Shifting our focus to cluster 0 and 2, it is surprising that cells in these two clusters share the same top 5 sgRNAs (sgCEP55, sgOGG1, sgPTGER2 sgCAPBP7, sgCIT), in RNA-seq but are perturbed with completely different sgRNAs in ATAC-seq. To understand what makes cluster 0 and 2 different, we performed a differential gene activity test using cells targeted by the top 5 sgRNAs in cluster 0 and 2 ATAC-seq data. We then examined cells targeted by the shared top 5 sgRNAs in cluster 0 and 2 RNA-seq, and we found that the differential genes we identified through ATAC-seq could partially explain the different clustering of these cells in RNA-seq (Figure 4.6F). Therefore, our integrated representation of these two independent datasets allows us to gain a better understanding of two subpopulations of cells that respond differently to the same gene perturbation.

Conclusion

In this study, we designed a novel adversarial approach for integration of single-cell chromatin accessibility and gene expression data. By benchmarking our method against 5 existing integration methods in 5 ATAC-seq/RNA-seq datasets, our showed that sciCAN and Seurat have overall superior performance of data integration. However, sciCAN shows good mutual label transferring either from RNA-seq to ATAC-seq or from ATAC-seq to RNA-seq, while this mutual information is lost via Seurat integration. In cases where researchers may want to translate ATAC-seq to RNA-seq for inferring gene expression,

sciCAN would have an advantage over Seurat. We further demonstrated that sciCAN can be applied to different integrative analyses, like co-trajectory, activity-expression velocity, and co-clustering. All these results above demonstrate that sciCAN could empower integrative single-cell analysis for novel biological discoveries.

Methods

Representation learning

Deep metric learning has shown effective representation learning without supervision. Chen et al. used a simple framework to learn visual representations in a self-supervised manner (Chen et al., 2020b). They duplicated each image into two counterparts through image perturbation. The goal of learning is to maximize the consistency of any paired replicates in the latent space z. To achieve this goal, NCE is applied as loss function as shown in (1). In an N-sample batch, there will be *2N* samples through data augmentation, and each augmented image *i* has its corresponding counterpart *j* which is the same, despite the added image perturbation. Then, *cos* quantifies the cosine similarity of image *i* and *j/k* in the latent space *z*. Chen et al. demonstrated that this simple framework turns out to be a highly effective way to learn the discriminative representation without supervision. We adapted this approach in our previous study and showed the sample framework can produce discriminative representations for single-cell data (Xu et al., 2021c). Because of the property of this metric learning, our method is fully unsupervised. Users do not need to provide cell-type labels to start model training.

$$(1)\ l_{i,j} = -\log \frac{\exp\left(cos(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \exp\left(cos(z_i, z_k)/\tau\right)}$$

Domain adaptation

Generative models with adversarial domain adaptation were successfully shown to transfer targets to source style and have general applications in image translation (Tzeng et al., 2017). Recently, both scGAN (Bahrami et al., 2020) and AD-AE (Dincer et al., 2020) incorporated adversarial domain adaption into a generative model for removing batch effects within single-cell expression data. For both studies, the goal is to find a batch-invariant representation for single-cell gene expression data from various sources. To achieve this, they stacked a discriminator to the encoder and trained the discriminator to distinguish which source the cell comes from using the latent space *z* projected by the encoder. Adversarial training, in this case, will push the encoder to approximate the joint distribution and become capable of projecting cells with data from different modalities to the same integrated representation. Here, we also used domain adaptation to train a discriminator to identify the modality source while the encoder is pushed to diminish modality difference.

Cycle-consistent adversarial network

Besides the use of adversarial domain adaptation above, we further introduced a cycle-consistent adversarial part. This practice stems from a method called cycleGAN, which presented a state-of-the-art outcome for the task of transferring image styles from one domain to another (Zhu et al., 2017). The success of establishing a connection between two image domains relies on the concept called "cycle consistency". Starting from the original image, a generator network translates the image to the other domain. Then, a second generator network translates the image back to its original domain. Through this cycle, the

translated-back image should be the same as the original image. Based on this information, adversarial training of generators can establish a reversible connection between two image domains. Different from the goal of cycleGAN, we aim to learn joint representation instead of translating chromatin accessibility to gene expression or vice versa. However, the fundamental concept is the same: we establish a cycle from encoder to generator, and from generator back to encoder. Then, the cycle-consistency loss is applied at the level of latent space *z*.

Data preprocessing

All methods benchmarked in our study require anchoring genes for integration. We used a common practice that transforms the sparse ATAC-seq peak matrix to a gene activity matrix (Fang et al., 2021; Stuart et al., 2019; Stuart et al., 2020; Wang et al., 2020b). Here, we briefly explain the rationale behind this transformation. RNA-seq measures gene expression, so in a matrix of single-cell gene expression data, each row represents one cell, and each column contains expression values of one gene. The whole matrix represents gene expression levels of all genes across all cells. ATAC-seq, on the other hand, quantifies how accessible genomic loci are to regulatory proteins. Therefore, in a matrix of single-cell chromatin accessibility data, each row is one cell (the same as single-cell gene expression data) and each column contains accessibility values of one genomic locus. The sum of accessibility values of all genomic loci upstream of and within one gene body may relate to the potential of transcription of that gene. Therefore, to convert ATAC-seq data to a form that can be compared to RNA-seq data (a matrix of cells by genes), all accessibility peaks upstream of and within each gene body are summed to represent gene activity. In the

converted gene activity matrix, each row is one cell, and each column is accessibility values of one gene. Therefore, after conversion, we can do a simple filtering and reordering to match features of chromatin accessibility and gene expression data. The Signac package provides this conversion process, and we ran the code available at https://satijalab.org/signac/articles/pbmc_vignette.html (Stuart et al., 2020). After we have both a gene activity matrix and a gene expression matrix, we normalize both modality data with (Log+1)-transformation, which adds 1 as a pseudo count to the matrix before log-transformation. Then, we identify the top 3000 highly variable genes (HVG) for each modality and use all identified HVG as features for integration. To identify the top 3000 HVG, we use Scanpy by calling the highly_variable_genes function (Wolf et al., 2018).

Model training

We trained sciCAN in all datasets for 100 epochs. The learning rate starts from 0.005 with 0.0005 weight decay. All weights in the sciCAN model are updated through stochastic gradient descending. In the *NCE* loss function, temperature τ is a crucial parameter that affects discriminative power of the final representation. We set as τ = 0.15 for the 32-dimension linear-transformed output and τ = 0.5 for the 25-dimension SoftMax activated output, which is consistent to the practice in our previous study (Xu et al., 2021c). Detailed training code is also provided on sciCAN GitHub (https://github.com/rpmccordlab/sciCAN).

Integration via LIGER

Multimodal single-cell data integration by LIGER was demonstrated in its published tutorial (Liu et al., 2020). We used default parameters to perform integration of chromatin accessibility and gene expression data, and the final dimension of integrated representation by LIGER is 20 for all 5 benchmark datasets. Briefly, LIGER uses integrative nonnegative matrix factorization (iNMF) to identify metagenes that are shared between ATAC-seq and RNA-seq (Yang and Michailidis, 2016). These metagenes are a weighted matrix of factor loadings of observed gene expression/activity. Then, cell loadings of these metagenes are used to perform joint clustering and other downstream analysis. Ideally, representations of cells from both modalities after iNMF should have been integrated in the same latent space and can be visualized via tSNE or UMAP (Becht et al., 2018; Kobak and Berens, 2019).

Integration via Harmony

Harmony is the second integration method benchmarked in our study. Originally, Harmony was designed to correct batch effects within single-cell RNA-seq datasets (Korsunsky et al., 2019). Later, the novel use of Harmony in multimodal single-cell data integration was discussed in reviews (Argelaguet et al., 2021; Forcato et al., 2021). Meanwhile, a batch-correction benchmark study showed that Harmony was ranked among the top 3 methods, with LIGER and Seurat, for integrating single-cell RNA-seq data (Tran et al., 2020). Therefore, we included Harmony in our benchmarking of multimodal single-cell data integration. Harmony learns the joint representation through an iterative k-means clustering, and the outcome is a linear correction function that transforms the original principal components (PCs) to the batch-corrected PCs. Batch information is necessary to

guide Harmony to distinguish what variation should be diminished during the k-means iterations. Principally, to integrate chromatin accessibility and gene expression data, modality information serves as the same role of batch information. Again, we used the default procedure of Harmony, in which we reduced the whole dataset into the first 30 PCs.

Integration via Seurat

Seurat uses canonical correlation analysis to learn the shared latent space between two modalities. This approach is different from LIGER, Harmony, and our method, in a way that Seurat will first identify confident cell pairs between the two modalities. Then, Seurat uses these paired cells as anchors to learn a mutual neighborhood graph. Finally, it computes a projection that brings all other cells to this shared latent space. Because of its "anchor" design, Seurat needs pairwise computation of anchor points when datasets come from more than two sources. Since we only deal with the modality difference between chromatin accessibility and gene expression in this study, we do not need to perform pairwise computation of anchor points with Seurat. For benchmarking, we ran Seurat v3 with the default tutorial, and the final dimension of integrated representation by Seurat would be 50.

Integration via ArchR

ArchR uses Seurat as infrastructure to integrate RNA-seq with ATAC-seq data. Different from Seurat, ArchR constrains the mapping from ATAC-seq to RNA-seq in a "subspace". An initial unconstrained mapping was done through Seurat. This step is aimed to estimate what clusters in ATAC-seq have good correspondence to a certain number of clusters in

RNA-seq. Then, the "subspace" or constrained mapping will only project a number of clusters in ATAC-seq onto their corresponded clusters in RNA-seq.

Integration via SMILE

We previously proposed SMILE to integrate multimodal single-cell data when cell anchor information was obtained from co-assay profiling. Because sciCAN and SMILE share the same architecture to learn lower dimension hidden space for single-cell data, SMILE also generate 128-dimension hidden space. To use SMILE for integration in this situation, we had to rely on external tool, like Seurat, to identify cell anchors. Once cell anchors identified, SMILE was trained based on anchored data and projected the rest of unanchored data into the joint representation space. A tutorial can be found at SMILE GitHub (https://github.com/rpmccordlab/SMILE).

Activity-expression velocity

Activity-expression velocity was calculated with scVelo (Bergen et al., 2020). We replaced the spliced layer with the gene activity matrix and the unspliced layer with the gene expression matrix, given the concept that increase of gene expression would follow increase in gene activity. To estimate first and second moments, we used the 128-dimension joint space learned by sciCAN, instead of PCA space.

Evaluation

To evaluate integration by each method, we proposed 4 metrics:

Modality and cell-type silhouette score. As we mentioned before, sciCAN and SMILE reduces each dataset into 128-dimension spaces, while LIGER reduces the data to 20

dimensions, Harmony to 30, and both Seurat and ArchR to 50. Since final dimensions of the integrated representations by the 6 methods are not the same, we further used Uniform Manifold Approximation and Projection (UMAP) to reduce them into 2-dimensions with the same UMAP running parameters (McInnes et al., 2018). Then, we calculated modality and cell-type silhouette scores on the 2D UMAP spaces. A typical silhouette score *S* ranges from -1 to 1. To better reflect the integration outcome, we define modality silhouette as $abs(S)$ and cell-type silhouette as $(1 + S)/2$. Of note, we used different labels to calculate modality and cell-type silhouette. For modality silhouette, the label used is modality information. A good integration should have chromatin accessibility and gene expression data largely overlapped. Therefore, 0 is the best outcome, and we ignore the positive/negative sign by using the absolute value of the typical silhouette score *S*. For cell-type silhouette, we used the author-reported annotation label to calculate *S* and then scale the output to the range from 0 to 1. Thus, cell-type silhouette 1 indicates the best integration that preserves cell-type structure.

F1 score from RNA-seq to ATAC-seq, and from ATAC-seq to RNA-seq. A useful integration of modalities should have the ability to transfer cell type labels from one datatype to another, either from RNA-seq to ATAC-seq or from ATAC-seq to RNA-seq. Given cell-type label availability from a single modality, the user should be able to predict cell-types for the other modality, with a fair accuracy. To evaluate how friendly the joint representation is for label transferring, we trained a Support Vector Machine (SVM) classifier with one modality and tested it with the other modality. The choice of SVM is simply based on a constant superior performance of SVM classifier across datasets. Then,

we used macro F1 and F1 score for each cell type to evaluate SVM classifiers trained with different joint representations by these 6 methods. Macro F1 score is the average of F1 scores for all cell-types, and it can help us reveal if integration is good for non-major cell-types. This is because cell-types are not balanced in most single-cell data and revealing non-major cell-types is critical for most single-cell analysis. A high macro F1 score can suggest that integration is also good for non-major cell-types. Meanwhile, individual F1 scores for all cell type also report which cell-type prediction is the hard case and what is the highest F1 score the classifier can reach to.

CHAPTER V
DIAGONAL INTEGRATION OF MULTIMODAL SINGLE-CELL DATA: AN ENCHANTING GOAL BUT A HAZARDOUS JOURNEY

A version of this chapter is a manuscript by Yang Xu and Rachel Patton McCord. This manuscript was published in *Nature communications*.

This chapter was revised to be different from the original manuscript. Y.X. conceived concept with guidance from R.P.M. and produced all analysis results. Y.X. and R.P.M. wrote the manuscript.

Introduction

With the advance of new single-cell technologies, single-cell computational analysis has moved into the multi-omics era. Integrating multi-omics single-cell data, therefore, has gained increasing attention from the single-cell community. This key research domain promises to help us understand complex cellular systems from different viewpoints, such as gene expression, chromosome structure, and even cellular imaging. Computational integration methods that match one modality with another can reveal a detailed picture of regulatory networks and cellular function. However, different types of 'omics data usually do not share the same features. For instance, transcriptomics describes expression of genes, while epigenomics measures histone modifications or accessibility across all regions of the genome. This feature discrepancy presents the first challenge to the development of integration methods. The other challenge stems from how single-cell data have been collected over the years. Though recent technologies enable multiple measurements to be made simultaneously on the same single cells ("joint-profiling") (Lee et al., 2019; Li et al., 2019; Ma et al., 2020b), most single-cell datasets profile different aspects of biology one at a time in independent groups of cells. Therefore, we lack ground truth about what is happening at the level of epigenetics, transcriptomics, and proteomics in the same single

cell. This makes it difficult to evaluate the quality of proposed integration methods. In recent years, many integration methods have been published to address different scenarios of multimodal single-cell data integration. A recent key review summarized three major approaches of multimodal single-cell data integration and outlined published methods in each category (Argelaguet et al., 2021). Of these categories, "horizontal integration" methods require anchored features to align up different modalities, while "vertical integration" methods need shared cells from multiple modalities as anchors. The "diagonal integration" approach requires neither anchoring cells nor features for integration, presenting a distinct advantage over horizontal and vertical methods. Because no prior knowledge is required, accurate diagonal integration is also challenging to achieve. Despite the rapid increase in new diagonal integration methods, there is not a single diagonal method that has been extensively examined and carefully benchmarked for its utility in multimodal integration in complex cellular systems.

The enchanting goal

In this comment, we focus solely on diagonal integration. Over the past three years, there has been a steady increase in publications describing new diagonal methods for the integration of multimodal single-cell data (Cao et al., 2020b; Cao et al., 2021; Demetci et al., 2020; Liu et al., 2019; Stark et al., 2020; Welch et al., 2017; Yang et al., 2021), indicating strong interest in the unique advantages of diagonal integration. Since horizontal and vertical methods require either anchored features or anchored cells, their application is limited to cases where it is feasible to engineer matched features (which is often quite difficult, particularly with disparate measures such as cell imaging and gene expression) or

where multiple modalities have been measured within the same cell. Therefore, an effective diagonal integration method would greatly expand the scope of possible data integration and is enchanting to the community. When we considered the mechanisms that previously published methods use to align modalities, we observed that they are all similarly built upon the foundation of manifold alignment, which projects data from different modalities into a common space while preserving the intrinsic structure within each modality. Therefore, these methods can generally be described in two steps: 1) preserving cell type structure within each modality; and 2) aligning cells across modalities. Each method differs with respect to the representation learning that preserves cell-type structure within each modality and the alignment approach to close the gap between modalities. Thus, they try to solve two problems at the same time and have varying performances of balancing representation learning and modality alignment. Nevertheless, they all share the same underlying principle.

The hazardous journey

Manifold alignment assumes that data from different modalities were generated from a similar distribution or through a similar process. In an ideal experiment, quantification of multi-omics data may satisfy this requirement. But, in reality, there are many unknown variations, and different research labs have different practices of data generation. Therefore, we need to ask how an algorithm distinguishes a true biological alignment that correctly matches the same cell types in different modalities from any other potential artificial alignments. The only judgment the algorithm can make is whether the alignment is the optimal solution. Thus, any artificial alignment that satisfies a mathematical optimum

can stand out as the best solution, but will not necessarily represent the accurate biological solution. There seems to lack ais no mechanism for diagonal algorithms to distinguish a true biological alignment from any artificial alignments without prior knowledge. To demonstrate this pitfall, we illustrate artificial and biologically incorrect alignments resulting from integration applied to a simulated multimodal dataset generated from real single-cell data where the ground truth is known. We began with single-cell RNA-seq data from mouse cortex and split the genes into two parts to represent two different "modalities" with different feature spaces, but which come from the same cell population (Figure 5.1) (Zeisel et al., 2015). We preserved some shared genes between the two modalities, and both modalities should have a similar power to distinguish the seven cell types. We tested five diagonal methods on five simulated scenarios (Cao et al., 2020b; Cao et al., 2021; Demetci et al., 2020; Liu et al., 2019; Yang et al., 2021). These methods can distinguish cell types in both modalities separately, and they all align both modalities with no noticeable gap. However, when we investigate cell type correspondence between modalities, we find that these methods all fail at least in one scenario in terms of accurately matching cell types. Since these methods share fundamentally the same mechanism for modality alignment, we conclude that such errors in alignment will be a widespread problem across diagonal methods. We propose that the use of such simulated data should provide a benchmark for future method developments. Developers can investigate in which scenarios their methods may fail and potential reasons for this failure.

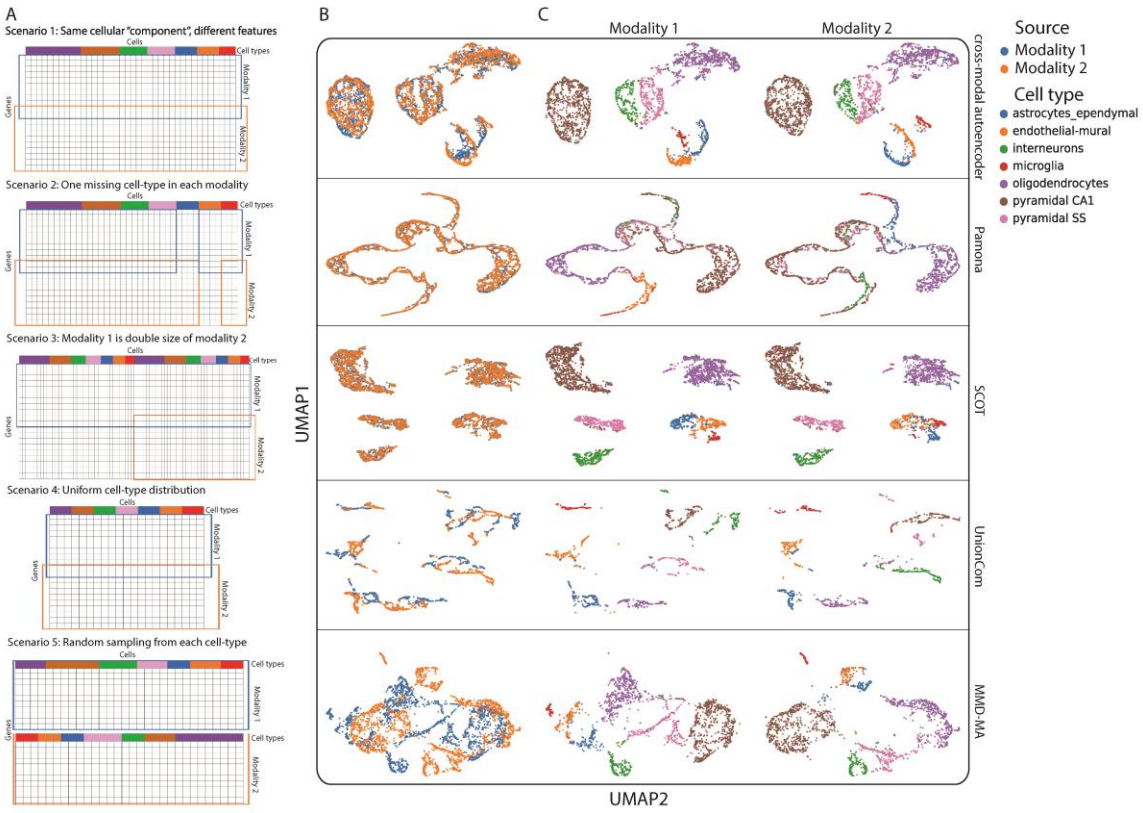


Figure 5.1. Artificial alignments by diagonal methods in 5 scenarios. A, 5 scenarios of simulated multimodal single-cell data, showing how each modality was generated. B and C, Visualization of integration by selected diagonal methods. Cells are colored by modality source (B) and cell type identity (C). The two modalities were split into separate visualizations in c to make artificial alignment errors visible.

Searching for solid ground

Given the outcomes above, we argue that a safe practice in applying diagonal methods is to incorporate certain prior knowledge. Indeed, Yang et al. briefly mentioned that more than one alignment can look equally optimal and incorporating prior knowledge can help deal with issue of artificial alignments (Yang et al., 2021). At the same time, Pamona only succeeds in complicated integration when it uses shared cells across modalities (Cao et al., 2021). Both publications briefly acknowledge the possibility of artificial alignment we comment on, but this issue has not been highlighted consistently as a key message for those who intend to apply these tools for data integration. Instead, the problem of diagonal integration may come across as solved, and users run the risk of pursuing hypotheses based on erroneous artificial alignment. For example, users could falsely think an enriched signature in one type of data is correlated with an enriched signature in another data type, even though the two aligned cell types in two modalities are not the same.

Considering the incorporation of prior knowledge into future method development, we suggest the following directions here. The first direction is to use partly shared features (Figure 5.2A). Incorporating shared features is feasible for datasets like RNA-seq, ATAC-seq, and other data that are quantified along the linear genome. A pioneering study proposed using partially shared features and extensively benchmarked this hybrid approach with well-established and reliable integration methods (Jain et al., 2021). Moving forward, we recommend additional work should continue to investigate how to achieve meaningful integration with minimal shared features. Along with our recommended simulated data above, there is a need for benchmarking datasets that can be used to evaluate the degree

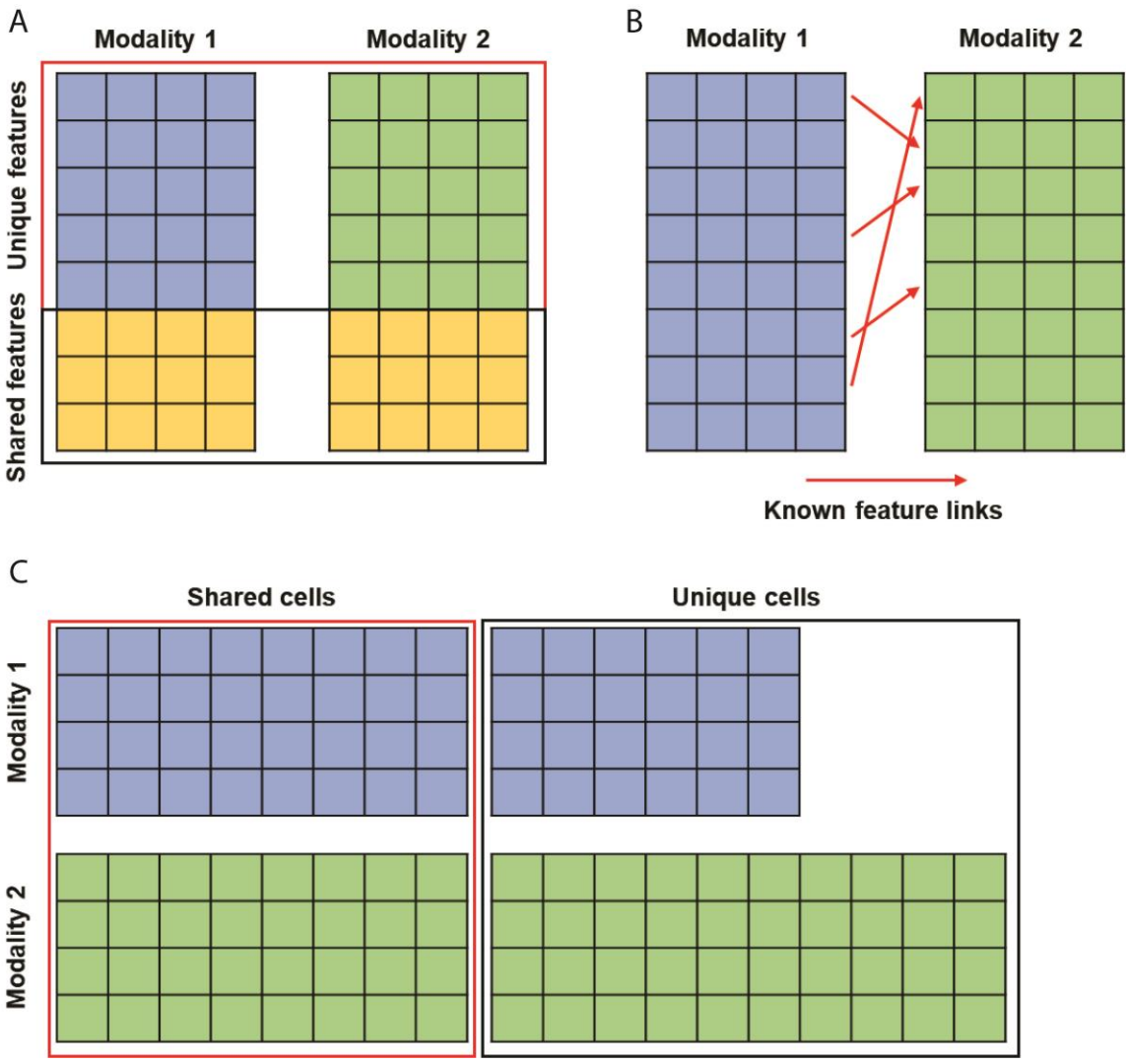


Figure 5.2. Conceptual models for integration of multimodal single-cell data. Models can be designed to consider (A) partially shared features, (B) known feature links between two modalities, and (C) shared cells as prior knowledge.

and type of shared features that are required to achieve accurate integration. Meanwhile, engineering different modalities to have shared features may not be applicable in cases like integrating gene expression data with chromatin structure data. In such cases, alternative approaches can be constructing a feature-relation matrix, which links features in one modality to possible corresponding features in the other (Figure 5.2B). For example, given an enhancer-promoter contact in Hi-C data, we can hypothesize which gene would be under impact and which histone mark may explain the regulation (Duren et al., 2021). However, this approach must be developed with substantial underlying knowledge to support the presumed feature connections. There are also cases in which the construction of feature-relations is not straightforward or lacks experimental support, as in the integration of single-cell omics and single-cell imaging data. This leads to our second recommended direction, using cell anchors or cell labels (Figure 5.2C). In this case, the integration task will be reframed into semi-supervised learning. In recent years, joint-profiling technologies generated multi-omics data at single cell resolution (Lee et al., 2019; Ma et al., 2020b; Zhu et al., 2021), and these joint-profiled single-cell data could serve as reference for learning the integrated space. We envision that combining joint-profiling technologies and diagonal methods would become a standard framework for multimodal single-cell data integration. Further work is needed to determine how many cells must be profiled by joint methods to represent sufficient complexity to facilitate integration of disparate datasets. Even so, algorithms could misalign cell types that do not show up in the training set. Thus, methods should be evaluated for whether they force all data to be aligned to the previously represented cell types or would allow them to be separate.

As diagonal integration gains more attention, the problem of artificial alignment and the two future directions we propose remain major challenges to overcome. When applying diagonal methods in complex situations, the community needs to cautiously evaluate conclusions generated by these methods. In a fast-moving and competitive field, there is strong temptation to show only the advantages of a new method and where it succeeds, making broad claims of general utility while minimizing any potential shortcomings. But it is equally valuable to clearly show scenarios where methods fail, both to inform potential users and to facilitate future research. We encourage the community to contribute additional guidelines for reliable use of diagonal integration methods and to propose additional challenging benchmark tests that will clearly reveal what problems are yet to be solved.

CONCLUSION THE DIVERSITY OF MULTI-OMICS

The conclusion chapter contains a partial manuscript by Yang Xu and Rachel Patton McCord, which was published in *BMC Bioinformatics*. This chapter also includes some unpublished results that will be considered for peer review.

1. Power single-cell transcriptome for diverse research projects

In Chapter I, I introduced a marker-based cell-type annotation tool MACA for single-cell transcriptome data (Xu et al., 2021b). In Chapter II, we built a marker-assisted integration tool termed MASI, based on the study of MACA (Xu et al., 2022b). We demonstrated that the marker-based integration approach outperforms model-based methods, even deep-learning models, regarding batch correction and cell-type annotation for multi-source single-cell transcriptome data. Many computational methods were proposed to address the issue of integration for single-cell transcriptome data. These methods used sophisticated designs to model batch effects and would require intensive computation. However, our methods, MACA and MASI, indicated that these methods may complicated the problem of batch correction. Instead, MACA and MASI used a general data processing pipeline, and we demonstrated this simple practice deals with batch effects in a wide variety of cases. We could propose these two marker-based approaches to diverse research projects that involve single-cell transcriptome data.

1.1. The diverse research projects

Combining single-cell transcriptome technology with chemical treatments could help us understand cellular functions under a range of conditions at system level. For example, Kang et al. treated PBMC cells with IFN-β and then performed scRNA-seq to study how

different immune cell types response to the same treatment (Kang et al., 2018). Another study combined small molecules that induce somatic cells back to pluripotent stem cells with scRNA-seq to identify key intermediate states of cell reprogramming (Guan et al., 2022). Studies like this could draw a picture of how a small molecule leads differentiated cells back to pluripotent cells and what gene modules are reprogrammed along the path. Besides treating cells with chemicals, combining single-cell transcriptome technology with genetic or molecular engineering tools could generate even more diverse single-cell data to understand more complicated functions of gene networks. A CRISPR-coupled scRNA-seq could be used to study cell-type-wise response of a cellular system to genetic perturbation (Dixit et al., 2016). Injecting cells with tracible tags and then using scRNA-seq to profile these tagged cells could reveal what the their lineage fates are and what intermediate states these cells went through (Bandler et al., 2021).

Studies can also be designed to cover multiple conditions. For example, we have just been through the COVID-19 pandemic, and we are still working on to reveal molecular the mechanism of SARS-CoV-2 in detail. Three independent studies from multiple research institutes recorded cellular profiles of patients, from mild, to moderate, and to severe with scRNA-seq, and this research group revealed how SARS-CoV-2 infection progresses in our immune systems in multiple tissues and even in different races (Chan Zuckerberg Initiative Single-Cell et al., 2020; Chua et al., 2020; Zhang et al., 2020). These studies together could draw a more comprehensive picture about COVID-19, instead of them alone. More beneficially, combining all these 3 studies, we would be more confident to

make a medical and public health plan to intervene immune defense, prevent a patient deteriorate to a severe stage, and lower the risk of death.

There are even more single-cell studies similar to those above. We are witnessing that numerous research groups across the globe empower single-cell transcriptome data for diverse purposes. Integrating these single-cell data serve as keys to unravel mysteries of complicated cellular systems and molecular networks.

1.2. Towards data-driven integration

In MACA, all marker genes were given equal weights for their contribution in defining cell types. In MASI, we differentiated marker weights given what ranking the marker is in the reference data. Nevertheless, both approaches of assigning marker weights didn't consider how much contribution a marker gene makes in a particular data. Learning marker weights based on data itself could enable more precisive integration analysis. Meanwhile, data-driven integration can also deal with the problem of generalization. In most supervised machine learning approaches, models learned from a reference are not applied well to new target data. Additional weight fine-tuning on the target data itself is required for the purpose of good generalization. It should be the same for our marker-based annotation. Neither assigning markers equal contribution nor ranking markers based on reference data could fit property of target data.

Thus, we wish to learn marker weights based on data itself in the future. This could be achieved through self-supervised or unsupervised learning. For self-supervised learning, marker weights can be fine-tuned with pseudo labels. In MASI, we have shown that annotation based on reference data correctly predicts cell-type labels for majority of cells.

Self-supervised learning could further take advantage of the existing correctness to further correct wrong labels (Asano et al., 2019; Caron et al., 2018). On the other model, we could also learn marker weights from scratch by an unsupervised approach. In recent years, contrastive learning has demonstrated its effectiveness to learn visual representation for unlabeled data (Chen et al., 2020a; Chen et al., 2020c). The same practice can be used here for learning marker weights that better represent the data itself. We have been working on both self-supervised and unsupervised approach towards data-driven integration, and data-driven integration will still be our next goal.

2. The diversity of multi-omics

From Chapter III to V, I discussed integration methods in 3 different categories for multimodal integration. We developed two different integration methods, SMILE and sciCAN, and they are aimed to address different integration difficulties. We aim to build to tools that cover single-cell multi-omics data from transcriptome to chromatin accessibility, DNA methylation and even to chromatin structure data. However, we still need more computational integration tools because of the diversity of multi-omics data. In this thesis, I primarily focused on data integration. However, analysis of each modality has its own unique challenges. Without solving what is the better computational analysis to represent each modality truthfully, integration could be distortion to each modality. One modality I did not cover extensively in this thesis is the spatial transcriptome. Different from conventional single-cell transcriptome, spatial transcriptome further incorporates spatial locations of cells in a tissue and provides extra information for us to understand regulatory landscape in situ. Meanwhile, single-cell chromatin structure data is also

standout example because it has such a distinct data format from any other modalities and has extreme data sparsity. Before I close this thesis, I would like to further acknowledge these two fields and their challenges. This includes one published work relevant to spatial transcriptome: a computational method I developed named CoSTA (Xu and McCord, 2021), and an ongoing project in which we examined impacts of feature extraction on data analysis.

2.1. Spatial transcriptome

I briefly mentioned integration of single-cell transcriptome and spatial transcriptome data in previous chapters. Here, let me elaborate more challenges of analyzing spatial data and additional effort we made. Evolving from single-cell transcriptomics, spatial transcriptomics further incorporated spatial information. This newly research domain has been attracting extensive attention from single-cell research community recently. Different spatial technologies have enabled high resolution measurements of how gene regulation is spatially organized but sacrifice balance between genome-wide transcriptome profiling and single-cell resolution (Burgess, 2019). While we consider integration of spatial transcriptomics data with other modalities, analyses and data practices for spatial transcriptome data deserve more careful consideration, in order to make full use of the extra spatial information. Thus, we need more wise data practices and analysis strategies for spatial transcriptome.

A few current analysis pipelines often treat each pixel in an expression matrix of spatial data as an independent feature, thus losing spatial information. For example, the seqFISH+ technique can fluorescently detect 10,000 mRNAs in situ at single cell resolution, and there

are often groups of cells that have correlated gene expression with their neighbors to make up larger structures. However, the original report analyzed these expression patterns using PCA and hierarchical clustering, treating each cell as an independent feature, rather than preserving spatial positions of cell neighbors (Eng et al., 2019). Slide-seq similarly produces high-throughput spatially resolved transcription information, using sequencing rather than fluorescence. Previous analyses of Slide-seq data first identified spatially non-random gene expression, but then looked for genes expressed in similar patterns using pixel-level overlap analysis rather than according to spatial features (Rodriques et al., 2019). Existing algorithms for analysis of spatial transcriptomics are based on statistical modeling and primarily propose to distinguish spatially expressing or variable genes from random spatial expression noise. For example, both SpatialDE and SPARK analysis approaches estimate how significant the spatial pattern of a gene is (Sun et al., 2020; Valentine et al., 2018). SpatialDE further builds in an unsupervised pattern detection algorithm to cluster significant SE genes into different groups which have certain spatial patterns in collective. SPARK, in contrast, was designed only for finding SE genes. To examine spatial relationships between genes, this method still relies on hierarchical clustering that uses individual pixels as features. Therefore, even though SPARK can identify genes with significant spatial patterns, the latter part of the SPARK analysis decouples the expression from its original spatial context.

Thus far, existing spatial transcriptomics analyses involve either multi-step complex feature engineering for spatial quantification or human-imposed rigid or statistical modeling-based screening of candidate SE genes. In the existing methods, the similarity of

expression pattern between two genes is either binary-- whether or not the genes cluster together-- or is quantified based on pixel-level correlation. To address spatial data analysis, I also proposed a computer-vision-inspired approach to examine relationships between spatial expression patterns of different genes while preserving the full spatial context. I adopt an unsupervised **Co**nvNet learning strategy for **S**patial **T**ranscriptomics **A**nalysis (CoSTA). This new method, named CoSTA, can find quantitative comparisons between gene expression patterns in a way that preserves spatial relationships between neighboring cells and tissue regions (Figure 6.1). Applying CoSTA to published MERFISH and Slide-seq data, we show that CoSTA identifies specific but biologically-relevant gene sets with significant spatial relationships.

2.2. 3D genome at single-cell resolution

Among all modalities, chromatin structure data stands out alone because it has distinct format from other modalities. Modalities, for example transcriptome and chromatin accessibility, represent the enrichment of a transcriptomic and epigenomic properties along the linear genome. These modalities are presented in the form of 1D information. Differently, chromatin structure data reflects how frequently one genomic locus contacts with others. A simple way to present chromatin structure data can be a 2D symmetrical contact matrix. However, this data representation can be misleading that contacts only happen between two loci. For example, we could observe contact between A and B, and contact between B and C. If we could represent chromatin structure data in a 3D form, it is possible that A, B, and C form the one contact simultaneously.

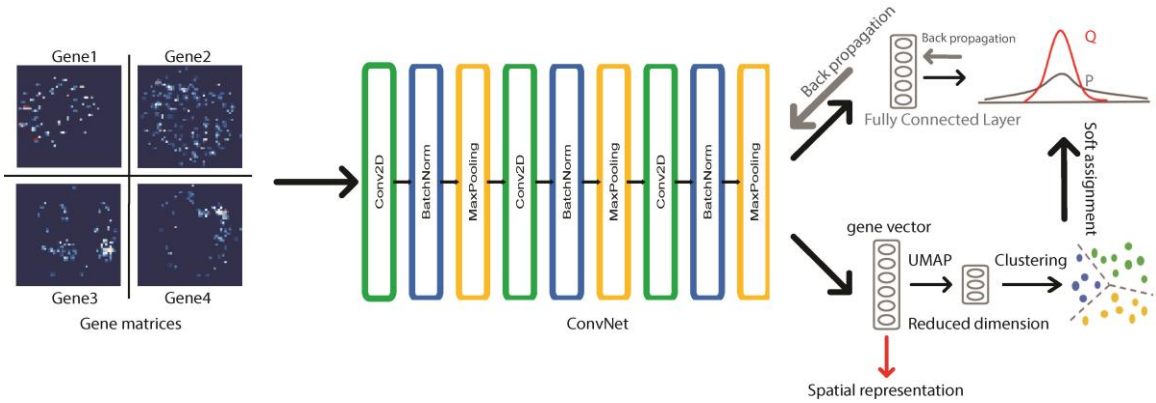


Figure 6.1. Unsupervised neural network model to learn spatial feature.

Because chromatin structure data goes beyond 1D information, there would be multiple ways of feature extraction and multiple angles of data interpretation (McCord et al., 2022). At the scale of whole genome, each chromosome may occupy a certain space within the nucleus (Cremer and Cremer, 2001). The arrangement of chromosome territories can vary among different cell types, and the cell-type specific chromosome arrangement could exhibit biological functions for purpose (Das et al., 2020; Parada et al., 2004). Within the scope of each chromosome, it is well known that chromosome is compartmentalized into active and inactive regions (Lieberman-Aiden et al., 2009). This compartmentalization is highly correlated to other linear features, like histone modification, CpG enrichment, and Lamin-associated domains (LADs) (Briand and Collas, 2020; Tan et al., 2021; van Steensel and Belmont, 2017; Xu et al., 2019). Taking advantage of this correlation could serves a link to integrate chromatin structure data with other epigenomic data. Going deeper, we can further reach to the scale to examine how chromatin is arranged within a compartment. This led to the concept of topological associated domains (TADs) (Beagan and Phillips-Cremins, 2020). Regulatory regions can be blocked by TADs and can't cast their influence on genes nearly. This local scale of chromatin arrange could endow cells with more precise gene regulation. Finally, the most refined scale would be specific contacts between two loci. Such contacts involving of enhancers and gene promoters could play a critical role in cellular programming.

However, there has not been a benchmark study to show which feature space better represent chromatin structure data so far. We have performed a preliminary examination on this issue (Figure 6.2). We found different feature spaces have different degrees of

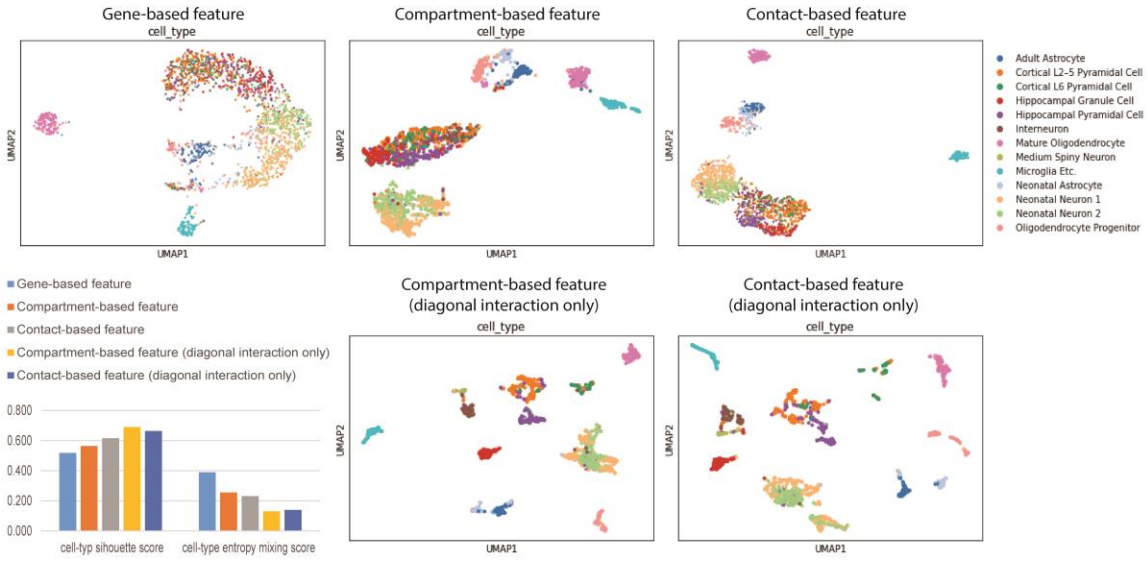


Figure 6.2. Single-cell 3D genome data analysis with 3 different levels of features. 3 different feature types are extracted from the same single-cell 3D genome data. Contact-based features is the most refined scale, gene-based feature reflects interactions at a local scale, while compartment-based feature represents a global scale. Both cell-type silhouette score and cell-type entropy mixing score measures how different cell types are separated out using these feature spaces. Higher is better for cell-type silhouette score, but lower is better for cell-type entropy mixing score.

discriminative power to reveal different cell types within a single-cell 3D genome data. Interestingly, if we only include self-interaction for data analysis, we can separate out every cell type, even for hard cases. This suggests a risk of identifying variation that does not involve long-range structural contacts, even though 3D genome data is aimed to reveal structural biological information. Moreover, we don't know if combining all features from the small to large scale of chromatin structure would be more beneficial than analyzing structure data with each feature space independently. Lack of this kind of benchmark motivated us to comprehensively examined impacts of different features on single-cell chromatin structure data analysis. Meanwhile, we aimed to address if the choice of feature from chromatin structure data would affect the integration with other single-cell modalities and how much the influence would be.

3. Conclusion

In this thesis, I presented multiple computational methods for integrating multi-source single-cell transcriptome data and multimodal single-cell data. I aimed to build a comprehensive toolbox to cover a wide range of integration applications. Building a comprehensive toolbox is not a lonely journey but a community effort. Over these years, the single-cell research community has grown into globe and built numerous tools for diverse analysis problems. To end this thesis, I would like to acknowledge that the diversity of single-cell multi-omics opens the opportunities to understand complex cellular systems from multiple levels as well as presenting grand challenges of developing suitable computational tools for precise and accurate data analysis.

REFERENCES

. The 1M Cell Evercode™ Whole Transcriptome Mega, Parse biosciences.

. 10x Datasets Single Cell Gene Expression, Official 10x Genomics Support.

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome biology *20*, 194-194.

Ahn, J.H., Kim, J., Hong, S.P., Choi, S.Y., Yang, M.J., Ju, Y.S., Kim, Y.T., Kim, H.M., Rahman, M.D.T., Chung, M.K., *et al.* (2021). Nasal ciliated cells are primary targets for SARS-CoV-2 replication in early stage of COVID-19. The Journal of clinical investigation *131*, 1-14.

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., *et al.* (2017). SCENIC: single-cell regulatory network inference and clustering. Nature methods *14*, 1083-1086.

Almanzar, N., Antony, J., Baghel, A.S., Bakerman, I., Bansal, I., Barres, B.A., Beachy, P.A., Berdnik, D., Bilen, B., Brownfield, D., *et al.* (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature *583*, 590-595.

Amid, E., and Warmuth, M.K. (2019). TriMap: Large-scale Dimensionality Reduction Using Triplets. arXiv e-prints, arXiv:1910.00204.

Arazi, A., Rao, D.A., Berthier, C.C., Davidson, A., Liu, Y., Hoover, P.J., Chicoine, A., Eisenhaure, T.M., Jonsson, A.H., Li, S., *et al.* (2019). The immune cell landscape in kidneys of patients with lupus nephritis. Nature Immunology *20*, 902-914.

Argelaguet, R., Cuomo, A.S.E., Stegle, O., and Marioni, J.C. (2021). Computational principles and challenges in single-cell data integration. Nature biotechnology.

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L.X. (2019). DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome biology *20*, 211-211.

Asano, Y.M., Rupprecht, C., and Vedaldi, A. (2019). Self-labelling via simultaneous clustering and representation learning. arXiv e-prints, arXiv:1911.05371.

Bahrami, M., Maitra, M., Nagy, C., Turecki, G., Rabiee, H.R., and Li, Y. (2020). Deep feature extraction of single-cell transcriptomes by generative adversarial network. Bioinformatics (Oxford, England).

Bandler, R.C., Vitali, I., Delgado, R.N., Ho, M.C., Dvoretskova, E., Ibarra Molinas, J.S., Frazel, P.W., Mohammadkhani, M., Machold, R., Maedler, S., *et al.* (2021). Single-cell delineation of lineage and genetic identity in the mouse brain. Nature.

Baron, M., Veres, A., Wolock, Samuel L., Faust, Aubrey L., Gaujoux, R., Vetere, A., Ryu, Jennifer H., Wagner, Bridget K., Shen-Orr, Shai S., Klein, Allon M., *et al.* (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell systems *3*, 346-360.e344.

Beagan, J.A., and Phillips-Cremins, J.E. (2020). On the existence and functionality of topologically associating domains. Nature Genetics *52*, 8-16.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology *37*, 38-44.

Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nature Biotechnology *38*, 1408-1414.

Bergen, V., Soldatov, R.A., Kharchenko, P.V., and Theis, F.J. (2021). RNA velocity—current challenges and future perspectives. Molecular Systems Biology *17*, e10282.

Bernstein, N.J., Fong, N.L., Lam, I., Roy, M.A., Hendrickson, D.G., and Kelley, D.R. (2020). Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. Cell Syst *11*, 95-101 e105.

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*, P10008-10012.

Briand, N., and Collas, P. (2020). Lamina-associated domains: peripheral matters and internal affairs. Genome Biology *21*, 85.

Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., Majeti, R., Chang, H.Y., and Greenleaf, W.J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell *173*, 1535-1548.e1516.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature (London) *523*, 486-490.

Burgess, D.J. (2019). Spatial transcriptomics coming of age. Nature reviews Genetics *20*, 317.

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*, 411-420.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., *et al.* (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science (American Association for the Advancement of Science) *361*, 1380-1385.

Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., *et al.* (2020a). A human cell atlas of fetal gene expression. Science (American Association for the Advancement of Science) *370*.

Cao, K., Bai, X., Hong, Y., and Wan, L. (2020b). Unsupervised topological alignment for single-cell multi-omics integration. Bioinformatics (Oxford, England) *36*, i48-i56.

Cao, K., Hong, Y., and Wan, L. (2021). Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. Bioinformatics.

Cao, Y., Wang, X., and Peng, G. (2020c). SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. Frontiers in genetics *11*, 490-490.

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep Clustering for Unsupervised Learning of Visual Features.

Carter, B., and Zhao, K. (2021). The epigenetic basis of cellular heterogeneity. Nature reviews Genetics *22*, 235-250.

Chan Zuckerberg Initiative Single-Cell, C.-C., Ballestar, E., Farber, D.L., Glover, S., Horwitz, B., Meyer, K., Nikolić, M., Ordovas-Montanes, J., Sims, P., Shalek, A., *et al.*

(2020). Single cell profiling of COVID-19 patients: an international data resource from multiple tissues. medRxiv, 2020.2011.2020.20227355.

Chen, S., Lake, B.B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature biotechnology *37*, 1452-1457.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A Simple Framework for Contrastive Learning of Visual Representations, pp. arXiv:2002.05709.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A Simple Framework for Contrastive Learning of Visual Representations. arXiv e-prints, arXiv:2002.05709.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. (2020c). Big Self-Supervised Models are Strong Semi-Supervised Learners. arXiv e-prints, arXiv:2006.10029.

Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M.T., *et al.* (2020). COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. Nature Biotechnology *38*, 970-979.

Ciortan, M., and Defrance, M. (2021). Contrastive self-supervised clustering of scRNA-seq data. BMC Bioinformatics *22*, 280.

Collombet, S., Ranisavljevic, N., Nagano, T., Varnai, C., Shisode, T., Leung, W., Piolot, T., Galupa, R., Borensztein, M., Servant, N., *et al.* (2020). Parental-to-embryo switch of chromosome organization in early embryogenesis. Nature (London) *580*, 142-146.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., *et al.* (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature Genetics *48*, 1193-1203.

Cortal, A., Martignetti, L., Six, E., and Rausell, A. (2021). Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. Nature Biotechnology.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature Reviews Genetics *2*, 292-301.

Cui, Y., Zheng, Y., Liu, X., Yan, L., Fan, X., Yong, J., Hu, Y., Dong, J., Li, Q., Wu, X., *et al.* (2019a). Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. Cell reports (Cambridge) *26*, 1934-1950.e1935.

Cui, Y., Zheng, Y., Liu, X., Yan, L., Fan, X., Yong, J., Hu, Y., Dong, J., Li, Q., Wu, X., *et al.* (2019b). Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. Cell Reports *26*, 1934-1950.e1935.

Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science (American Association for the Advancement of Science) *348*, 910-914.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., *et al.* (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell (Cambridge) *174*, 1309-1324.e1318.

Dai, M., Pei, X., and Wang, X.-J. (2022). Accurate and fast cell marker gene identification with COSG. Briefings in Bioinformatics, bbab579.

Das, P., Shen, T., and McCord, R.P. (2020). Inferring chromosome radial organization from Hi-C data. BMC Bioinformatics *21*, 511.

Demetci, P., Santorella, R., Sandstede, B., Noble, W.S., and Singh, R. (2020). Gromov-Wasserstein optimal transport to align single-cell multi-omics data. bioRxiv, 2020.2004.2028.066787.

Dincer, A.B., Janizek, J.D., and Lee, S.-I. (2020). Adversarial deconfounding autoencoder for learning robust gene expression embeddings. Bioinformatics (Oxford, England) *36*, i573-i582.

Ding, J., Adiconis, X., Simmons, S.K., Kowalczyk, M.S., Hession, C.C., Marjanovic, N.D., Hughes, T.K., Wadsworth, M.H., Burks, T., Nguyen, L.T., *et al.* (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nature biotechnology *38*, 737-746.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., *et al.* (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853-1866.e1817.

Duren, Z., Lu, W.S., Arthur, J.G., Shah, P., Xin, J., Meschi, F., Li, M.L., Nemec, C.M., Yin, Y., and Wong, W.H. (2021). Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. Nature Communications *12*, 4763.

Efremova, M., and Teichmann, S.A. (2020). Computational methods for single-cell omics across modalities. Nature Methods *17*, 14-17.

Efroni, I., Ip, P.-L., Nawy, T., Mello, A., and Birnbaum, K.D. (2015). Quantification of cell identity from single-cell gene expression profiles. Genome biology *16*, 9-9.

Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., *et al.* (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature *568*, 235.

Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., and Theis, F.J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nature communications *10*, 390-390.

Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X., Xie, F., *et al.* (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nature communications *12*, 1337-1337.

Farrell Jeffrey, A., Wang, Y., Riesenfeld Samantha, J., Shekhar, K., Regev, A., and Schier Alexander, F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science *360*, eaar3131.

Forcato, M., Romano, O., and Bicciato, S. (2021). Computational methods for the integrative analysis of single-cell data. Briefings in bioinformatics *22*, 20-29.

Franzén, O., Gan, L.-M., and Björkegren, J.L.M. (2019a). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database *2019*, baz046.

Franzén, O., Gan, L.-M., and Björkegren, J.L.M. (2019b). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database : the journal of biological databases and curation *2019*.

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 [version 2; peer review: 3 approved]. F1000 research *7*, 1297-1297.

Fudenberg, G., Kelley, D.R., and Pollard, K.S. (2020). Predicting 3D genome folding from DNA sequence with Akita. Nature Methods *17*, 1111-1117.

Gierahn, T.M., Wadsworth, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nature Methods *14*, 395-398.

Goetz, J.J., and Trimarchi, J.M. (2012). Transcriptome sequencing of single cells with Smart-Seq. Nature Biotechnology *30*, 763-765.

Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature Genetics *53*, 403-411.

Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., *et al.* (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nature Biotechnology *37*, 1458-1465.

Grün, D., Muraro, Mauro J., Boisset, J.-C., Wiebrands, K., Lyubimova, A., Dharmadhikari, G., van den Born, M., van Es, J., Jansen, E., Clevers, H., *et al.* (2016). De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell stem cell *19*, 266-277.

Guan, J., Wang, G., Wang, J., Zhang, Z., Fu, Y., Cheng, L., Meng, G., Lyu, Y., Zhu, J., Li, Y., *et al.* (2022). Chemical reprogramming of human somatic cells to pluripotent stem cells. Nature.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology *36*, 421-427.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., *et al.* (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091-1107.e1017.

Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B.J., Bader, G.D., Barker, R.A., Camara, P.G., Camp, J.G., Chédotal, A., Copp, A., *et al.* (2021). A roadmap for the Human Developmental Cell Atlas. Nature *597*, 196-205.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Reports *2*, 666-673.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nature biotechnology *37*, 685-691.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome research *21*, 1160-1167.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLOS ONE *9*, e98679.

Jain, M.S., Conde, C.D., Polanski, K., Chen, X., Park, J., Botting, R.A., Stephenson, E., Haniffa, M., Lamacraft, A., Efremova, M., *et al.* (2021). MultiMAP: Dimensionality Reduction and Integration of Multimodal Data. bioRxiv, 2021.2002.2016.431421.

Jaitin Diego, A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., *et al.* (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science *343*, 776-779.

Jin, S., Zhang, L., and Nie, Q. (2020). scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. Genome Biology *21*, 25.

Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., *et al.* (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology *36*, 89-94.

Kelsey, G., Stegle, O., and Reik, W. (2017). Single-cell epigenomics: Recording the past and predicting the future. Science (American Association for the Advancement of Science) *358*, 69-75.

Kim, H.J., Wang, K., Chen, C., Lin, Y., Tam, P.P.L., Lin, D.M., Yang, J.Y.H., and Yang, P. (2021). Uncovering cell identity through differential stability with Cepo. Nature Computational Science *1*, 784-790.

Kimmel, J.C., and Kelley, D.R. (2021). Semi-supervised adversarial neural networks for single-cell classification. Genome research.

Klein, Allon M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, David A., and Kirschner, Marc W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell *161*, 1187-1201.

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nature communications *10*, 5416-5416.

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics *28*, 573-580.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nature methods *16*, 1289-1296.

Kuppe, C., Ibrahim, M.M., Kranz, J., Zhang, X., Ziegler, S., Perales-Patón, J., Jansen, J., Reimer, K.C., Smith, J.R., Dobie, R., *et al.* (2021). Decoding myofibroblast origins in human kidney fibrosis. Nature *589*, 281-286.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., *et al.* (2018). RNA velocity of single cells. Nature *560*, 494-498.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., *et al.* (2020). Eleven grand challenges in single-cell data science. Genome biology *21*, 31-35.

Lakkis, J., Wang, D., Zhang, Y., Hu, G., Wang, K., Pan, H., Ungar, L., Reilly, M.P., Li, X., and Li, M. (2021). A joint deep learning model enables simultaneous batch effect

correction, denoising, and clustering in single-cell transcriptomics. Genome Research *31*, 1753-1766.

Lawlor, N., George, J., Bolisetty, M., Kursawe, R., Sun, L., Sivakamasundari, V., Kycia, I., Robson, P., and Stitzel, M.L. (2017). Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. Genome research *27*, 208-222.

Lee, D.-S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J.R., Fitzpatrick, C., O'Connor, C., Dixon, J.R., *et al.* (2019). Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nature methods *16*, 999-1006.

Lee, I.T., Nakayama, T., Wu, C.-T., Goltsev, Y., Jiang, S., Gall, P.A., Liao, C.-K., Shih, L.-C., Schürch, C.M., McIlwain, D.R., *et al.* (2020). ACE2 localizes to the respiratory cilia and is not increased by ACE inhibitors or ARBs. Nature Communications *11*, 5453.

Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M., and Ren, B. (2019). Joint profiling of DNA methylation and chromatin architecture in single cells. Nature methods *16*, 991-993.

Li, H.-S., Ou-Yang, L., Zhu, Y., Yan, H., and Zhang, X.-F. (2021). scDEA: differential expression analysis in single-cell RNA-sequencing data via ensemble learning. Briefings in Bioinformatics, bbab402.

Li, H., Janssens, J., De Waegeneer, M., Kolluru Sai, S., Davie, K., Gardeux, V., Saelens, W., David Fabrice, P.A., Brbić, M., Spanier, K., *et al.* Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. Science *375*, eabk2432.

Li, Y., Hu, P., Liu, Z., Peng, D., Tianyi Zhou, J., and Peng, X. (2020). Contrastive Clustering. arXiv e-prints, arXiv:2009.09687.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY) *326*, 289-293.

Lin, Y., Wu, T.-Y., Wan, S., Yang, J.Y.H., Wong, W.H., and Wang, Y.X.R. (2021). scJoint: transfer learning for data integration of single-cell RNA-seq and ATAC-seq. bioRxiv, 2020.2012.2031.424916.

Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C.L., Lindberg, E.L., Kanda, M., Polanski, K., Heinig, M., Lee, M., *et al.* (2020). Cells of the adult human heart. Nature.

Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J.K., Nery, J.R., Chen, H., *et al.* (2021). DNA methylation atlas of the mouse brain at single-cell resolution. Nature *598*, 120-128.

Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E.Z., and Welch, J.D. (2020). Jointly defining cell types from multiple single-cell datasets using LIGER. Nature protocols *15*, 3632-3662.

Liu, J., Huang, Y., Singh, R., Vert, J.-P., and Noble, W.S. (2019). Jointly embedding multiple single-cell omics measurements. bioRxiv, 644310.

Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S., Chittenden, T.W., D'Gama, A.M., Cai, X., *et al.* (2015). Somatic mutation in single human

neurons tracks developmental and transcriptional history. Science (American Association for the Advancement of Science) *350*, 94-98.

Longo, S.K., Guo, M.G., Ji, A.L., and Khavari, P.A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nature Reviews Genetics.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nature methods *15*, 1053-1058.

Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., *et al.* (2021). Mapping single-cell data to reference atlases by transfer learning. Nature Biotechnology.

Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nature methods *16*, 715-721.

Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020a). Integrative Methods and Practical Challenges for Single-Cell Multi-omics. Trends in biotechnology (Regular ed) *38*, 1007-1022.

Ma, F., and Pellegrini, M. (2020). ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics (Oxford, England) *36*, 533-538.

Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., *et al.* (2020b). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell (Cambridge) *183*, 1103-1116.e1120.

Ma, W., Su, K., and Wu, H. (2021). Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. Genome Biology *22*, 264.

Macaulay, I.C., Ponting, C.P., and Voet, T. (2017). Single-Cell Multiomics: Multiple Measurements from Single Cells. Trends in genetics *33*, 155-168.

Macosko, Evan Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, Allison R., Kamitaki, N., Martersteck, Emily M., *et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

Mancarci, B.O., Toker, L., Tripathy, S.J., Li, B., Rocco, B., Sibille, E., and Pavlidis, P. (2017). Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data. eNeuro *4*, ENEURO.0212-0217.2017.

McCord, R.P., Xu, Y., Li, H., Das, P., and San Martin, R. (2022). SnapShot: Chromosome organization. Molecular Cell *82*, 2350-2350.e2351.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv e-prints, arXiv:1802.03426.

Meng, C., Kuster, B., Culhane, A.C., and Gholami, A.M. (2014). A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics *15*, 162.

Miao, Z., Balzer, M.S., Ma, Z., Liu, H., Wu, J., Shrestha, R., Aranyi, T., Kwan, A., Kondo, A., Pontoglio, M., *et al.* (2021). Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. Nature Communications *12*, 2277.

Miao, Z., Moreno, P., Huang, N., Papatheodorou, I., Brazma, A., and Teichmann, S.A. (2020). Putative cell type discovery from single-cell gene expression data. Nature Methods *17*, 621-628.

Mou, T., Deng, W., Gu, F., Pawitan, Y., and Vu, T.N. (2020). Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. Frontiers in Genetics *10*, 1331.

Muraro, Mauro J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, Marten A., Carlotti, F., de Koning, Eelco J.P., *et al.* (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell systems *3*, 385-394.e383.

Oetjen, K.A., Lindblad, K.E., Goswami, M., Gui, G., Dagur, P.K., Lai, C., Dillon, L.W., McCoy, J.P., and Hourigan, C.S. (2018). Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. JCI insight *3*.

Parada, L.A., McQueen, P.G., and Misteli, T. (2004). Tissue-specific spatial organization of genomes. Genome Biology *5*, R44.

Peng, T., Chen, G.M., and Tan, K. (2021). GLUER: integrative analysis of single-cell omics and imaging data by deep neural network. bioRxiv, 2021.2001.2025.427845.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature Methods *10*, 1096-1098.

Pierce, S.E., Granja, J.M., and Greenleaf, W.J. (2021). High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. Nature Communications *12*, 2969.

Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., *et al.* (2018). Cicero

Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular cell *71*, 858-871.e858.

Pliner, H.A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. Nature Methods *16*, 983-986.

Polański, K., Young, M.D., Miao, Z., Meyer, K.B., Teichmann, S.A., and Park, J.-E. (2020). BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics *36*, 964-965.

Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. Nature *553*, 212-216.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science (New York, NY) *363*, 1463.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., *et al.* (2018a). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science (American Association for the Advancement of Science) *360*, 176-182.

Rosenberg, A.B.A.B., Roco, C.M.C.M., Muscat, R.A.R.A., Kuchina, A.A., Sample, P.P., Yao, Z.Z., Gray, L.L., Peeler, D.J.D.J., Mukherjee, S.S., Chen, W.W., *et al.* (2018b). SPLiT-seq reveals cell types and lineages in the developing brain and spinal cord. Science (American Association for the Advancement of Science) *360*, 176-182.

Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. (2017). The Human Cell Atlas: from vision to reality. Nature *550*, 451-453.

Rubin, A.J., Parker, K.R., Satpathy, A.T., Qi, Y., Wu, B., Ong, A.J., Mumbach, M.R., Ji, A.L., Kim, D.S., Cho, S.W., *et al.* (2019). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. Cell *176*, 361-376.e317.

Saunders, A., Macosko, E.Z., Wysoker, A., Goldman, M., Krienen, F.M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., *et al.* (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell *174*, 1015-1030.e1016.

Schaum, N., Neff, N.F., May, A.P., Quake, S.R., Darmanis, S., Batson, J., Chen, M.B., Chen, S., Green, F., Penland, L., *et al.* (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature (London) *562*, 367-372.

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., *et al.* (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell metabolism *24*, 593-607.

Shaham, U., Stanton, K.P., Zhao, J., Li, H., Raddassi, K., Montgomery, R., and Kluger, Y. (2017). Removal of batch effects using distribution-matching residual networks. Bioinformatics *33*, 2539-2546.

Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., and Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience *23*, 100882.

Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods *15*, 255-261.

Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., *et al.* (2021). Confronting false discoveries in single-cell differential expression. Nature Communications *12*, 5692.

Stark, S.G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., Tumor Profiler, C., Rätsch, G., and Lehmann, K.-V. (2020). SCIM: universal single-cell matching with unpaired feature sets. Bioinformatics *36*, i919-i927.

Stewart Benjamin, J., Ferdinand John, R., Young Matthew, D., Mitchell Thomas, J., Loudon Kevin, W., Riding Alexandra, M., Richoz, N., Frazer Gordon, L., Staniforth Joy, U.L., Vieira Braga Felipe, A., *et al.* (2019). Spatiotemporal immune zonation of the human kidney. Science *365*, 1461-1466.

Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nature Biotechnology *39*, 313-319.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e1821.

Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. Nature reviews Genetics *20*, 257-272.

Stuart, T., Srivastava, A., Lareau, C., and Satija, R. (2020). Multimodal single-cell chromatin analysis with Signac. bioRxiv, 2020.2011.2009.373613.

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature Methods *17*, 193-200.

Sun, Z., Chen, L., Xin, H., Jiang, Y., Huang, Q., Cillo, A.R., Tabib, T., Kolls, J.K., Bruno, T.C., Lafyatis, R., *et al.* (2019). A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. Nature communications *10*, 1649-1649.

Tan, L., Ma, W., Wu, H., Zheng, Y., Xing, D., Chen, R., Li, X., Daley, N., Deisseroth, K., and Xie, X.S. (2021). Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. Cell *184*, 741-758.e717.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., *et al.* (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods *6*, 377-382.

Tedesco, M., Giannese, F., Lazarević, D., Giansanti, V., Rosano, D., Monzani, S., Catalano, I., Grassi, E., Zanella, E.R., Botrugno, O.A., *et al.* (2021). Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. Nature Biotechnology.

Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T.S., Seidi, A., Jabbari, J.S., *et al.* (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nature methods *16*, 479-487.

Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome biology *21*, 12-12.

Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., *et al.* (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619-625.

Tucker, N.R., Chaffin, M., Fleming, S.J., Hall, A.W., Parsons, V.A., Bedi, J.K.C., Akkad, A.-D., Herndon, C.N., Arduini, A., Papangeli, I., *et al.* (2020a). Transcriptional and Cellular Diversity of the Human Heart. Circulation (New York, NY).

Tucker, N.R., Chaffin, M., Fleming, S.J., Hall, A.W., Parsons, V.A., Bedi, K.C., Akkad, A.-D., Herndon, C.N., Arduini, A., Papangeli, I., *et al.* (2020b). Transcriptional and Cellular Diversity of the Human Heart. Circulation (New York, NY) *142*, 466-482.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial Discriminative Domain Adaptation. arXiv e-prints, arXiv:1702.05464.

Valentine, S., Sarah, A.T., and Oliver, S. (2018). SpatialDE: identification of spatially variable genes. Nature Methods *15*.

van Steensel, B., and Belmont, A.S. (2017). Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell *169*, 780-791.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., *et al.* (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nature Cell Biology *19*, 271-281.

Vieira Braga, F.A., Kar, G., Berg, M., Carpaij, O.A., Polanski, K., Simon, L.M., Brouwer, S., Gomes, T., Hesse, L., Jiang, J., *et al.* (2019). A cellular census of human lungs identifies novel cell states in health and in asthma. Nature medicine *25*, 1153-1163.

Wagner Daniel, E., Weinreb, C., Collins Zach, M., Briggs James, A., Megason Sean, G., and Klein Allon, M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science *360*, 981-987.

Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. Nature reviews Genetics *21*, 410-427.

Wang, A., Chiou, J., Poirion, O.B., Buchanan, J., Valdez, M.J., Verheyden, J.M., Hou, X., Kudtarkar, P., Narendra, S., Newsome, J.M., *et al.* (2020a). Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes. eLife *9*, e62522.

Wang, C., Sun, D., Huang, X., Wan, C., Li, Z., Han, Y., Qin, Q., Fan, J., Qiu, X., Xie, Y., *et al.* (2020b). Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome biology *21*, 1-198.

Wang, L., Yu, P., Zhou, B., Song, J., Li, Z., Zhang, M., Guo, G., Wang, Y., Chen, X., Han, L., *et al.* (2020c). Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. Nature cell biology *22*, 108-119.

Wang, T., Johnson, T.S., Shao, W., Lu, Z., Helm, B.R., Zhang, J., and Huang, K. (2019). BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome biology *20*, 165-165.

Wang, X., Wang, J., Zhang, H., Huang, S., and Yin, Y. (2021a). HDMC: a novel deep learning-based framework for removing batch effects in single-cell RNA-seq data. Bioinformatics, btab821.

Wang, Z., Xie, L., Ding, G., Song, S., Chen, L., Li, G., Xia, M., Han, D., Zheng, Y., Liu, J., *et al.* (2021b). Single-cell RNA sequencing of peripheral blood mononuclear cells from acute Kawasaki disease patients. Nature Communications *12*, 5444.

Welch, J.D., Hartemink, A.J., and Prins, J.F. (2017). MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. Genome Biology *18*, 138.

Wilson, P.C., Wu, H., Kirita, Y., Uchimura, K., Ledru, N., Rennke, H.G., Welling, P.A., Waikar, S.S., and Humphreys, B.D. (2019). The single-cell transcriptomic landscape of early human diabetic nephropathy. Proceedings of the National Academy of Sciences *116*, 19619.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome biology *19*, 15-15.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biology *20*, 59.

Wu, H., Malone, A.F., Donnelly, E.L., Kirita, Y., Uchimura, K., Ramakrishnan, S.M., Gaut, J.P., and Humphreys, B.D. (2018a). Single-Cell Transcriptomics of a Human Kidney

Allograft Biopsy Specimen Defines a Diverse Inflammatory Response. Journal of the American Society of Nephrology *29*, 2069.

Wu, K.E., Yost, K.E., Chang, H.Y., and Zou, J. (2021). BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. Proceedings of the National Academy of Sciences *118*, e2023070118.

Wu, Z., Xiong, Y., Yu, S., and Lin, D. (2018b). Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination, pp. arXiv:1805.01978.

Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy of Sciences *116*, 19490.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021a). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Molecular Systems Biology *17*, e9620.

Xu, Y., Baumgart, S.J., Stegmann, C.M., and Hayat, S. (2021b). MACA: Marker-based automatic cell-type annotation for single cell expression data. Bioinformatics, btab840.

Xu, Y., Das, P., and McCord, R.P. (2021c). SMILE: Mutual Information Learning for Integration of Single Cell Omics Data. bioRxiv, 2021.2001.2028.428619.

Xu, Y., Das, P., and McCord, R.P. (2022a). SMILE: mutual information learning for integration of single-cell omics data. Bioinformatics *38*, 476-486.

Xu, Y., Kramann, R., McCord, R.P., and Hayat, S. (2022b). Fast model-free standardization and integration of single-cell transcriptomics data. bioRxiv, 2022.2003.2028.486110.

Xu, Y., and McCord, R.P. (2021). CoSTA: unsupervised convolutional neural network learning for spatial transcriptomics analysis. BMC Bioinformatics *22*, 397.

Xu, Y., Shen, T., and McCord, R.P. (2019). 3D Genome Structure Variation Across Cell Types Captured by Integrating Multi-omics. bioRxiv, 784223.

Yang, K.D., Belyaeva, A., Venkatachalapathy, S., Damodaran, K., Katcoff, A., Radhakrishnan, A., Shivashankar, G.V., and Uhler, C. (2021). Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nature Communications *12*, 31.

Yang, Z., and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics (Oxford, England) *32*, 1-8.

Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R.S., Aldridge, A.I., Ament, S.A., Bartlett, A., Behrens, M.M., Van den Berge, K., *et al.* (2021). A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature *598*, 103-110.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., *et al.* (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (American Association for the Advancement of Science) *347*, 1138-1142.

Zhang, A.W., O'Flanagan, C., Chavez, E.A., Lim, J.L.P., Ceglia, N., McPherson, A., Wiens, M., Walters, P., Chan, T., Hewitson, B., *et al.* (2019a). Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nature methods *16*, 1007-1015.

Zhang, J.-Y., Wang, X.-M., Xing, X., Xu, Z., Zhang, C., Song, J.-W., Fan, X., Xia, P., Fu, J.-L., Wang, S.-Y., *et al.* (2020). Single-cell landscape of immunological responses in patients with COVID-19. Nature Immunology *21*, 1107-1118.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., *et al.* (2019b). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic acids research *47*, D721-D728.

Zhang, Z., Luo, D., Zhong, X., Choi, J.H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E.W., Modrusan, Z., *et al.* (2019c). SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. Genes *10*, 531.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., *et al.* (2017). Massively parallel digital transcriptional profiling of single cells. Nature communications *8*, 14049-14049.

Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T.J., Dixon, J.R., and Ecker, J.R. (2019). Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. Proceedings of the National Academy of Sciences - PNAS *116*, 14011-14018.

Zhu, C., Zhang, Y., Li, Y.E., Lucero, J., Behrens, M.M., and Ren, B. (2021). Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Nature Methods.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A.A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv e-prints, arXiv:1703.10593.

VITA

Yang Xu was born in Chongqing, China on November 4th, 1992, and spent his childhood in the rural area in Chongqing. After finishing high school in Chongqing, he went to Wuhan for university. He earned a B.Sc. in Biotechnology at Huazhong Agricultural University in 2015. In 2017, Yang joined the UT-ORNL Graduate School of Genome Science and Technology at the University of Tennessee, Knoxville. He received a Ph.D. in 2022. His research focuses on 1) unsupervised representation learning for biomedical data, 2) deep learning approach to integrate multimodal single-cell data, and 3) marker or knowledge-based approach to annotate single-cell data. Yang will begin working as Machine Learning Scientist at the Broad Institute of MIT and Harvard in July 2022.