Statistical and Computational Methods for Single Cell and Spatial Transcriptomics

by

Jingyue Xi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2023

Doctoral Committee:

Professor Hyun Min Kang, Chair
Associate Professor Kelly Kidwell
Professor Jun Hee Lee
Research Professor Laura Scott
Professor Xiang Zhou

Jingyue Xi

jyxi@umich.edu

ORCID iD: 0000-0003-3216-1776

For Yawei, Hanmei, and Bangqi

ACKNOWLEDGEMENTS

These projects would not have been possible without the support of many people.

First of all, I would like to express my deepest gratitude to my advisor Professor Hyun Min Kang for his exceptional guidance, mentorship, encouragement and continuous support. His insight, immense knowledge, rich experience and passion for research guided me through my Ph.D. studies. I am always grateful for his understanding and tremendous support when I was going through tough medical challenges.

I am fortunate to have my thesis committee. I would like to thank Dr. Jun Hee Lee for his expertise in molecular physiology and his guidance in biological understanding has greatly helped this thesis. The projects would not have been possible without his guidance. Dr. Laura Scott for all the valuable feedback in committee meetings. I am inspired from her dedication to science and attention to detail. Dr. Xiang Zhou for his expertise and insightful comments on my research. Dr. Kelly Kidwell for all the support in committee meetings and wonderful feedback on thesis structure.

I have been very fortunate to have wonderful friends for both physical and mental support during my Ph.D. studies. In no order of preference, I would like to thank: Zhi Li for passing me her sense of humor; Han Fu, for encouraging me when I was recovering from medical surgery; Yichen Si for her generous suggestions on my research work. Ketian Yu for taking care of my fur babies. I will always cherish the wonderful memories we shared and the invaluable help they gave me.

Finally, I would like to say thank you to my parents whose constant love and support keep me motivated and confident. To my fur babies Daodao, Momo, Yoyo,

Umi and Pi, your presence has enriched my existence with boundless happiness. Last but not least, Yawei, thank you for always being there for me, whenever and wherever.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES x

LIST OF ABBREVIATIONS xi

ABSTRACT xiv

CHAPTER

I. Introduction 1

1.1 Background 1
1.2 Importance of Upstream Quality Control in Single-cell RNA Sequencing 3
1.3 Limited Resolution of Existing Spatial Transcriptomics Technology 4
1.4 Need for Software Pipelines to Handle Ultra-high-resolution Spatial Transcriptomics Data 5
1.5 Future Directions 6

II. *SiftCell*: A Robust Framework to Identify and Filter Cell-free and Cell-containing Droplets from Single-cell RNA Sequence Reads 7

2.1 Background 7
2.1.1 Droplet Barcoding and Ambient RNAs 7
2.1.2 Literature Review on Droplet Filtering Methods . . 9
2.1.3 *SiftCell* Framework 10
2.2 Results 12

2.2.1 *SiftCell-Shuffle* Visually Distinguishes Cell-free Droplets from Cell-containing Ones . . . 12
2.2.2 Evaluating the Performance of Droplet Filtering Using *SiftCell-Shuffle* . . . 17
2.2.3 *SiftCell-Boost* Robustly Filters Cell-containing and Healthy Droplets . . . 20
2.2.4 *SiftCell-Mix* Estimates the Contribution from Ambient RNAs in Each Droplet . . . 26
2.2.5 Evaluation of Computational Cost . . . 29
2.3 Materials and Methods . . . 29
2.3.1 *SiftCell-Shuffle*: Visualizing Cell-free and Cell-containing Droplets in a Manifold Space . . . 29
2.3.2 Evaluation of Existing Methods for Filtering Cell-containing Droplets with *SiftCell-Shuffle* . . . 32
2.3.3 *SiftCell-Boost*: Automated Machine Learning Method to Identify Cell-containing Droplets . . . 33
2.3.4 *SiftCell-Mix*: Model-based Approach for Inferring the Fraction of Ambient RNAs in Each Droplet . . . 38
2.4 Summary . . . 40
2.5 Appendix: Sparse Quantile Aggregation Test (SQuAT) . . . 40

III. Microscopic Examination of Spatial Transcriptome Using Seq-Scope . . . 46

3.1 Background . . . 46
3.2 Results . . . 47
3.2.1 Seq-Scope Has an Outstanding Transcriptome Capture Performance . . . 48
3.2.2 Seq-Scope Captures Transcriptome Information with High Efficiency . . . 48
3.2.3 Seq-Scope Reveals Nuclear-cytoplasmic Transcriptome Architecture from Tissue Sections . . . 50
3.2.4 Seq-Scope Performs Spatial Single-cell Analysis of Hepatocytes . . . 52
3.2.5 Seq-Scope Detects Non-parenchymal Cell Transcriptome from Liver Section . . . 54
3.2.6 Seq-Scope Visualizes Histological Layers of Colonic Wall . . . 57
3.2.7 Seq-Scope Identifies Individual Cellular Components from Colon Tissue . . . 60
3.2.8 Seq-Scope Performs Microscopic Analysis of Colonic Spatial Transcriptome . . . 62
3.3 Materials and Methods . . . 62
3.3.1 Seq-Scope Technology . . . 62
3.3.2 Tissue Boundary Estimation . . . 67

3.3.3 Read Alignment and Generation of Digital Gene Expression Matrix . . . 67
3.3.4 Error Correction Methods for Spatial Barcodes . . . 69
3.3.5 Analysis of Spliced and Unspliced Gene Expression 69
3.3.6 H&E Based Image Segmentation for Spatial Single Cell Analysis . . . 70
3.3.7 Simple Aggregation . . . 70
3.3.8 Clustering Analysis . . . 71
3.3.9 Analysis of Transcripts Discovered Outside of Tissue-overlaid Region . . . 71
3.3.10 Multi-scale Sliding Window Analysis . . . 72
3.3.11 Visualization of Spatial Gene Expression . . . 73
3.3.12 Benchmark Analysis . . . 75
3.4 Summary . . . 76

IV. STtools: Comprehensive Software Pipeline for Ultra-high-resolution Spatial Transcriptomics Data . . . 77

4.1 Background . . . 77
4.2 Results . . . 78
4.2.1 STtools Enables High Resolution Cell Type Mapping 78
4.2.2 STtools Visualizes Spatial Gene Expression at Various Scales . . . 80
4.2.3 STtools Can Efficiently Process Spatial Transcriptomic Data Scaling with Millions of Spatially Resolved Barcodes . . . 82
4.3 Materials and Methods . . . 85
4.3.1 Alignment . . . 85
4.3.2 Two-track Approach for Clustering . . . 87
4.3.3 Visualization . . . 89
4.4 Summary . . . 89

V. Discussion . . . 90

5.1 Summary . . . 90
5.2 Upstream Quality Control in Single-cell RNA Sequencing with Ambient RNAs . . . 90
5.3 Spatial Transcriptomics Technique with High Resolution . . . 93
5.4 Tools for High Resolution Spatial Transcriptomics . . . 93

BIBLIOGRAPHY . . . 95

LIST OF FIGURES

Figure

2.1	Droplet barcoding and ambinet RNA contamination	8
2.2	Overview of *SiftCell* Framework	11
2.3	Visualization of *SiftCell-Shuffle* results in t-SNE manifold space	14
2.4	Feature plot of cell type-specific marker genes in PBMC dataset	15
2.5	Feature plot of cell type-specific marker genes in colon cell line mixture dataset	16
2.6	Evaluation of droplet filtering methods with *SiftCell-Shuffle*	18
2.7	%NN-concordance evaluation plot	19
2.8	Knee plots across PBMC, brain nuclei and colon cell line mixture datasets	20
2.9	Venn Diagram of number of droplets that are classified as cell-containing or cell-free across all the methods	23
2.10	Annotation of comparison of cell filtering methods by *SiftCell-Shuffle*	24
2.11	Visualization of contribution of ambient RNAs from scRNA-seq and snRNA-seq datasets	27
2.12	Annotated visualization of contribution of ambient RNAs from scRNA-seq and snRNA-seq datasets	28
2.13	QQ plots of overdispersion tests on shuffled scRNA-seq data	35
2.14	Evaluation of overdispersion tests on the original and shuffled PBMC scRNA-seq dataset	36
3.1	Schematic diagram depicting tile arrangement in MiSeq regular flow cell	48
3.2	Seq-Scope capture performance	49
3.3	Benchmark analysis	51
3.4	Seq-Scope visualizes subcellular spatial transcriptome	53
3.5	Seq-Scope performs spatial single-cell analysis in normal mouse liver	55
3.6	Spatial expressions of individual genes	56
3.7	Normal liver Seq-Scope dataset analyzed by data binning with $10\mu m$-sided square grids	58
3.8	Detection of non-parenchymal (NPC) transcriptome through histology-agnostic segmentation with $10\mu m$ grids	59
3.9	Seq-Scope identifies various cell types from colonic wall histology	61

3.10 Seq-Scope enables microscopic analysis of colon spatial transcriptome 63
3.11 Seq-Scope overview 65
3.12 Seq-Scope data structure 66
3.13 HDMI discovery plot for Seq-Scope liver and colon data 67
3.14 Saturation analysis of liver and colon 69
3.15 Schematic diagrams depicting the sliding windows analysis methodology 72
3.16 Cell type mapping by Multi-scale Sliding Window analysis 74
4.1 Visualization of spatial transcriptomics data with STtools 79
4.2 Multi-scale Sliding Window analysis enables micrometer-resolution cell-type mapping 80
4.3 Spatial RGB visualization of marker gene sets by STtools 81
4.4 Additional visualization of spatial transcriptomics data produced by STtools 82
4.5 STtools workflow 86
4.6 Multi Scale Sliding Window (MSSW) Algorithm 88

LIST OF TABLES

Table

2.1 Number of droplets that are classified as cell-containing or cell-free exclusively to a specific method 20
2.2 Accuracy and recall of *SiftCell-Boost* evaluated by cross-validation . 25
2.3 Table for computational and memory usage across all methods . . . 30
4.1 Comparison between STtools and other related tools (spacemake and squidpy). 84

LIST OF ABBREVIATIONS

scRNA-seq single-cell RNA sequencing

ST Spatial Transcriptomics

SQuAT Sparse Quantile Aggregation Test

MSSW Multi-scale Sliding Window

RNA-seq RNA sequencing

DEG differentially expressed genes

snRNA-seq Single nucleus RNA-sequencing

scATAC-seq single cell ATAC sequencing

snATAC-seq single nuclei ATAC sequencing

DGE Digital Gene Expression

UMI Universal Molecular Identifier

E-M Expectation-maximization

mtRNAs mitochondrial RNA

UMAP Uniform Manifold Approximation and Projection

t-SNE t-Distributed Stochastic Neighbor Embedding

SNN shared nearest neighbor

MLE maximum likelihood estimates

HDMI high-definition map coordinate identifier

PC pericentral

PP periportal

NPC non-parenchymal

HSC hepatic stellate cells

ENDO endothelial cells

RBC red blood cells

DCSC deep crypt secretory cell

EEC enteroendocrine cells

CCK Cholecystokinin

SBS sequencing-by-synthesis

RTA realtime analysis

RGB red, green, and blue

PBMC peripheral blood mononuclear cells

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) and Spatial Transcriptomics (ST) have become instrumental tools to understand cellular dynamics and heterogeneity in disease-related tissues. As these technologies rapidly advance, many statistical and computational challenges arise in analyzing them, and relatively few methods address the challenges in the upstream processing of data from rapidly evolving technologies. This dissertation investigates computational challenges of upstream data analysis for scRNA-seq and ST, including quality control to identify and filter out droplets comprised of ambient RNAs, enabling high-resolution inference of ST data from a new submicrometer resolution technology, and developing robust computational tools and pipelines capable of handling ST platforms at various resolutions.

Following a brief overview of scRNA-seq, ST technologies and related challenges in Chapter I, in Chapter II, we focus on the problem of distinguishing cell-containing droplets from cell-free droplets that mostly contain ambient RNAs in scRNA-seq data from multiple angles. By leveraging efficient randomization, manifold visualization, statistical test tailored for sparse scRNA-seq data, and machine learning methods, we develop *SiftCell*, a suite of software tools to identify and visualize cell-containing and cell-free droplets in manifold space via randomization, to classify between the two types of droplets, and to quantify the contribution of ambient RNAs for each droplet. We also develop Sparse Quantile Aggregation Test (SQuAT), a statistical test designed to aggregate quantile-based summary statistics from many sparse

discrete datasets for meta-analysis. SQuAT robustly identifies likely cell-containing droplets and highly variable genes across cell types in sparse scRNA-seq data and is integrated as a core statistical method in *SiftCell*. Through a comprehensive evaluation of three scRNA-seq or snRNA-seq datasets we demonstrate that *SiftCell* enables new visualization of locating cell-free droplets in the manifold space and outperforms existing methods in filtering cell-containing droplets and in quantifying ambient RNA contribution.

In Chapter III, we introduce Seq-Scope, a new submicrometer resolution ST technology that repurposes the Illumina sequencing platform to achieve high resolution and scalability. Unlike other ST technologies, Seq-Scope does not require cumbersome image processing steps and leverages the existing sequencing platform to obtain spatial barcodes that are $0.5 - 0.8\mu m$ apart from each other, achieving a resolution comparable to that of an optical microscope. We performed the complete Seq-Scope experimental and analytical procedure on two representative gastrointestinal tissues (liver and colon). This chapter focuses on the computational aspects that enable the analysis of data produced from the new Seq-Scope technology.

In Chapter IV, we build a comprehensive software pipeline STtools that provides a versatile framework to handle ST datasets with various resolutions. STtools is designed to efficiently align, cluster and visualize ST data scaling with millions of spatially resolved barcodes. STtools improves the resolution of spatial inference compared to typical segmentation-based approaches by leveraging the Multi-scale Sliding Window (MSSW) algorithm. We applied STtools to several ST platforms, including Seq-Scope, Slide-seq and VISIUM and showed that STtools enables both analysis and visualization at various resolutions.

CHAPTER I

Introduction

1.1 Background

RNA sequencing (RNA-seq) is an important tool in molecular biology and biomedical research that allows researchers to quantify gene expression levels in a sample of cells or tissue(*Wang et al.* (2009);*Li and Wang* (2021)). By analyzing RNA-seq data, researchers can identify which genes are being actively transcribed and at what levels, providing insight into the molecular mechanisms behind disease development or progression and aids in the development of targeted therapeutics(*Ozsolak and Milos* (2011)).

RNA-seq technology has advanced significantly in recent years, evolving from classic bulk RNA-seq, popular single-cell RNA-seq(scRNA-seq) to newly emerging Spatial Transcriptomics(ST).

Bulk RNA-seq involves sequencing the RNA extracted from a population of cells, providing average global gene expression levels across all cells in a sample(*Wang et al.* (2009),*Stark et al.* (2019)). It is useful for identifying differentially expressed genes (DEG) between different tissues, conditions, or time points. However, bulk RNA-seq cannot capture cell-to-cell variability, or distinguish gene expression differences between individual cells within the population and can mask rare cell populations.

The development of scRNA-seq is motivated by the imperative need to unravel

the heterogeneity of gene expression patterns across individual cells within a population(*Jovic et al.* (2022)). scRNA-seq technologies allow us to simultaneously profile the transcriptomes of thousands of individual cells, enabling us to understand the regulatory impact of genetic, developmental, environmental, and clinical determinants in a single cell resolution (*Klein et al.* (2015); *Macosko et al.* (2015)). Single nucleus RNA-sequencing (snRNA-seq), Single cell ATAC sequencing (scATAC-seq), Single nuclei ATAC sequencing (snATAC-seq) and other high throughput single cell genomic profiling technologies (*Buenrostro et al.* (2015); *Habib et al.* (2017); *Preissl et al.* (2018)) can also scale to thousands of cells or nuclei, providing us with the comprehensive epigenomic landscape of individual cells or cell types. The rapid advances of single-cell genomic technologies have revolutionized our ability to understand the dynamics of individual cells and have given us unprecedented opportunities to characterize cellular heterogeneity, which is essential for understanding and treating human disease.

scRNA-seq is an effective technique for revealing gene expression profiles of individual cells in liquid tissues such as blood. However, it is not adequate for revealing the cellular heterogeneity in solid tissues. scRNA-seq for solid tissues requires extensive tissue dissociation and single-cell sorting procedures. The rigorous techniques employed during these procedures can create harsh conditions that may eliminate labile cell populations and induce stress responses(*Volovitz et al.* (2016); *O'Flanagan et al.* (2019)). The practical obstacle drives the rapid developments in spatially resolved, high dimensional assessment of gene transcriptome, known as 'spatial transcriptomics' (ST) that combines the power of scRNA-seq and the ability to spatially map gene expression patterns within a tissue without inducing stress, cell death, and/or cell aggregation (*Williams et al.* (2022)). By capturing the transcriptome directly from a frozen tissue slice, ST can be used to characterize transcriptional patterning and regulation in tissues, reveal tissue neighborhoods and local features contributing to

disease and be used to study individual cells and cell types in detail (*Garcia-Alonso et al.* (2021)).

1.2 Importance of Upstream Quality Control in Single-cell RNA Sequencing

As scRNA-seq has become essential for biomedical research over the past decade, many software tools have been developed for downstream analysis such as to identify cell types(*Grün and van Oudenaarden* (2015);*Jiang et al.* (2016); *Kiselev et al.* (2017); *Tsoucas and Yuan* (2018)) or developmental trajectories (*Qiu et al.* (2017); *Schiebinger et al.* (2019); *Setty et al.* (2016); *Welch et al.* (2016)), to account for systematic differences across experimental batches or technologies, to identify differentially regulated genes by clinical variables or genotypes (*Ntranos et al.* (2019); *Soneson and Robinson* (2018)), or to enable efficient single cell experiment via multiplexing (*Stoeckius et al.* (2018)). Whereas only few methods were developed for upstream quality control(*Lun et al.* (2019),*Alvarez et al.* (2020),*Fleming et al.* (2019)). In fact, such quality control is crucial to ensure that the downstream analysis is not misled by potential technical artifacts, such as sequence alignment or Digital Gene Expression (DGE) matrix generation. Incorrectly filtered scRNA-seq/snRNA-seq may lead to identifying spurious clustering or false positive cell types.

A key computational challenge for upstream quality control is to identify libraries for droplets containing real cells. Ideally, scRNA-seq reads from individual cells or nuclei are distinctly barcoded, however, in reality, each observed barcode may not correspond to a single cell or nucleus. One of the the reason lies in that a droplet may fail to encapsulate the entirety of single cell, and instead captures "cell debris" or "ambient mRNAs" produced from the damaged and lysed cells. Barcodes derived from such droplets containing defective or ambient mRNAs may be mistaken to represent

single cell transcriptome while they are not. We will denote such a barcoded droplet enriched for ambient RNAs as "cell-free droplets". Because these cell-free droplets enriched for ambient mRNAs do not represent single cells, failure in filtering out such droplets produces misleading interpretations in the downstream analysis. Hence, it is an essential quality control procedure to filter out cell-free droplets to ensure that scRNA-seq and snRNA-seq analysis produce biologically relevant information. However, distinguishing cell-free and cell-containing droplets is often challenging, and incorrect discrimination may mislead the downstream analysis substantially.

In Chapter II, we focus on the challenges of contamination from ambient RNAs in single-cell and single-nucleus RNA-seq experiments and propose *SiftCell*, a suite of software tools to visualize cell-free and cell-containing droplets in a more intuitive way, to robustly classify between the two types of droplets and to quantify the fraction of ambient RNAs contamination in each droplet.

1.3 Limited Resolution of Existing Spatial Transcriptomics Technology

Spatial Transcriptomics (ST) is a newly developed technology for analyzing gene expression patterns in tissues while preserving their spatial organization (*Asp et al.* (2020)). There are three main experimental methods for ST: (1) the sequential *in situ* hybridization method which can increase the number of RNA species that can be detected from a single histological section. (2) *in situ* sequencing which can identify RNA sequences from the tissue by fluorescence-based direct sequencing. (3) spatial barcoding methods that associate RNA sequences and their spatial locations by capturing tissue RNA using a spatially barcoded oligonucleotide array. The ST field continues to grow rapidly and in 2021 and was named 'Method of the Year 2020' by Nature Methods (*Marx* (2021)).

Despite the fast pace of ST technology development, there is still an intrinsic limitation due to the low-resolution specifications of current ST technologies. For example, VISIUM from 10X Genomics has a center-to-center resolution of $100\mu m$ (*Asp et al.* (2020)), which is worse than that of the naked eye ($40\mu m$). More recent technologies, such as Slide-Seq, HDST, and DBiT-Seq, improved the resolution (*Rodriques et al.* (2019); *Vickovic et al.* (2019); *Liu et al.* (2020)); however, their resolutions are still far coarser than optical microscope that has submicrometer resolution. In addition to the technical challenge, there is also the needf for computational tools to efficiently map between millions of spatial barcodes and RNA sequences and to deal with the sparsity of sequence reads when focusing on small regions corresponding to single cell or subcellular regions.

In chapter III, we introduced a ST technology "Seq-Scope" that achieves submicrometer resolution, comparable to an optical microscope. We conducted a series of computational analysis to show that Seq-Scope visualizes ST heterogeneity at multiple histological scales.

1.4 Need for Software Pipelines to Handle Ultra-high-resolution Spatial Transcriptomics Data

With the rapid development of ST technologies, many analytical software pipelines/tools (10x Genomics, 2022; *Palla et al.* (2022); *Petukhov et al.* (2022), etc.) have been developed for researchers to analyze, interpret and gain insights from the large and complex ST datasets. Current software tools analyzing spatially resolved transcriptomes are primarily designed for relatively coarse resolution technologies such as VISIUM or Slide-Seq, where each spatial barcode typically represents more than a single cell. However, when analyzing transcriptome spatially resolved at a micrometer or a submicrometer resolution, current tools perform poorly due to several challenges. Low-

vs high-resolution ST datasets are very different with orders of magnitude differences in the number of spatial barcodes per mm^2. As the resolution increases, the reads per spatial barcode become sparser, which necessitates the development of robust methods to make inferences from sparse data at high resolution. Furthermore, scalable methods are needed to process hundreds of millions of barcodes and RNA sequences together and new software tools/pipelines are needed to support emerging new ST technology with high resolution.

In Chapter IV, we present STtools, a comprehensive ST pipeline that provides a versatile framework to handle ST platforms with various resolutions including but not limited to VISIUM, Slide-Seq and Seq-Scope. STtools is designed to efficiently align, cluster and visualize ST scaling with millions of spatially resolved barcodes.

1.5 Future Directions

RNA sequencing has revolutionized human genetics research by enabling researchers to study the expression and regulation of genes at the transcriptome level. Over the past decade, RNA-seq has evolved from bulk RNA-seq, which provides an average measure of gene expression across all cells in a sample, to scRNA-seq, which enables the measurement of gene expression in individual cells. More recently, ST has emerged as a powerful technology that enables the analysis of gene expression in a tissue context. And it is still rapidly evolving with emergence of new technologies and studies accompanied by new statistical and computational challenges. In Chapter V, we review our work that focus on the problem of distinguishing cell-containing droplets from cell-free droplets that mostly contain ambient RNAs in scRNA-seq and snRNA-seq data from multiple angles, computational/pipeline challenges for analyzing submicrometer-resolution data produced from the new Seq-Scope spatial transcriptomics technology. In addition, we examine the limitations of current studies, discuss remaining challenges and explore future opportunities.

CHAPTER II

SiftCell: A Robust Framework to Identify and Filter Cell-free and Cell-containing Droplets from Single-cell RNA Sequence Reads

2.1 Background

2.1.1 Droplet Barcoding and Ambient RNAs

In single-cell RNA sequencing(scRNA-seq), droplet barcoding is a technique to uniquely label individual cells or nuclei so that thousands or millions of cells or nuclei can be simultaneously sequenced in a single library. The sequenced reads can be grouped into the originating cells or nuclei according to the barcodes, and the grouped information is used for the downstream single cell analysis. However, each observed barcode may not correspond to a single cell or nucleus due to several reasons. First, sequencing errors may lead to incorrect assignment of each read into its originating cells or nuclei. Second, two or more cells or nuclei can be encapsulated within the same barcoded droplet, forming "multiplets", either stochastically or due to imperfect dissociation of tissues (*Cao et al.* (2017); *Klein et al.* (2015); *Macosko et al.* (2015)). Third, a droplet may fail to encapsulate the entirety of a single cell, and instead captures "cell debris" or "ambient mRNAs" produced from the damaged

and lysed cells (Figure 2.1). Barcodes derived from such droplets containing defective or ambient mRNAs may be mistaken to represent single cell transcriptome while they are not. We will denote such a barcoded droplet enriched for ambient RNAs as "cell-free droplets". It is reported that scRNA-seq from solid tissues are more enriched for ambient mRNAs, particularly when incubated at high temperature (*O'Flanagan et al.* (2019)), which renders cells to be more vulnerable and therefore producing more cell death and lysis. Different technologies have different susceptibilities of contaminating their datasets with cell-free droplets. For example, snRNA-seq technologies inherently produce more ambient mRNAs, therefore are more likely to generate cell-free droplets compared to conventional scRNA-seq (*Alvarez et al.* (2020)). Because these cell-free droplets enriched for ambient mRNAs do not represent single cells, failure in filtering out such droplets produces misleading interpretation in the downstream analysis. Therefore, filtering out cell-free droplets are an essential quality control procedure to make sure scRNA-seq and snRNA-seq analysis produces biologically relevant information.

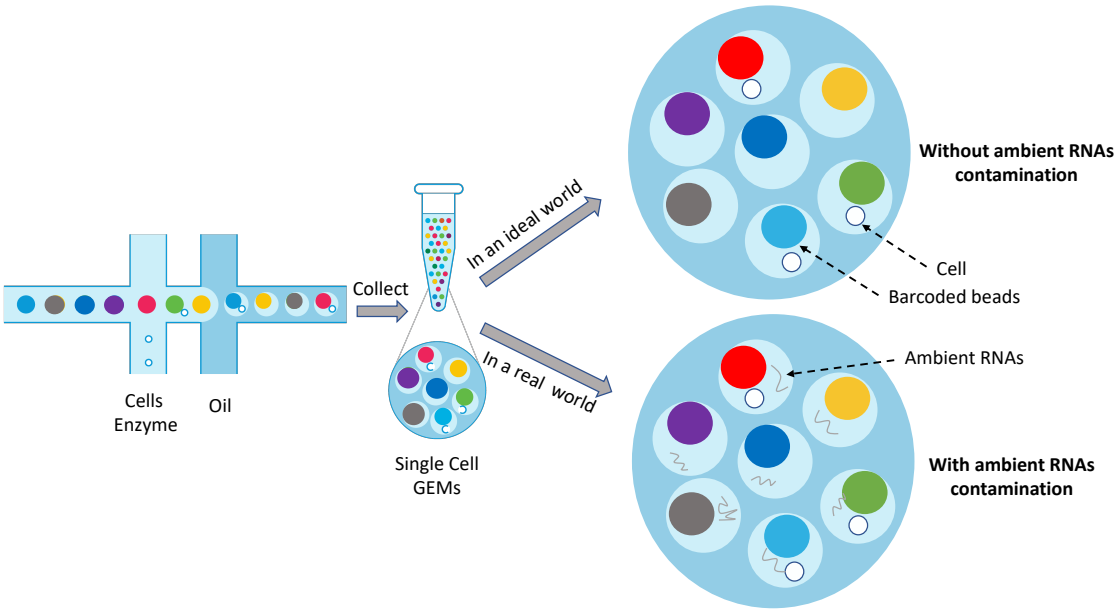

Figure 2.1: Droplet barcoding and ambinet RNA contamination. In droplet barcoded microfluidoc devices, both cells and ambient RNAs that were released from damaged or lysed cells can be encapsulated in a droplet representing a cell-containing droplet.

2.1.2 Literature Review on Droplet Filtering Methods

In real-world scRNA-seq data, many barcoded droplets do not contain cells, but instead, they capture a fraction of ambient RNAs that were released from damaged or lysed cells. A typical first step to analyze droplet-based scRNA-seq data is to filter out cell-free droplets and isolate cell-containing droplets. The simplest and most widely used strategy to filter out cell-free droplets is to remove droplets with very low number of unique reads or Universal Molecular Identifier (UMI) counts. This strategy is based on the simple fact that, compared to cell-containing droplets, cell-free droplets contain less mRNAs because ambient mRNAs will be excessively diluted in the media surrounding the cells, leading to low UMI counts. Earlier versions of cellRanger and DropseqTools software tools filter out droplets below a certain UMI cutoff determined from the distribution of UMI counts across the barcoded droplets and from a user-specified parameter of expected number of cells using knee plots(*Macosko et al.* (2015); *Zheng et al.* (2017)). While this strategy works quite effectively in practice, it relies on a simplistic assumption that all droplets containing individual cells will have higher UMI counts than other barcoded droplets. Because UMI counts result from a stochastic procedure involving multiple factors, this simplistic assumption does not always hold.

Recently, alternative approaches have been developed to identify and filter out cell-free droplets using more sophisticated statistical models. For example, EmptyDrops method (*Lun et al.* (2019)), which is adopted to the newer version of cellRanger (v3), first determines a UMI cutoff from the knee plot to identify cell-containing droplets, and then attempts to rescue droplets below the UMI cutoff using a statistical test. The assumption is that the expression profile of cell-free droplets is homogeneous, which can be estimated as a Dirichlet-Multinomial distribution. If the likelihood of observed read count from a barcoded droplet is significantly lower than those from simulated reads, EmptyDrops identifies them as cell-containing droplets. EmptyDrops is useful

only for rescuing cell-containing droplets with lower UMI counts and cannot filter out cell-free droplets with high UMI counts. DecontX and SoupX (*Yang et al.* (2020); *Young and Behjati* (2020)), on the other hand, assumes that every droplet contains a certain fraction of ambient RNAs, and attempts to estimate the proportion of ambient RNA contamination, and determines cell-free droplets if the estimated proportion is above a specific threshold. A recently developed method DIEM (*Alvarez et al.* (2020)), uses an Expectation-maximization (E-M) algorithm (*Dempster et al.* (1977)) to cluster barcoded droplets into cell types while modeling cell-free droplets as a separate cluster using Dirichlet-Multinomial distribution. CellBender uses a deep generative model implemented by neural auto-encoders to model scRNA-seq data and applies a variational mix to evaluate the posterior probability of cell-free droplet. Finally, DropletQC (*Muskovic and Powell* (2021)), estimates proportion of intronic reads from sequence reads and use the information to separate droplets containing damaged cells or ambient RNAs.

2.1.3 *SiftCell* Framework

While these droplet filtering methods demonstrated their utility in some of the real datasets, it is often not clear what are objective criteria to evaluate their performances in distinguishing cell-containing droplets from cell-free droplets.

Here, we propose *SiftCell*, a suite of three software tools (*SiftCell-Shuffle*, *SiftCell-Boost* and *SiftCell-Mix*) to address challenges due to cell-free droplets in conceptually unique ways (Figure 2.2).

The first tool *SiftCell-Shuffle*, allows us to visually distinguish cell-free and cell-containing droplets in sc/snRNA-seq experiments to help filter out cell-free droplets. *SiftCell-Shuffle* takes an arbitrary DGE matrix to visualize the distribution of potentially cell-free barcoded droplets in a manifold space using randomization. The second tool, *SiftCell-Boost* distinguishes cell-containing droplets from cell-free droplets using

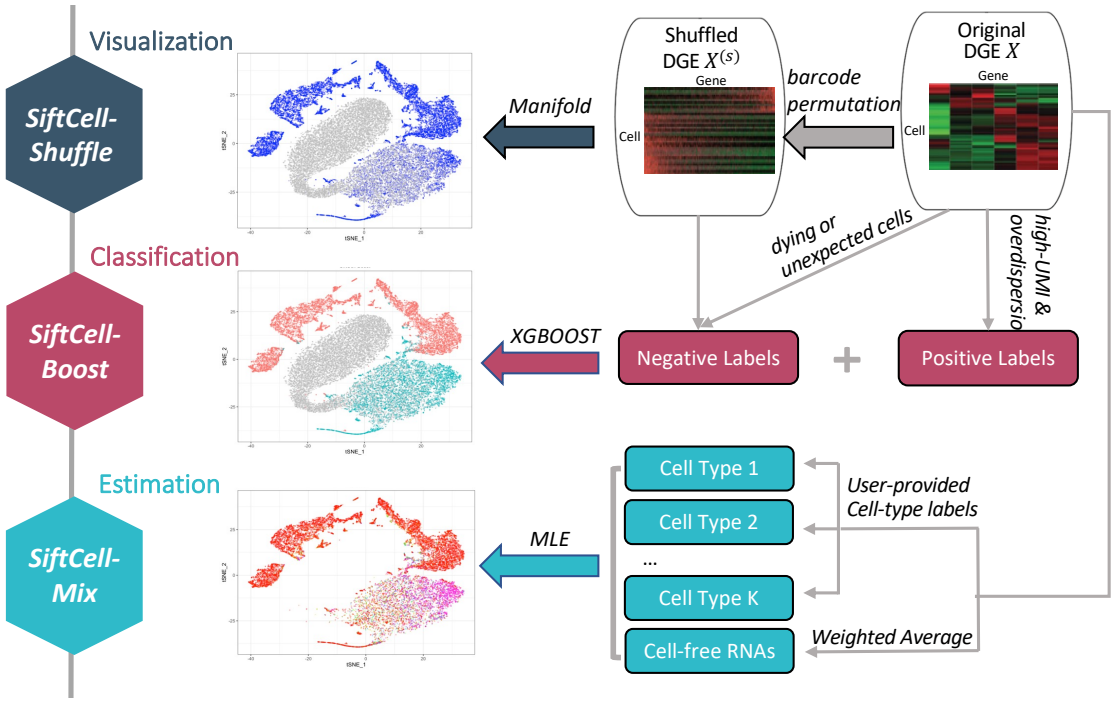

Figure 2.2: Overview of *SiftCell* Framework. The *SiftCell* software package includes three tools for visualizing and filtering barcoded droplets from scRNA- or snRNA-seq experiments: *SiftCell-Shuffle* visualizes original barcoded droplets with randomized droplets on a manifold space to distinguish cell-containing and cell-free droplets visually in the manifold space; *SiftCell-Boost* classifies cell-containing droplets and cell-free droplets by leveraging the results from *SiftCell-Shuffle* results with gradient boosting algorithm; *SiftCell-Mix* estimate the proportion of ambient RNAs in each barcoded droplet.

a semi-supervised learning algorithm guided by the labels generated with *SiftCell-Shuffle*. The third tool, *SiftCell-Mix* is a model-based method that allows estimation the proportion of "ambient RNAs" in each barcoded droplet. We also provide a comprehensive evaluation of existing methods to identify cell-free droplets. Therefore, in addition to providing an intuitive way of eliminating cell-free droplets and selecting cell-containing droplets, our method can evaluate and visualize the strength of each different previously available method to inform users to guide the best practice for handling cell-free droplets.

2.2 Results

2.2.1 *SiftCell-Shuffle* Visually Distinguishes Cell-free Droplets from Cell-containing Ones

Even though there are multiple methods to determine cell-containing droplets from sc/snRNA-seq data, currently there is no systematic way to evaluate whether one method more robustly distinguish cell-containing and cell-free droplets than the other method with real data. Previous studies utilized indirect measurement, such as fraction of mitochondrial RNA (mtRNAs) (*Alvarez et al.* (2020)) or UMI counts (*Lun et al.* (2019)), but they can be confounded by cell types (e.g. some cell types may contain more mtRNAs and less UMIs) or technical factors (e.g. some scRNA-seq preps contain high amount of ambient mRNAs). Other studies compared manifold plots (such as Uniform Manifold Approximation and Projection (UMAP) or t-Distributed Stochastic Neighbor Embedding (t-SNE)) after applying different filtering method and argue for one over the other based on their visual patterns of clustered cell types (*Fleming et al.* (2019)), but such interpretations can easily become subjective.

We developed *SiftCell-Shuffle*, a randomization-based scRNA-seq visualization tools focusing on distinction between cell-containing and cell-free droplets. *SiftCell-Shuffle* assumes that the ambient RNAs are distributed as a pseudo-bulk (i.e. in a single distribution across all dataset) while cell-containing RNAs are distributed in a cell-type-specific manner. Based on this assumption, *SiftCell-Shuffle* creates a digital expression matrix that mimics the "bulk" distribution by randomizing the droplet barcode assignments across the UMIs. After randomization, the original and randomized DGE matrices are jointly analyzed using a standard scRNA-seq workflow (e.g. Seurat) and individual droplets are visualized in a t-SNE and/or UMAP manifold space (Figure 2.2, See Materials and Methods for further details). For cell-containing droplets, the original and randomized data should have very different transcriptomic

profiles and will be located at very distant points to each other in the manifold space. For cell-free droplets containing mostly ambient RNAs, the original and randomized data are more likely to be located in close proximity, so the cluster of cell-free droplets can be clearly visualized.

We first assessed the performance of *SiftCell-Shuffle* in the three experimental scRNA-seq or snRNA-seq datasets. First is scRNA-seq of $10,000$ peripheral blood mononuclear cells (PBMC) using 10X Chromium v3 chemistry. Second is snRNA-seq of $1,000$ E18 mouse brain nuclei using 10X Chromium, available at `https://www.10xgenomics.com/resources/datasets`. Third is scRNA-seq of $1,000$ cultured colon mixture data pooled across 3 cell lines (RKO, HCT116, SW480), profiled using Drop-Seq technique(*Park et al.* (2020)). We expect that the PBMC dataset is more straightforward to distinguish cell-containing droplets from cell-free droplets than the other two datasets because snRNA-seq or Drop-Seq are known to be more enriched for ambient RNAs.

When we applied *SiftCell-Shuffle* on the unfiltered PBMC dataset together with unsupervised clustering produced by Seurat (*Butler and Satija, 2017*), we observed a clear separation between the "original" (clusters 2, 3, 5, 6, 7, 8) and "shuffled" (clusters 1, 4) droplets in both t-SNE and UMAP manifolds (Figure 2.3 A), except for the cluster 0, which had much lower UMI counts than other clusters (Figure 2.3B, 2.3 C). The original droplets that belongs to cluster 0 showed larger dispersion of UMIs across genes (Figure 2.3 C), and is also enriched with mtRNAs (18.5% of UMIs compared to 10.1% in other clusters; Figure 2.3 D). Altogether, these observations strongly suggest that the Cluster 0 represents cell-free droplets enriched for ambient mRNAs. On the other hand, the rest of clusters containing original droplets (cluster 2, 3, 5, 6, 7, and 8) contained very few randomized droplets (0 - 0.5%), suggesting that they likely represent cell-containing droplets with different cell types. Using known marker genes specific to immune cell types, we demonstrated that each of these

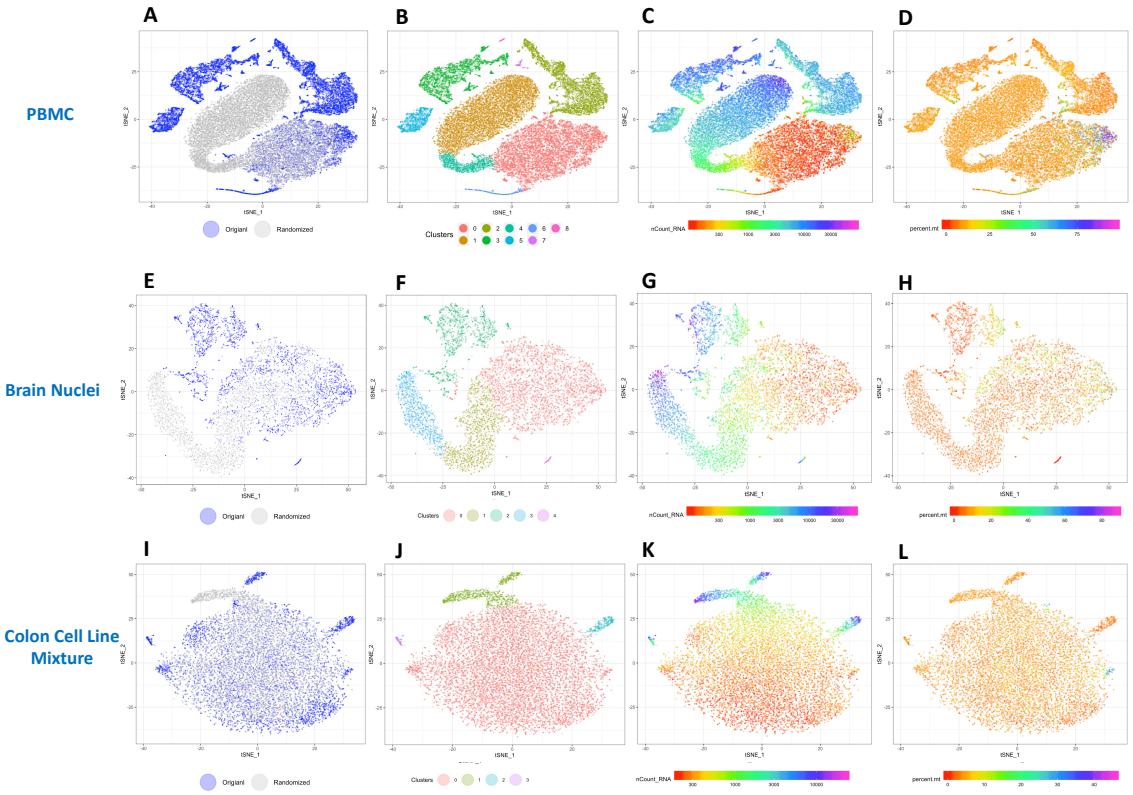

Figure 2.3: Visualization of *SiftCell-Shuffle* results in t-SNE manifold space. These four panels visualize original and randomized droplets from the result of *SiftCell-Shuffle* for PBMC dataset(A-D), brain nuclei dataset(E-H) and colon cell line mixture dataset(I-L) in t-SNE manifold space generated using Seurat v3(*Butler et al.* (2018)). The t-SNE manifolds were colored by (A ,E, I) original (blue) vs. randomized (grey) droplets, (B, F, J) clusters produced by Seurat with with FindNeighbors and FindClusters functions, (C, G, K) the total number of UMIs in logarithmic scale corresponding to the droplet across all genes, and (D, H, L) the fraction of mitochondrial RNAs in logarithmic scale. For PBMC dataset, In (A), we see clear separation between the original (blue) and randomized (gray) droplets except for the cluster (cluster 0 in (B)) in the lower-right quadrant, which we believe to be enriched for cell-free droplets. This cluster tends to have lower UMI counts in (C) and contains droplets with higher proportion of mitochondrial RNAs in (D). However, it is important to know that not all randomized droplets are clustered together in (A). Randomized droplets with higher UMI counts tend to form their own clusters (cluster 1 and 4 in (B)). This is because randomized droplets with higher UMI counts do not necessarily share similarities with cell-free droplets, because UMI count plays a role as a confounding variable. Similarly, for brain nuclei data, In panel (E), original (blue) and randomized (gray) droplets are separated clearly except for the droplets in cluster 0 in (F). This cluster is believed to be enriched for cell-free droplets with lower UMI count which can be shown in (G). However, not all randomized (gray) droplets are clustered together. Droplets with higher UMI count tend to form a separate cluster (cluster 1,3 in (F)). Similar observation can be found for colon mixture cell line dataset.

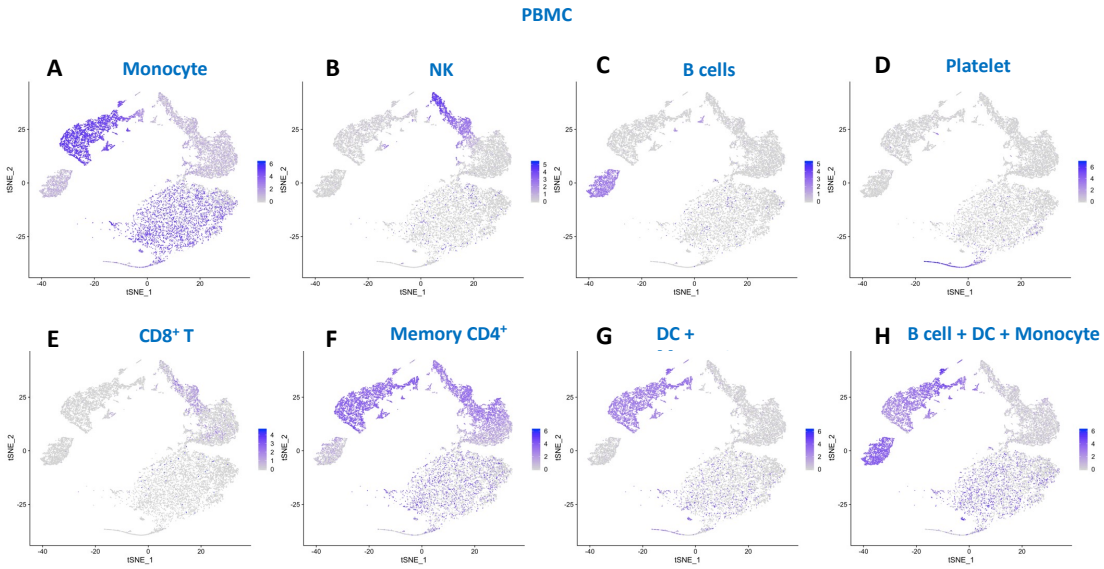

Figure 2.4: Feature plot of cell type-specific marker genes in PBMC dataset. The eight panels (A-H) visualize the feature plot using known marker genes specific to eight immune cell types, *LYZ* for Monocyte, *NKG7* for NK cells, *MS4A1* for B cells, *PPBP* for Platelet cells, *CD8A* for $CD8^+$ and T cells, *S100A4* for memory $CD4^+$ cells, *CST3* for Dendritic Cell (DC) and Monocyte, *HLA-DRA* for B cell, DC and Monocyte.

clusters indeed show specific enrichment for specific immune cell types, while cluster 0 shows non-specific expression across most of these genes (Figure 2.4). By visualizing both original and randomized droplets together in a single manifold space, our results suggest that *SiftCell-Shuffle* distinguishes clusters of cell-containing droplets from cell-free droplets in a straightforward and visually interpretable/inspectable way.

We made similar observations when applying *SiftCell-Shuffle* on the other two datasets. For example, among the 5 clusters of brain nuclei, original droplets (clusters 0, 2, 4) and randomized droplets (clusters 0, 1, 3) were well-separated except for cluster 0, suggesting that it represents cell-free droplets (Figure 2.3 E,F). Cluster 0 also tend to have the lower UMI counts and high mtRNAs overall. Interestingly, In cluster 2, we also observed that a substantial fraction of the droplets with high mtRNAs (Figure 2.3 H). These droplets may represent nuclei undergoing necrosis (*Young*

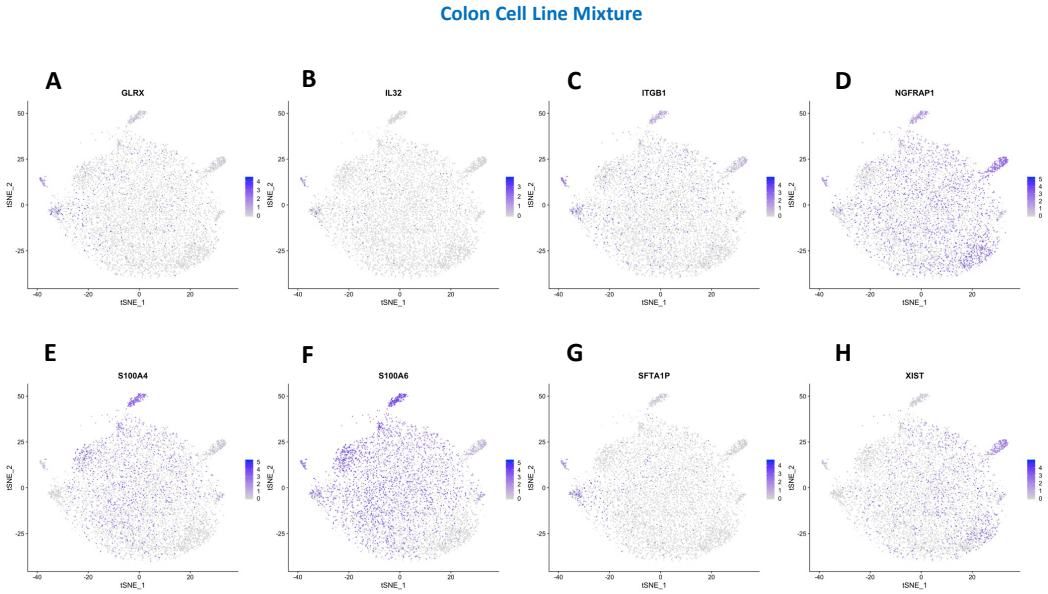

Figure 2.5: Feature plot of cell type-specific marker genes in colon cell line mixture dataset. The eight panels visualize the feature plot using known marker genes specific to three colon cell line mixture - HCT116, SW480 and RKO. The eight selected genes are specifically enriched for RKO (A, B, C, G), SW480 (D, H), or HCT116 (E, F).

and Behjati (2020)) or "nucleus-containing" droplets that also contain a substantial amount of ambient RNAs. When applying *SiftCell-Shuffle* on the Drop-Seq dataset of colon cell line mixture, we observed four distinct clusters, three representing each of three cell lines, and the largest cluster representing cell-free droplets with much lower UMI counts (Figure 2.3 I-L). Clusters representing each cell line were enriched for genes specific to the cell line (*Park et al.* (2020)), while the cell-free cluster tends to express most of these genes at a lower expression levels, suggesting that they contain ambient mRNAs as a mixture of multiple cell lines (Figure 2.5).

Across the three datasets, visualizing the original and randomized droplets in a low-dimensional manifold space provided us with a straightforward way to distinguish cell-free and cell-containing droplets. Applying unsupervised clustering based on shared nearest neighbor (SNN) (*Waltman and Van Eck* (2013)) was effective in distinguishing clusters of cell-free droplets from cell-containing droplets in PBMC

(Figure 2.3 A-D) and brain nuclei (Figure 2.3 E-H). However, in colon cell line mixture, one of the clusters (cluster 1) largely contained both cell-containing (mostly HCT116) and randomized droplets together (Figure 2.3 I,J), suggesting that unsupervised clustering does not always distinguish clusters of cell-free droplets automatically. Moreover, distinguishing cell-containing droplets from cell-free droplets by visual inspection from *SiftCell-Shuffle* without additional "gold-standard" labels involves subjective decision by users and may be hard to be automated in a software tool. Therefore, while *SiftCell-Shuffle* is an intuitive and human-interpretable approach to identify clusters of cell-free droplets, it does not completely replace existing methods to filter cell-containing droplets in a more systematic fashion.

2.2.2 Evaluating the Performance of Droplet Filtering Using *SiftCell-Shuffle*

Our *SiftCell-Shuffle* framework can also be used to evaluate different approaches to filter DGE matrix that allow us to focus on cell-containing droplets in the downstream analysis. While this approach would not be as accurate as evaluation based on "gold-standard" labels, in the absence of knowledge of true cell-free and cell-containing droplets, *SiftCell-Shuffle* can provide quasi-groud truth as a silver standard. We applied four existing filtering methods and visualized the filtering results in the manifold space produced by *SiftCell-Shuffle*. By contrasting the distribution of original and randomized droplets in the manifold space, it clearly demonstrates that cellRanger2-filtered (by UMI-cutoff) cell-containing droplets more specifically than the other methods in PBMC dataset (Figure 2.6 A-D). In the brain nuclei dataset, CellRanger/UMI-cutoff and EmptyDrops much more stringently filtered cell-containing droplets than DIEM and CellBender (Figure 2.6 F-I). In the mixture of three colon cancer cell lines, all of the four methods filtered the cell-containing droplets too stringently (CellRanger/UMI-cutoff) or too leniently (EmptyDrops, DIEM, Cell-

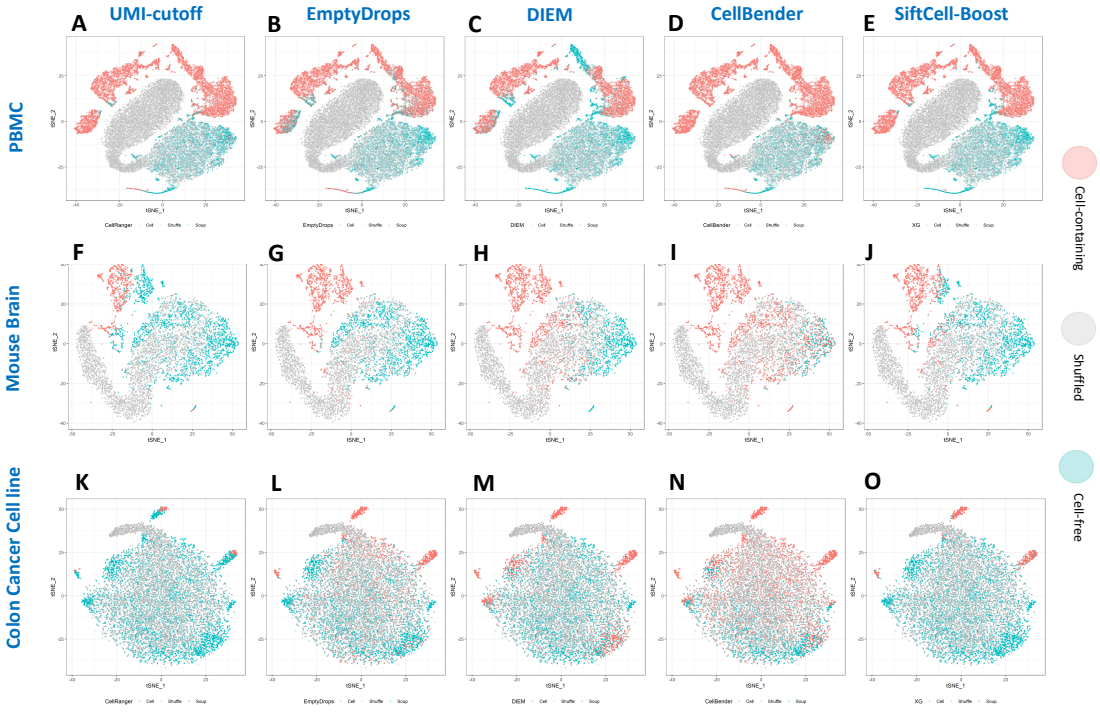

Figure 2.6: Evaluation of droplet filtering methods with *SiftCell-Shuffle*. Each panel visualizes the results of droplet filtering methods in the same manifold spaces described in Figure 2.3. Each colored points represents predicted cell-containing droplets (red), predicted cell-free droplets (cyan), or randomized droplets (grey). Each row corresponds to PBMC (A-E), brain nuclei (F-J), and colon cell line mixture (K-O) datasets, respectively. Each column visualizes the results from different droplet filtering methods, including CellRanger/UMI-cutoff (A,F,K), EmptyDrops (B,G,L), DIEM (C,H,M), CellBender (D,I,N), and *SiftCell-Boost* (E,J,O).

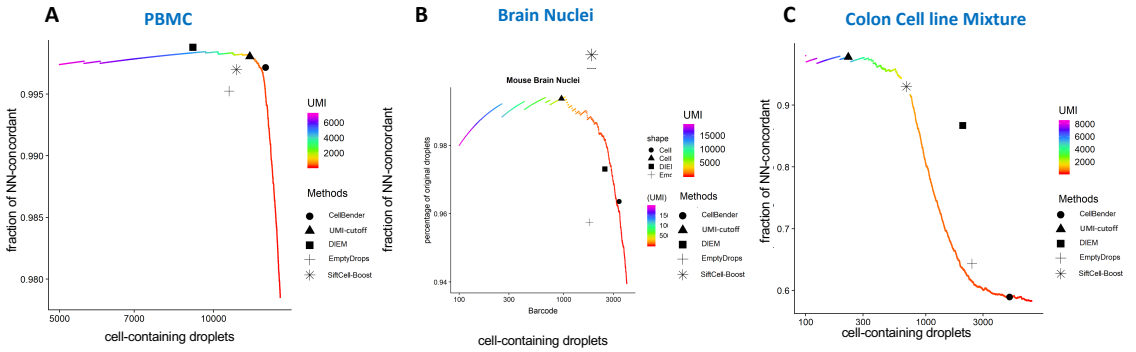

Figure 2.7: %NN-concordance evaluation plot. The three panels visualize the %NN-concordance among filtered droplets across (A) PBMC, (B) Brain nuclei, and (C) Colon cell line mixtures. In each plot, %NN-concordance was evaluated in two ways. The colored lines represent %NN-concordance in logarithmic scale when selecting top x droplets with highest UMIs as cell-containing droplets. Each point represents the number of filtered droplets (x-axis) and the corresponding %NN-concordance (y-axis) across five filtering methods - CellBender, UMI cutoff, DIEM, EmptyDrops and *SiftCell-Boost*. With this definition, the CellRanger/UMI-cutoff method will always be located on the colored line.

Bender) (Figure 2.6 K-N).

Besides visual inspections, we can also quantitatively evaluate droplet filtering methods using *SiftCell-Shuffle*. For a filtered droplet, we can quantify how often its nearest neighbor is an original droplet as opposed to a randomized droplet (named as % NN-concordance) as a metric. A high %NN-concordance suggest that the filtered droplets are well-separated from randomized droplets (Figure 2.7). A typical method to determine the number of cell-containing droplet is knee plot (Figure 2.8). However, our %NN-concordance plot is more informative to pinpoint where ambient RNAs start to increase. Each filtering method can be placed on this operating characteristic curve for evaluation, too. For example, in PBMC dataset, it is clear that EmptyDrops is worse than other methods in terms of % NN-concordance (Figure 2.7 A). Among the other three methods, DIEM appears to filter too few droplets ($n = 9,112$ droplets) even though %NN-concordant droplets remained high even after $10,000$ droplets. In brain nuclei and colon cell line mixture, we observed that CellRanger/UMI-cutoff

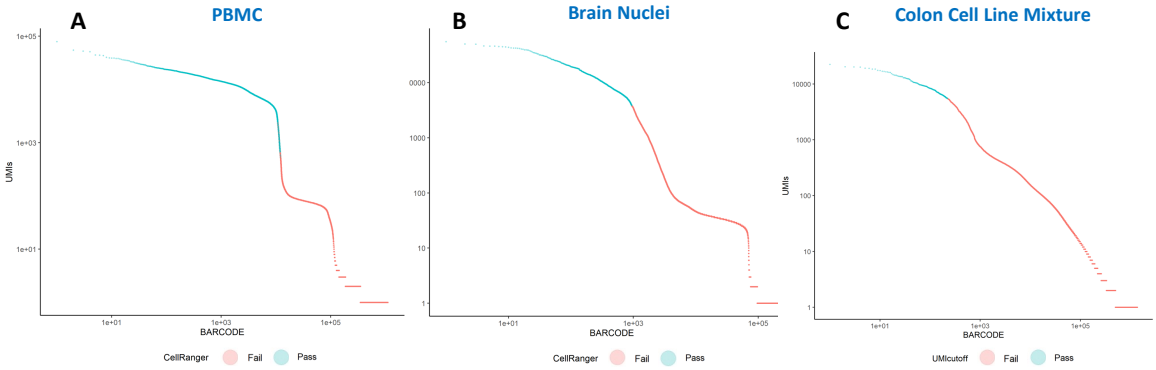

Figure 2.8: Knee plots across PBMC, brain nuclei and colon cell line mixture datasets. The three panels show the knee plots showing the number of UMIs for each barcode, ordered by the decreasing order of UMI counts. Both x- and y-axis are in log-scale, and the color of the plot represents whether the barcoded droplet passed (green) or failed (red) based on the CellRanger/UMI-cutoff method. Each panel represents (A) PBMC, (B) brain nuclei and (C) colon cell line mixture datasets.

appears to filter stringently while others filter leniently (Figure 2.6 F,K).

Table 2.1: Number of droplets that are classified as cell-containing or cell-free exclusively to a specific method. Each row represents a method classify cell-containing droplets, and each column represents "outliers", meaning the number of droplets classified differently from the other four methods. All methods, except for *SiftCell-Boost*, have at 2 of 3 datasets that have > 200 droplets classified exclusively (i.e. discordant to all other methods) to them. *SiftCell-Boost* showed very small number (< 10) of droplets classified differently from all the other methods.

Method	**PBMC**		**Brain Nuclei**		**Colon Cell Line Mixture**	
	Exclusive Cell-containing	**Exclusive Cell-free**	**Exclusive Cell-containing**	**Exclusive Cell-free**	**Exclusive Cell-containing**	**Exclusive Cell-free**
CellRanger/UMI-cutoff	0	0	0	268	0	455
CellBender	646	0	836	0	2059	0
DIEM	0	1427	0	17	381	3
EmptyDrops	188	647	28	22	249	0
SiftCell-Boost	0	5	0	3	0	12

2.2.3 *SiftCell-Boost* Robustly Filters Cell-containing and Healthy Droplets

Because none of the existing droplet filtering methods always provided satisfactory performance across all datasets in our evaluation, we next attempted to develop a

method to filter cell-containing droplets by leveraging results from *SiftCell-Shuffle*. Our approach applies a gradient boosting classification algorithm XGBoost (*Chen and Guestrin* (2016)) by assigning randomized droplets as negative labels (representing ambient RNAs) and droplets confidently predicted to contain cells as positive labels using an overdispersion test (see Materials and Methods for details). *SiftCell-Boost* assumes that the positively or negatively labeled droplets are confident cell-containing or cell-free droplets, respectively, and focuses on classifying the unlabeled droplets (10% in PBMC, 71% in brain nuclei, and 66% in colon cell line mixture) into either cell-containing or cell-free droplets.

This is because our method assumes that the distribution of ambient RNAs are random samples from existing reads, but in fact they tend to be enriched for higher proportion of mtRNAs due to necrosis. To address this challenge, we marked droplets with excessive proportion of mtRNAs as additional negative labels (see Materials and Methods). In addition, for PBMC dataset, to avoid including unintended cell types (i.e., platelets), we also marked droplets with excessive proportion of PPBP as negative labels. With these additional negative labels, *SiftCell-Boost* clearly outperformed existing methods on PBMC and colon cell line mixture and was comparable with other methods for brain nuclei (Figure 2.10). We also evaluated the concordance of droplet classification between the five evaluated methods. We counted how often a specific method exclusively classified each droplet into either cell-containing or cell-free droplets discordantly from all the other methods (Figure 2.9, Table 2.1). For example, we found that cell-containing droplets identified from CellRanger/UMI-cutoff were always consistent with at least one of the other methods. However, 288 and 455 cell-free droplets determined by CellRanger/UMI-cutoff were discordant with all the other methods for brain nuclei and colon cell line mixture data, suggesting that the method has high specificity but poor sensitivity. With this criteria, all four methods except for *SiftCell-Boost* had two or more datasets where > 200 droplets were discor-

dantly classified with all the other methods. However, *SiftCell-Boost* had 12 or less droplets discordantly classified with all other methods, suggesting that classification is more consistent to the consensus among all methods. We also evaluate the accuracy of *SiftCell-Boost* using 5-fold cross validation and obtained an average accuracy of 99.92% for PBMC data, 99.83% for brain nuclei data and 98.96% for colon cell line mixture(Table 2.2).

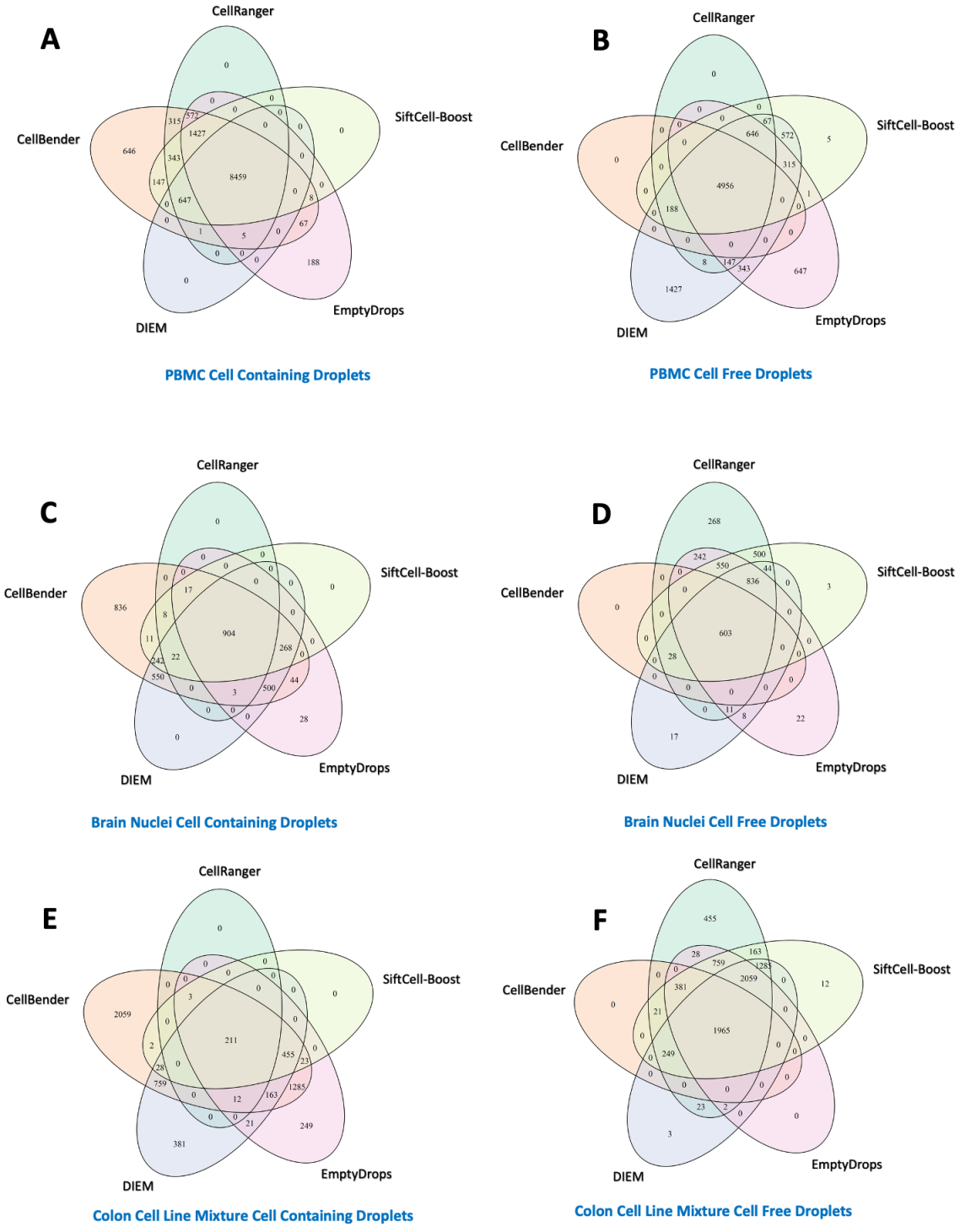

Figure 2.9: Venn Diagram of number of droplets that are classified as cell-containing or cell-free across all the methods. These panels contain full 5-way comparisons of the number of droplets classified each of the 5 methods, accounting all possible combinations. The Venn Diagram represents classifications of (A) cell-containing and (B) cell-free droplets for the PBMC dataset, (C) cell-containing and (D) cell-free droplets for the brain nuclei dataset, (E) cell-containing and (F) cell-free droplets for the colon cell line mixture dataset.

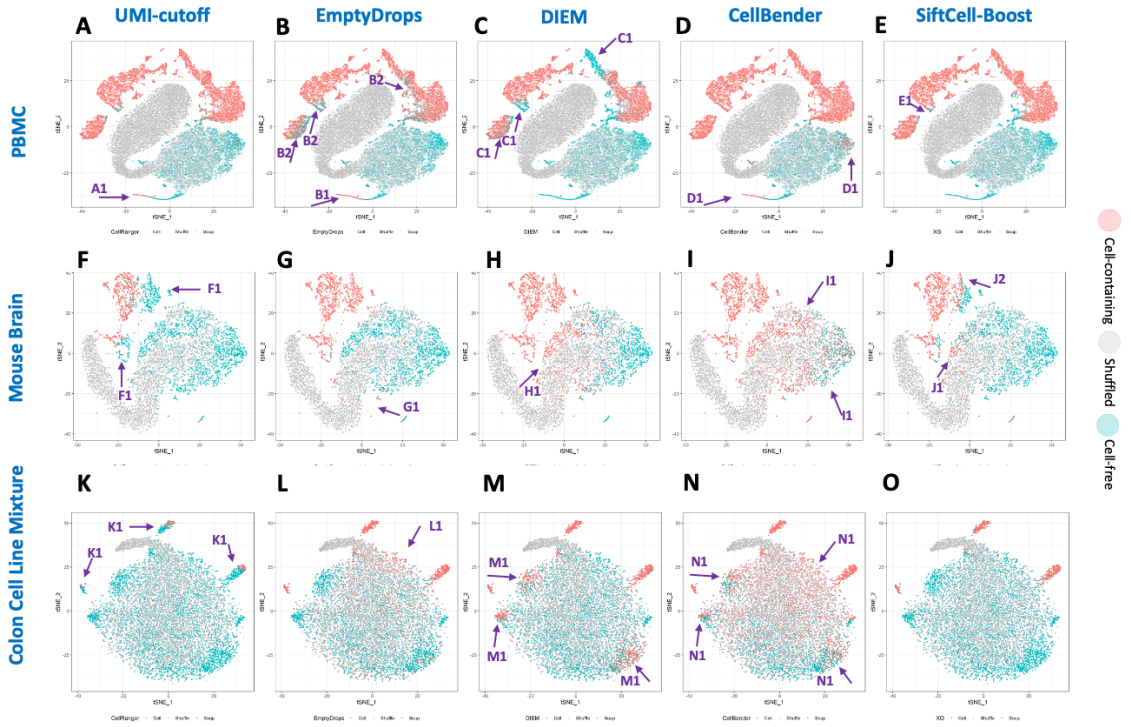

Figure 2.10: Annotation of comparison of cell filtering methods by *SiftCell-Shuffle*. This figure is identical with Figure 2.6, except that it is annotated (with arrows and labels) in multiple places to explain the difference between different methods. (A1) CellRanger/UMI-cutoff method classifies only a fraction of platelets as cell-containing droplets in PBMC dataset. (B1) EmptyDrops also classifies only a fraction of platelets as cell-containing droplets in PBMC dataset (B2, C1) EmptyDrops and DIEM classify some of likely true cell-containing droplets are cell-free droplets. (D1) CellBender classifies a fraction of platelets and a fraction of droplets with higher proportion of mt-RNAs as cell-free droplets. (E1) *SiftCell-Boost* classifies a small fraction of cell containing-droplets into cell-free ones. In brain nuclei dataset, (F1) CellRanger/UMI-cutoff method classifies a portion of potentially cell-containing droplets as cell free ones. EmptyDrops performs good in this case except a tiny proportion of cell-free droplets (G1) classified as cell-containing ones. (H1, I1, J1) DIEM, CellBender, and *SiftCell-Boost* classifies a fraction of cell-free droplets as cell-containing droplets. Also, (J2) *SiftCell-Boost* misclassifies cell-containing droplets as cell free ones. In colon cell line mixture, (K1) UMI-cutoff filters a large portion of cell-containing droplets. (L1, M1, N1) EmptyDrops, DIEM, and CellBender classifies a large proportion of cell-free droplets as cell-containing droplets.

Table 2.2: Accuracy and recall of *SiftCell-Boost* evaluated by cross-validation. This table represents the full results of 5-fold cross-validation from *SiftCell-Boost* across PBMC, brain nuclei and colon cell-line mixture datasets

Data	**Fold iteration**	**True Positives**	**False Positives**	**True Negatives**	**False Negatives**	**Accuracy(%)**	**Recall (%)**
PBMC	1	1132	0	2680	1	99.97	99.91
	2	1130	2	2678	3	99.86	99.74
	3	1131	0	2680	2	99.95	99.82
	4	1129	3	2678	3	99.84	99.73
	5	1132	2	2678	0	99.95	100.00
	Mean	**1132**	**1**	**2679**	**2**	**99.91**	**99.84**
Brain Nuclei	1	79	1	618	0	99.85	100.00
	2	78	0	619	1	99.85	98.73
	3	79	0	619	0	100.00	100.00
	4	78	1	618	1	99.71	98.73
	5	78	2	618	0	99.71	100.00
	Mean	**78**	**1**	**618**	**0**	**99.83**	**99.49**
Colon Cell Line Mixture	1	73	8	1283	9	98.76	89.02
	2	80	6	1285	2	99.41	97.56
	3	72	11	1280	10	98.47	87.80
	4	78	6	1284	8	98.98	90.36
	5	78	6	1284	5	99.20	93.98
	Mean	**76**	**7**	**1283**	**7**	**98.96**	**91.74**

2.2.4 *SiftCell-Mix* Estimates the Contribution from Ambient RNAs in Each Droplet

Even though classifying each droplet into two categories is practically useful to determine droplets for downstream analysis, it is reasonable to assume that each cell-containing droplet may also contain a certain amount of reads from ambient RNAs considering the overall procedure of droplet-based scRNA-seq experiment (*Heaton et al.* (2020); *Yang et al.* (2020); *Young and Behjati* (2020)). While *SiftCell-Boost* accurately classify cell-containing and cell-free droplets, it is important to estimate the proportion of ambient RNAs to inform downstream analysis. Once cell-containing droplets are clustered into cell types by users, *SiftCell-Mix* models the distribution of UMIs as a multinomial mixture of a single cell type and ambient RNAs to quantify contribution of ambient RNAs using maximum likelihood estimates (MLE). Across the three datasets – PBMC, brain nuclei, and colon cell line mixture – *SiftCell-Mix* corroborates the results from *SiftCell-Boost*, in the sense that the cell-containing droplets identified from *SiftCell-Boost* are estimated to have very small contribution from ambient RNAs, except for brain nuclei snRNA-seq that are expected to have contamination from ambient RNAs event for cell-containing droplets(Figure 2.11, Figure 2.12). Compared to DecontX with the default option, *SiftCell-Mix* provides more consistent estimates of % contribution from ambient RNAs across 3 datasets. While DecontX performed robustly for PBMC, it provided almost uniform estimates of % ambient RNAs across all droplets and failed to distinguish cell-containing and cell-free droplets. When *SiftCell-Mix* was compared to DecontX specifying the cell types as external variables, we observed that the results became much more similar than DecontX with default option. In *SiftCell-Mix*, it should be noted that not all cell-free droplets had high estimates of % ambient RNAs. We suspect that this is a result of multiple factors, such as non-random contribution from individual cell types to constitute ambient RNAs in specific droplets, systematic difference of mt-RNAs

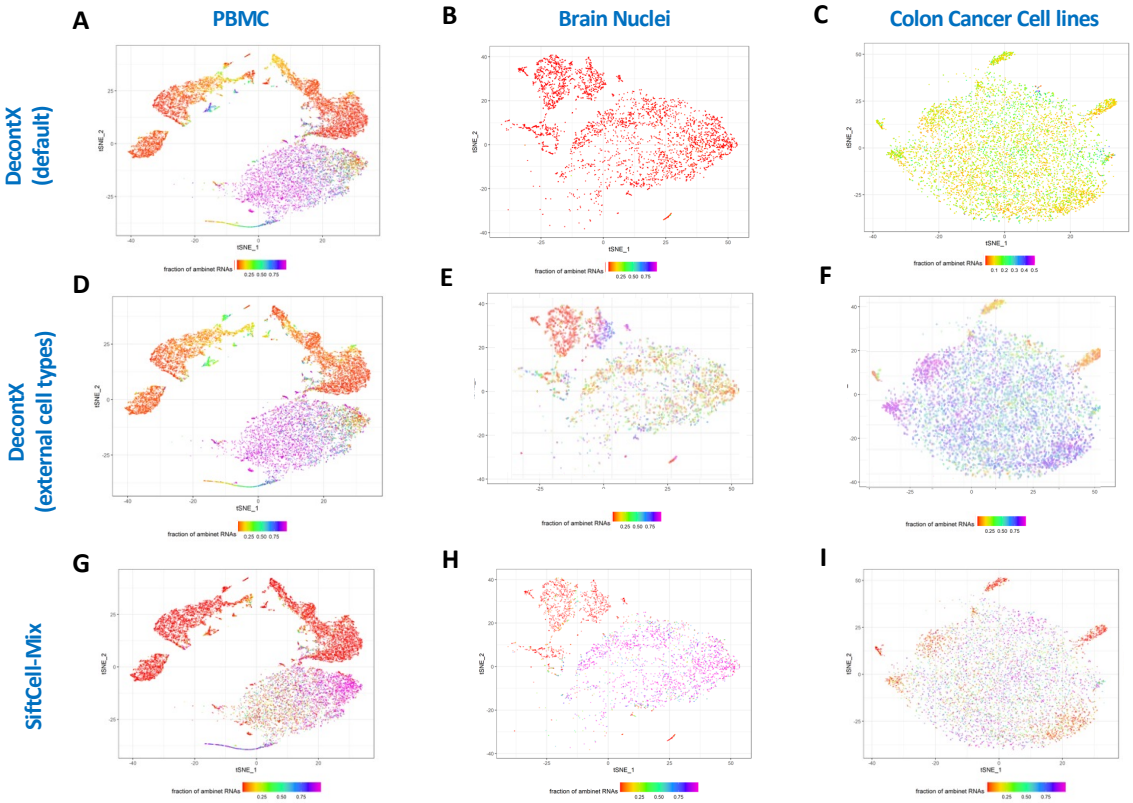

Figure 2.11: Visualization of contribution of ambient RNAs from scRNA-seq and snRNA-seq datasets. The six panels visualize the estimates of ambient RNA contamination in a linear scale among droplets in PBMC (A,D,G), brain nuclei (B,E,H), and colon cell line mixture (C,F,I) by DecontX with default option (A-C), DecontX with external cell types (D-F) and *SiftCell-Mix* (G-I) in the t-SNE manifold space excluding randomized droplets. Figure A and G show that the performance between DecontX with default option and *SiftCell-Mix* is comparable in PBMC dataset. In (B), DecontX with default option suggests that there is very little contamination of ambient RNAs in brain nuclei data, which is inconsistent to the expectation for typical snRNA-seq. DecontX with defualt option estimated that 0.2% of cell-free droplets (inferred by *SiftCell-Boost*) have >10% of ambient RNAs present. On the other hand, in (H), *SiftCell-Mix* suggests a large amount of ambient RNA contamination in the same data. *SiftCell-Mix* estimates that 81.8% of cell-free droplets have >10% ambient RNAs present. In colon cell line mixture, we do not expect a large contamination from ambient RNAs, However, in (C), DecontX with default option estimated that 56.8% of cell-containing droplets (inferred by *SiftCell-Boost*) have >10% of ambient RNAs present while, in (I), the estimation from *SiftCell-Mix* is only 9.9%. When *SiftCell-Mix* was compared to DecontX specifying the cell types as external variables, the results became much more similar than DecontX with default option.

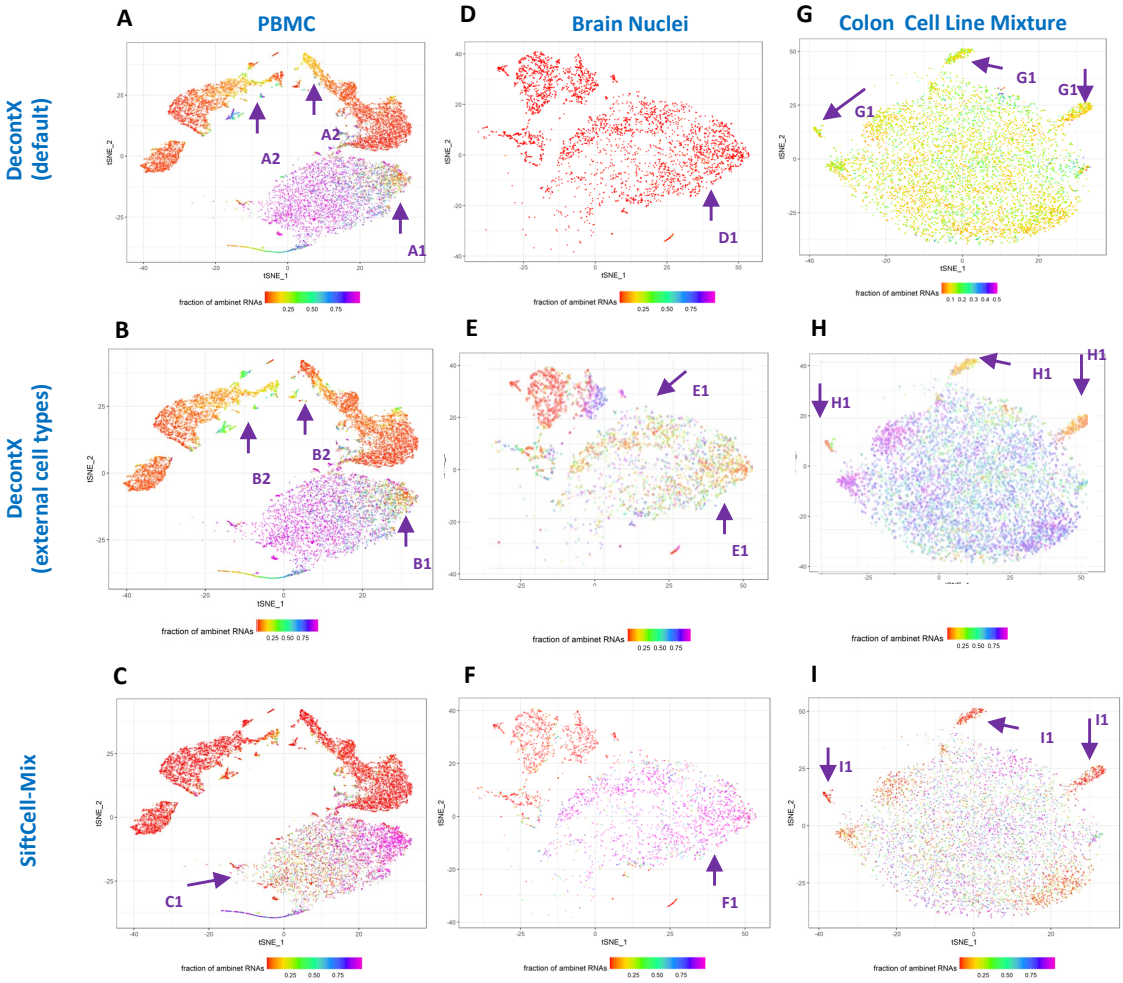

Figure 2.12: Annotated visualization of contribution of ambient RNAs from scRNA-seq and snRNA-seq datasets. This figure is identical with Figure 2.11, except that it is annotated (with arrows and labels) in multiple places to explain the difference between DecontX with default option, DecontX with external cell types and *SiftCell-Mix*. (A1, B1, C1) All the three methods estimated that a fraction of likely cell-free droplets have low proportion of ambient RNAs in PBMC data, although this is not necessarily unexpected. (A2,B2) In DecontX(default) and DecontX(with external cell type), there were a large fraction of likely cell-containing droplets that have elevated contribution from ambient RNAs. (D1)In brain nuclei, DecontX estimates that the proportion of ambient RNAs are very low for almost all droplets, which is unlikely. (E1) In DecontX (with external cell type), there were a large fraction of likely cell-free droplets that have low contribution from ambient RNAs. (F1) On the other hand, *SiftCell-Mix* show clear separation of the eastimated proportion of ambient RNAs between cell-containing and cell-free droplets. (G1) In colon cell line mixture, DecontX(default) estimated the contamination of ambient RNAs in cell-containing droplets to very similarly (10-30%) across all droplets. (H1, I1) On the other hand, DecontX with external cell type and *SiftCell-Mix* estimated ambient RNA contribution very differently between cell-containing and cell-free droplets.

by droplets (particular for snRNA-seq of brain nuclei), estimation errors due to low UMI counts in certain droplets.

2.2.5 Evaluation of Computational Cost

We evaluated the computational cost, in terms of wall time (i.e. elapsed time) and peak memory usage for *SiftCell* and other methods we evaluated above (Table 2.3). Both computational time and memory usage increased as the number of droplets increased across all methods evaluated. Each of the *SiftCell* methods typically finished the analysis within minutes. For the largest dataset (PBMC), *SiftCell* could take up to 15 minutes and consume up to 4.5GB of memory. Among the other methods, EmptyDrops and DecontX consumed a smaller memory footprint and computational time compared to *SiftCell*. DIEM was slower than *SiftCell* by a factor of 2-5. CellBender was evaluated in a GPU-enabled environment; nevertheless, its computational cost was the largest among all methods evaluated.

2.3 Materials and Methods

2.3.1 *SiftCell-Shuffle*: Visualizing Cell-free and Cell-containing Droplets in a Manifold Space

Our methods assume that we have a raw DGE matrix $X \in \{0, 1, 2, ...\}^{B \times G}$, where B is the total number of unique barcodes representing individual droplets, and G is the total number of genes or features. We assume that DGE matrix counts unique RNA molecules without redundancy using UMIs. Then $X_{b\cdot} = \sum_{g=1}^{G} X_{bg}$ is the total UMI counts per each barcoded droplet, and $X_{\cdot g} = \sum_{b=1}^{B} X_{bg}$ is the total number of reads covering each gene. The *SiftCell-Shuffle* algorithm takes an original DGE matrix X and outputs a permuted DGE matrix $X^{(S)}$ while preserving $X_{b\cdot}$ and $X_{\cdot g}$ to simulate cell-free droplets containing pseudo-bulk RNAs only to approximate ambient

Table 2.3: Table for computational and memory usage across all methods. Each row represents an evaluated method and each column represents either the memory usage or the wall time across the PBMC, brain nuclei and colon cell line mixture datasets. The computational experiments for EmptyDrops, DIEM, DecontX and *SiftCell* were conducted on a macOS Ventura system (version 13.2.1) with 1.4 GHz Quad-Core Intel Core i5 and 8 cores with 8GB of RAM. CellBender's computational cost was intractable in CPU-only machine. Therefore, we evaluated CellBender in a GPU-accelerated system in Google Colab Notebooks with 12GB of RAM. Due to the technical difficulty in tracking the memory footprint in the environment, only wall time was reported for CellBender.

Method	**PBMC**		**Brain Nuclei**		**Colon Cell Line Mixture**	
	Peak RAM (MB)	**Wall Time (Seconds)**	**Peak RAM (MB)**	**Wall Time (Seconds)**	**Peak RAM (MB)**	**Wall Time (Seconds)**
EmptyDrops	3437	155	1389	63	1943	76
DIEM	3578	2141	1197	223	2251	519
CellBender	N/A	4938	N/A	2978	N/A	3494
DecontX	3291	625	1620	78	2250	101
SiftCell-Shuffle	4459	118	1363	14	1897	25
SiftCell-Boost	4238	651	3162	112	2020	164
SiftCell-Mix	4231	994	1422	213	2407	416

RNAs. Specifically, let $U = \sum_{g=1}^{G} \sum_{b=1}^{B} X_{bg}$ be the total number of UMIs, and $J_u \in \{1, ..., B\} \times \{1, ..., G\}$ represents the (barcode, gene) pair each UMI belongs to, so that $X_{bg} = \sum I(J_u = (b, g))$is always true.

The *SiftCell-Shuffle* algorithm simply permutes the barcodes and genes in J_u independently (i.e. randomizes the relationship between barcodes and genes) across all Unique Molecular Identifiers (UMIs) to produce $J_u^{(S)}$; then the DGE matrix after *SiftCell-Shuffle* becomes $X_{bg}^{(S)} = \sum I(J_u^{(S)} = (b, g))$. As a result, the total number of UMIs for each barcode and each gene remains unchanged, because $X_b^{(S)} = X_b$ and $X_{\cdot g}^{(S)} = X_{\cdot g}$ hold as long as $J_u^{(S)}$is a permutation of J_u. The main idea of this procedure is that the distribution of UMI counts for each barcode in $X_{bg}^{(S)}$ is uniform, as if the droplet barcodes are randomly assigned from a bulk RNA-seq (i.e. aggregate of all reads ignoring barcode assignment), which we assume to represent the distribution of ambient RNAs.

To visualize whether each barcoded droplet likely contains ambient RNAs or not, we construct a low-dimensional manifold plots, such as UMAP or t-SNE, after combining X_{bg} and $X_{bg}^{(S)}$ into one DGE. *SiftCell-Shuffle* uses Seurat software with default parameters, except for no minimum UMI counts and 10 PCs, on the merged DGE matrix to generate t-SNE and UMAP manifolds and visualize it. The visualized manifold distinguishes the barcodes from X_{bg} and $X_{bg}^{(S)}$ in different colors (Figure 2.3 A,E,I). If a barcode contains ambient RNAs only, we expect it to appear proximal to $X_{bg}^{(S)}$ in the manifold space. For barcodes representing cell-containing droplets, we expect it to be located in a separate cluster from $X_{bg}^{(S)}$ (Figure 2.3 B,F,J). These plots allow us to quickly visualize how many cell-containing and cell-free droplets exist in a scRNA-seq or snRNA-seq dataset. When the randomized and original droplets are clustered together with Seurat(*Butler et al.* (2018)), the putative cell-free droplets tend to be assigned the same cluster label with shuffled droplets (Figure 2.3B). Visualizing the total UMI counts (Figure 2.3 C,G,K) or the proportion of mitochondrial

reads (Figure 2.3 D,H,L) also illustrate the cluster of shuffled droplets are enriched for lower total UMIs and high proportion of mitochondrial reads. This visualization was used to visually evaluate how well a specific quality control method classifies cell-containing and cell-free droplets across all datasets.

2.3.2 Evaluation of Existing Methods for Filtering Cell-containing Droplets with *SiftCell-Shuffle*

We evaluated existing methods for classifying cell-free droplets from cell-containing droplets by visualizing the results from each method in the t-SNE manifold plots generated by *SiftCell-Shuffle*. We used t-SNE instead of UMAP because it distributes the cell-free droplets more widely in the manifold space, which fits for the purpose of our evaluation. Four methods are used for evaluation: (1) CellRanger/UMI-cutoff method that determines cell-containing droplets based on a UMI count threshold, which is determined from knee plot (and a few other criteria), as implement in CellRanger 2 (`https://github.com/10XGenomics/cellranger`), (2) EmptyDrops (*Lun et al.* (2019)), implemented in DropletUtils R package(*Griffiths et al.* (2018)), which uses a likelihood-based permutation test to determine cell-containing droplets, (3) DIEM (*Alvarez et al.* (2020)), which uses E-M with Dirichlet distribution to identify droplets contaminated by ambient RNAs or extranuclear RNAs. (4) CellBender (*Fleming et al.* (2019)), which uses a generative model based on deep neural network, to identify cell-free and cell-containing droplets.

For UMI cutoff method, we used the default output from CellRanger 2 for PBMC and brain nuclei as they were generated from 10X Chromium. For colon cell line mixture, which is produced with DropSeq platform, we determined the UMI cutoff determined by the knee plot (UMI $\geq$ 5440, Figure 2.8). For EmptyDrops, which is expected to be similar to CellRanger 3, we used the default parameters for PBMC and brain nuclei, which is UMI$\leq$ 100 to represent ambient RNAs. For colon cell line

mixture, because UMI cutoff was high, we used UMI $\leq$ 200 to determine ambient RNAs. For DIEM and CellBender, we used the default parameters across all three datasets.

To illustrate the performance of each method with *SiftCell-Shuffle*, we visualized original and shuffled droplets in t-SNE spaces with three categories (1) shuffled droplets (2) original droplets classified as cell-containing droplets (3) original droplets classified as cell-free based on each algorithm (Figure 2.3, 2.6). This illustration is used to visually evaluate the performance of each algorithm to filter droplets.

We also developed a new metric, "% NN-concordance", as an alternative to the knee plot to estimate the number of cell-containing droplets by leveraging shuffled droplets. For each original droplet that is classified cell-containing, the nearest droplet (in terms of Euclidean distance of top 100 PCs of highly variable genes) among original + shuffled dataset is selected. The % NN concordance metric quantifies, across all filtered droplets, how often their nearest droplet is an original droplet as opposed to a shuffled droplets. This can be done for an arbitrary subset of droplets. This metric is intended to quantify how well the filtered droplets are separated from the shuffled droplets in a high-dimensional space.

2.3.3 *SiftCell-Boost*: Automated Machine Learning Method to Identify Cell-containing Droplets

SiftCell-Boost employs automated machine learning classification method(XGBoost) to classify each barcoded droplet into cell-containing (positive label) or cell-free (negative label) droplets using a training set consisting of permuted droplets from *SiftCell-Shuffle* and a subset of original droplets that are likely cell-containing droplets.

Given the absence of a definitive ground truth for droplet labels, we leveraged *SiftCell-Shuffle* result and proposed Sparse Quantile Aggregation Test(SQuAT) to generate labeled training data.

Assigning Negative Labels

To take advantage of *SiftCell-Shuffle* results, we used all permuted droplets from *SiftCell-Shuffle* as negative labels. Except for a few specified examples, we also include additional negative labels from original droplets based on the proportion of reads from unwanted genes (3 standard deviation above the median). The unwanted genes include all mitochondrial genes across 3 datasets. For PBMC, we also included *PPBP*, a marker gene for platelet cell type. This is to avoid classifying platelets, which is not supposed be a part of PBMC cell types, as cell-containing droplets.

Assigning Positive Labels

It is expected that cell-containing droplets in scRNA-seq experiments would exhibit higher levels of overdispersion compared to cell-free droplets due to the inherent variability in gene expression levels among cells. Existing statistical test $C(\alpha)$ (*Kim and Margolin* (1992)) test can be applied for identification of cell-containing dropelts with overdispersion, however, it generated inflated p-values among droplets with lower UMI counts, resulting in identifying false positive labels. When testing overdispersion in shuffled PBMC data, we expect to see no significant overdispersion in these droplets. As shown in the the QQ-plot (Figure 2.13 A), C-alpha test (*Kim and Margolin* (1992)) deviate greatly from the expected line. Such huge inflation can be attributed to the presence of extremely small gene profiles in the test parameters.

Therefore, we developed SQuAT (see Appendix for details) aimed at identifying potentially positive labels in sparse scRNA-seq datasets. SQuAT conducts bidirectional binomial overdispersion test using the interval of quantiles inputs instead of point estimators with variance adjustment of the test statistics. We compared the performance of SQuAT test with/without variance adjustment and $C(\alpha)$ test on shuffled (Figure 2.13) and original (Figure 2.14) PBMC data. SQuAT test with or without variance adjustment closely follow the expected line in the QQ-plot (Figure 2.13 A)

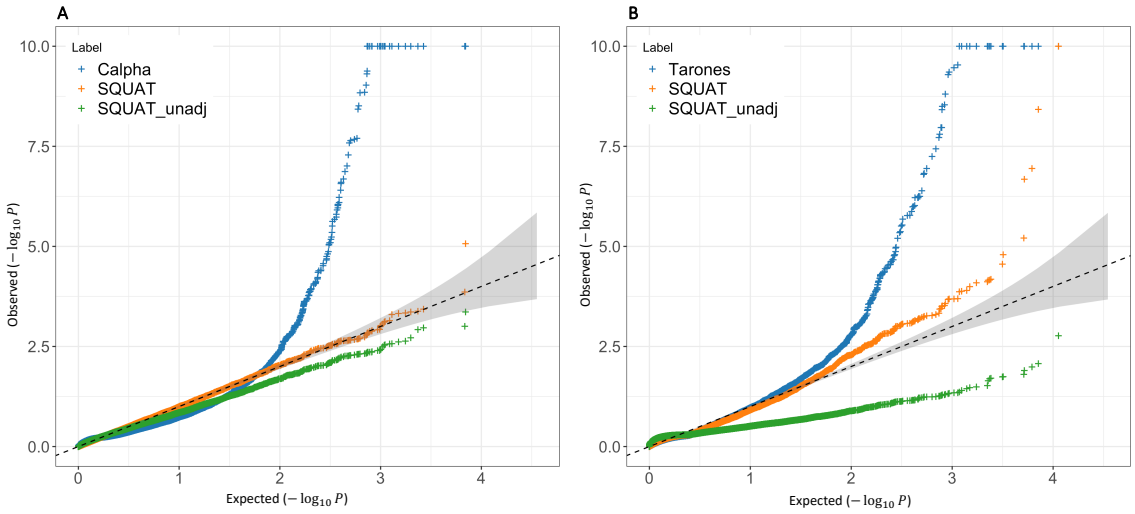

Figure 2.13: QQ plots of overdispersion tests on shuffled scRNA-seq data (i.e. under the null distribution). These two panels show the QQ-plot on shuffled PBMC data generated from *SiftCell-Shuffle*. Panel A compares C-alpha test (blue), SQuAT with (yellow) and without (green) variance adjustment to identify droplets with overdispersion. Panel B compares Tarone's test(blue), SQuAT with (yellow) and without (green) variance adjustment to identify highly variable genes.

and produce satisfactory results on shuffled PBMC whereas $C(\alpha)$ test deviate greatly from the expected line. We plotted the fraction of significant observations after Bonferroni correction against binned UMIs and compared the result of each method on shuffled and original datasets. For $C(\alpha)$ test, the statistical power increases with the number of UMI count, but at the cost of higher false positive rate for droplets with low UMI counts. On the other hand, the SQuAT test with variance adjustment is robust and efficient in detecting droplets overdispersion. Overall, the overdispersion in sparse scRNA-seq datasets can be addressed through bidirectional binomial overdispersion test, and the SQuAT test provided calibrated p-values with excellent performance, while C-alpha were found to be anti-conservative with low UMI count.

To generate highly confident positive labels for cell-containing droplets, we conducted non-parametric ranking among UMIs and z-scores derived from SQuAT with selected top N expected number of cells, provided by the user. In our experiment, we used $N = 10,000$ for PBMC, $N = 1,000$ for brain nuclei, and $N = 800$ for colon cell

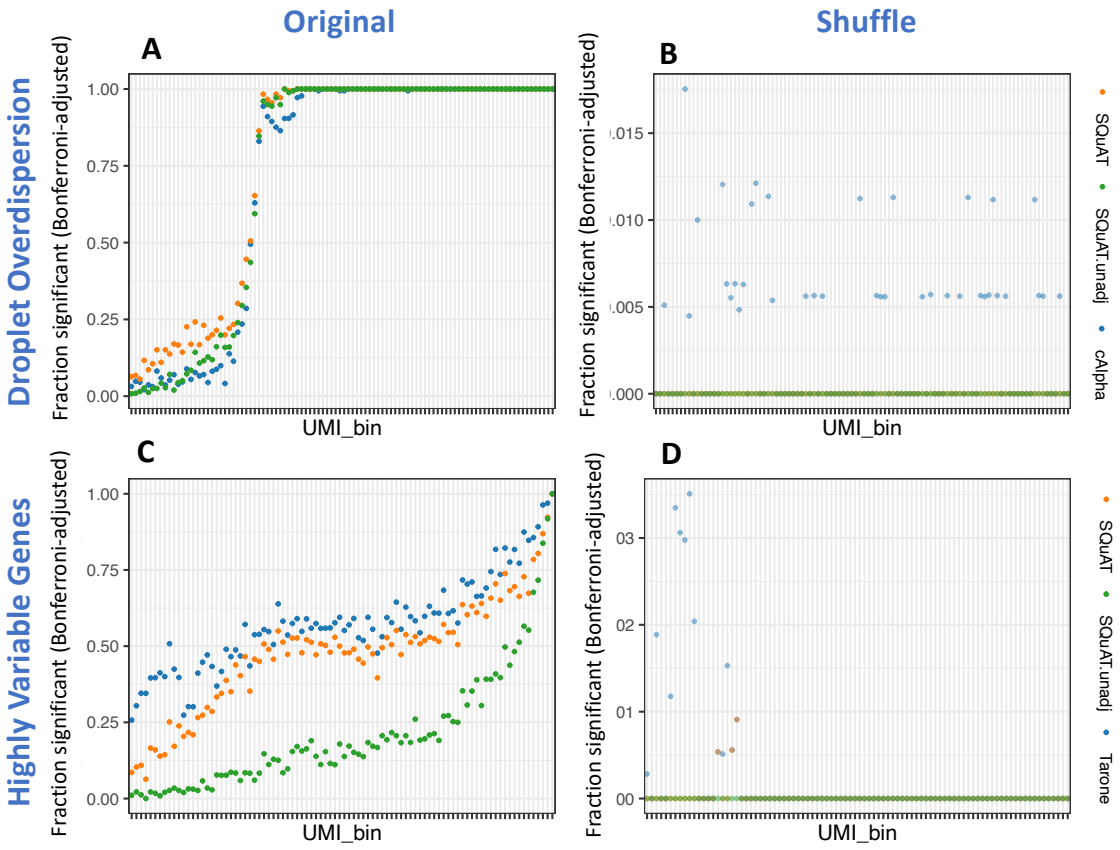

Figure 2.14: Evaluation of overdispersion tests on the original and shuffled PBMC scRNA-seq dataset, stratified by the UMI counts. These four panels show the scatter plot between the percentage of significant observations after Bonferonni correction against UMI groups (size 100 each) in ascending order. Panels A and B show the result of these methods in identifying droplets with overdispersion, respectively for the original and shuffled PBMC data. Similarly, Panels C and D present the results of these methods for identifying highly variable genes, separately for the original and shuffled PBMC data.

line mixture as suggested by CellRanger 2 or the published data.

Summary of Training Data

To summarize, the training dataset comes from three sources: (1) negative labels are obtained from randomized droplets (2) positive labels are obtained from confident cell-containing droplets estimated from SQuAT, and (3) additional negative labels are obtained from the original droplets based on excessive contribution from mtRNAs and/or enrichment of marker genes representing unwanted cells (e.g. Platelet in PBMC datasets). The test data is the rest of the unlabeled droplets which is not part of the training data.

Classification Model via Extreme Gradient Boosting (XGBoost)

SiftCell-Boost uses XGBoost to train the classification model with the positively and negatively labeled droplets. To generate the features for XGBoost training, We used $1,000$ most variable ones identified by SQuAT to generate top 100 principal components from the log-normalized digital expression matrix. Note that similarly, detection of highly variable genes can also be achieved by applying SQuAT on to the scRNA-seq data(Figure 2.13 B and 2.14 C and D).

Using *SiftCell-Boost*, we classified original droplets into cell-containing and cell-free droplets and evaluated its performance with other methods (Figure 2.6), using t-SNE visualization from *SiftCell-Shuffle*, as well as %-NN concordance metrics(Figure 2.7). %-NN-concordance metrics were evaluated for arbitrary UMI cutoff, as well as for the 5 methods to filter cell-containing droplets.

2.3.4 *SiftCell-Mix*: Model-based Approach for Inferring the Fraction of Ambient RNAs in Each Droplet

It is possible that some of the reads in a droplet are "contaminated" by ambient RNAs floating outside individual cells(*Heaton et al.* (2020); *Yang et al.* (2020)). For cell-free droplets, most reads are contributed by ambient RNAs, but even for cell-containing droplets, ambient RNAs may present within barcoded reads assigned for the droplets. We assume that the read counts for a droplet follow a mixture of multinomial distributions, one for each cell type and an additional one representing ambient RNAs. Prior to *SiftCell-Mix* analysis, we assume that a large fraction of cell-containing droplets are assigned to specific cell types using other software tools (e.g. Seurat or scanpy) or by an domain expert so that the distribution of each cell type can be modeled reliably.

Let $n_1, n_2, \ldots, n_K$ be the number of droplets assigned to each of the K cell types and $n_0 = N$ be the total number of droplets in the single cell RNA-seq dataset, including cell-containing and cell-free droplets. Let $D_0, D_1, \ldots, D_k$ denote the set of droplets corresponding to $n_0, n_1, \ldots, n_k$, respectively. For a given droplet $i \in \{1, 2, \ldots, N\}$, the reads count across the G genes (with nonzero read count) is a vector: $x_i = (x_i^1, x_i^2, \ldots, x_i^G)$. Let $\pi_k = (\pi_k^1, \pi_k^2, \ldots, \pi_k^G)$ for $k \in \{0, 1, \ldots, K\}$ be the multinomial probabilities representing the distribution of each cell type (or ambient RNAs for $k = 0$). We model x_i as a multinomial mixture between ambient RNAs (π_0) and one of the cell types ($\pi_k, k > 0$) as we will describe later. For $k > 0$, we define π_k as an arithmetic mean of the proportion of reads of the gene across the droplets of the cell type k:

$$\pi_k^j = \frac{1}{n_k} \sum_{i \in D_k} \left[\frac{x_i^j}{\sum_{g=1}^{G} x_i^g} \right], j \in \{1, 2, \ldots, G\}$$

.

We define π_0 in a similar way, but across all droplets regardless of their cell types, slightly up-weighting droplets with high total UMI count, but ensuring minimum weight λ ($\lambda = 100$ in our experiments) for droplets with low UMI counts according to $w_i = \min(\lambda, \sum_{g=1}^{G} x_i^g)$. We weight π_0 based on $\log w_i$ to better represent the distribution of ambient RNAs enriched for low-UMI count droplets:

$$\pi_0^j = \frac{\sum_{i \in D_k} \left[\frac{x_i^j}{\sum_{g=1}^{G} x_i^g} \log w_i \right]}{\sum_{i \in D_k} \log w_i}, j \in \{1, 2, ..., G\}$$

To avoid the corner case that $\pi_k^j = 0$ for some (j, k), we adjust π_k^j to contain a small fraction (α) of ambient RNAs as $(1 - \alpha)\pi_k^j + \alpha\pi_0^j$, and used $\alpha = 0.01$ in our experiments. In summary, $\pi_1, \pi_2, \ldots, \pi_K$are defined as arithmetic mean of reads within each cell type, and π_0 is defined logarithmic mean of reads across all droplets with a threshold.

We model the log likelihood of the read count in a droplet as a mixture of multinomial distributions of ambient RNAs and one of the cell types. Let γ_{ik} be the fraction of contributions from the kth category to the ith droplet where $k = 1, 2, \ldots, K$. Thus, the objective function is the log likelihood of the reads coming from the different cell types and ambient RNAs, which can be formulated as:

$$f_i = -\log \left[\sum_{k=0}^{K} \gamma_{ik} \exp(\sum_{j=1}^{G} x_i^j log \pi_k^j) \right]$$

Subject to :

$$\sum_{k=0}^{k} \gamma_{ik} = 1$$

$$0 \leq \gamma_{ik} \leq 1, \forall k \in \{0, 1, \ldots, K\}, \text{ and } \forall i \in \{1, 2, \ldots, N\}$$

The nonlinear optimization problem is solved using augmented Lagrange multiplier method with an sequential quadratic programming interior algorithm as imple-

mented in Rsolnp package (v1.16) available in CRAN.

2.4 Summary

In this chapter, we describe *SiftCell* framework, a suite of software tools implementing methods including *SiftCell-Shuffle*, *SiftCell-Boost* and *SiftCell-Mix*, focusing on the challenges of contaminations from ambient RNAs in single-cell and single-nucelus RNA-seq experiments. *SiftCell-Shuffle* works with DGE matrix and aids the investigators to visually distinguish cell-free and cell-containing droplets by contrasting with a randomized digital expression matrix. *SiftCell-Boost* takes the output of *SiftCell-Shuffle* as input and applies a machine learning method to classify cell-containing droplets and cell-free droplets. *SiftCell-Mix* is a model-based tool that allows quantitative estimation the contribution of "ambient RNAs" in each droplet.

We believe that *SiftCell* will facilitate more holistic understanding of scRNA-seq from upstream and reduce the chance that upstream technical issues such as ambient RNA contamination obscure novel scientific discovery.

2.5 Appendix: Sparse Quantile Aggregation Test (SQuAT)

Aggregating summary statistics from multiple datasets into a single meta-analyzed statistic has been widely useful in multiple areas of scientific research for more than a century (*Shannon*, 2016; *Borenstein et al.*, 2021). Most widely-used meta-analysis methods such as Fisher's method of combining p-values(*Fisher*, 1925), Liptak and Stouffer's method(*Lipták*, 1958; *RMJ*, 1949) of weighted sum of z-scores, and inverse-variance-weighted meta-analysis(*Fleiss*, 1993) rely on asymptotic distributions that assumes a large sample sizes. When asymptotic distribution approximates the underlying data well, it is shown that meta-analysis is as powerful as jointly analyzing individual-level data together(*Lin and Zeng*, 2010).

Recently, the volume of sparse data available for scientific research is rapidly increasing with technological advances. For example, in single-cell or spatial genomics data, a small fraction of genes have informative reads per barcode, which represents an individual cell or a spatial location, so the data is typically highly sparse(*Zheng et al.*, 2017; *Cho et al.*, 2021). When meta-analyzing sparse scRNA-seq datasets, asymptotic approximation may no longer hold. In such cases, it has been demonstrated that the estimated effect may be biased and the estimated variance may be misleading(*Martin and Austin*, 2000; *Richardson et al.*, 2021). Meta-analysis methods based on exact p-value or continuity correction may improve its accuracy but benefits of such heuristics are still limited(*Rubin-Delanchy et al.*, 2019; *Liu et al.*, 2014; *J. Sweeting et al.*, 2004).

In this Appendix, we propose a new framework for meta-analysis that can capture the uncertainty inherent in the sparse data, and introduce SQuAT (Sparse Quantile Aggregation Test) as an efficient implementation of the proposed framework. The framework uses an interval of quantiles as inputs for meta-analysis. The interval of quantiles is then projected onto a continuous distribution to be meta-analyzed accounting for its uncertainty represented by the interval. We apply SQuAT in the meta-analysis of binomial overdispersion test in sparse single-cell RNA-sequencing dataset to demonstrate the practical utility of our approach.

SQuAT Framework

More specifically, Let $x \in \mathcal{X}$ be the observed data generated from a, possibly sparse, discrete distribution with probability mass function $f(x) = \Pr(X = x)$ and cumulative density function $F(x) = \Pr(X \leq x)$.

Let $g(y)$, $G(y)$, and $G^{-1}(y)$ be the probability density function, cumulative distribution function, and inverse cumulative distribution function of a continuous distribution $\mathcal{G}$, respectively. Define

$$g(y; a, b) = \frac{1}{G(b) - G(a)} g(y) \qquad (a < y \leq b)$$

be the truncated distribution of $\mathcal{G}$ conditional on $y \in (a, b]$.

When we observe a discrete value x, we consider a random variable $Y(x)$ that follows the truncated distribution $g\left(y; G^{-1}(F(x-1)), G^{-1}(F(x))\right)$, which corresponds to the same interval of quantiles associated with x in $F(\cdot)$.

Bidirectional SQuAT with Normal Distribution

Using quantile projection described above, consider a specific case where $g(y)$ is a normal distribution. Let $g(\cdot) = \phi(\cdot)$ be the standard normal probability density function, $G(\cdot) = \Phi(\cdot)$ be the its cumulative distribution function. Then for a random variable $X \in \mathcal{X}$, the marginal distribution of $Y(X)$ follows $\mathcal{N}(0, 1)$.

Therefore, given observations $x_1, \ldots, x_k$, and known weights $w_1, \ldots, w_k$, Liptak-Stouffer meta-analyzed Z-score

$$Z(Y_1(x_1), \ldots, Y_k(x_k)) = \frac{\sum_{i=1}^{k} w_i Y_i(x_i)}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

will be marginally distributed $\mathcal{N}(0, 1)$. However, the problem with $Z(Y)$ is that $Z(Y)$ is a random variable that does not have fixed value given x. Therefore the meta-analyzed statistic does not provide a deterministic value.

To obtain a deterministic meta-analyzed statistic, we define $S(x)$ as an expectation of $Y(x)$, we define a basic unit of SQuAT: $S(x)$, an expectation of $Y(x)$, then we have

$$S_b(x) = \begin{cases} \dfrac{\phi(\Phi^{-1}(2F(x))) - \phi(\Phi^{-1}(2F(x-1)))}{2f(x)} & F(x) \leq \frac{1}{2} \\ \dfrac{\phi(\Phi^{-1}(2F_c(x-1))) - \phi(\Phi^{-1}(2F_c(x)))}{2f(x)} & F(x-1) \geq \frac{1}{2} \\ -\dfrac{\phi(\Phi^{-1}(2F(x-1))) + \phi(\Phi^{-1}(2F_c(x)))}{2f(x)} & F(x-1) < \frac{1}{2} < F(x) \end{cases}$$

where $F_c(x) = 1 - F(x)$. We can show that $\mathrm{E}\left[S_b(X)\right] = 0$. $\mathrm{Var}\left[S_b(X)\right] = \mathrm{E}\left[S_b^2(X)\right]$ follows

$$\begin{aligned} \mathrm{Var}\left[S_b(X)\right] &= \mathrm{E}\left[S_b^2(X)\right] = \sum_{x \in \mathcal{X}} S^2(x) f(x) \\ &= \sum_{\forall x,\ F(x) \leq \frac{1}{2}} \frac{\left[\phi(\Phi^{-1}(2F(x))) - \phi(\Phi^{-1}(2F(x-1)))\right]^2}{4f(x)} \\ &\quad + \sum_{\forall x,\ F(x-1) \geq \frac{1}{2}} \frac{\left[\phi(\Phi^{-1}(2F_c(x-1))) - \phi(\Phi^{-1}(2F_c(x)))\right]^2}{4f(x)} \\ &\quad + \sum_{\forall x, F(x-1) < \frac{1}{2} < F(x) \geq \frac{1}{2}} \frac{\left[\phi(\Phi^{-1}(2F_c(x))) + \phi(\Phi^{-1}(2F(x-1)))\right]^2}{4f(x)} \\ &= \sum_{x \in \mathcal{X}} V_b(x) \end{aligned}$$

SQuAT in scRNA-seq

We incorporated SQuAT as a statistical tool within *SiftCell* to estimate how likely each original droplet contains a cell and to identify highly variable genes. As previously discussed, *SiftCell-Boost* is a machine learning based method to distinguish cell-containing and cell-free droplets. During the training process, Positive labels of cell-containing droplets and features of highly variable genes are determined through the utilization of SQuAT by identifying genes or droplets that deviates from the expectation from the null distribution.

To assess the performance of the proposed framework on scRNA-seq data, we first applied bidirectional SQuAT on the shuffled PBMC dataset generated by *SiftCell-Shuffle* and assume that the shuffled data is under a Binomial distribution. We ran SQuAT on two different types of test. First is to determine the over dispersion of each droplet . In this case, each gene was the component for SQuAT. Second is to determine the highly variable genes, and each droplet was the component for SQuAT. In both tests, we used observed UMI count as the input, and the marginal sum of UMI (per droplet or per gene) was used as the total count for the binomial distribution. We compared the performance of three tests: SQuAT test with and without variance adjustment, and Tarone's test (*Tarone* (1979)) or the C-alpha (*Kim and Margolin* (1992)) test, a generalization of Tarone's test, to determine droplets with significant overdispersion and genes with high variability (Figure 2.13). The QQ-plots revealed that both C-alpha and Tarone's test deviate greatly from the expected line. This huge inflation can be attributed to the presence of extremely small gene profiles in the test parameters. As shown in Figure 2.13 A, the variance adjustment did not impact the test results much in identifying droplets with overdispersion. Both the tests closely followed the expected line and produced satisfactory results on shuffled data. In Figure 2.13 B when testing for highly variable genes, SQuAT test without variance adjustment exhibited great deflation. However, after adjusting for the variance, the SQuAT produced slightly inflated test statistics. Since SQuAT for highly variable genes was used to select the top $1,000$ genes, we determined that the slight level of inflation was acceptable, but there are rooms for improvement in SQuAT.

Similiarly, we then applied these methods to the original PBMC dataset, and the results are depicted in Figure 2.14. We plotted the fraction of significant observations after Bonferroni correction against binned UMIs and compared the result between shuffled and original datasets. For Tarone's test, the statistical power increases with the number of UMI count, but at the cost of higher false positive rate

when UMI count is low. On the other hand, the SQuAT test with variance adjustment was robust and efficient in detecting overdispersion. It was concluded that the overdispersed droplets or highly variable genes in sparse scRNA-seq datasets can be identified through meta-analysis of bidirectional binomial data, and the SQuAT test provided excellent performance, while C-alpha and Tarone's tests were found to be anti-conservative with low UMI count.

CHAPTER III

Microscopic Examination of Spatial Transcriptome Using Seq-Scope

3.1 Background

Standard immunohistochemistry and RNA in situ hybridization can examine only one or a handful of target molecular species at a time; therefore, the amount of information obtained from a single experimental session is limited. To overcome this, emerging spatial transcriptomics (ST) techniques aim to examine all genes expressed from the genome from a single histological slide (*Asp et al.* (2020)). There are three major methodologies to experimentally implement ST. First, the sequential in situ hybridization method, often combined with combinatorial multiplexing, can increase the number of RNA species that can be detected from a single histological section. Second, in situ sequencing can identify RNA sequences from the tissue through fluorescence-based direct sequencing. Finally, spatial barcoding methods associate RNA sequences and their spatial locations by capturing tissue RNA using a spatially barcoded oligonucleotide array. Among these three major methodologies, the spatial barcoding method is the most straightforward, comprehensive, widely used, and commercially available method easily accessible by many laboratories (*Asp et al.* (2020)). Spatial barcoding technologies have the potential to reveal histological

details of transcriptomic profiles; however, they are currently limited by their low resolution. For example, VISIUM from 10X Genomics has a center-to-center resolution of 100 m (*Bergenstråhle et al.* (2020)), which is worse than that of the naked eye (40 μm). More recent technologies, such as Slide-Seq, HDST, and DBiT-Seq, improved the resolution (*Liu et al.* (2020); *Rodriques et al.* (2019); *Stickels et al.* (2021); *Vickovic et al.* (2019)); however, their resolutions are still far coarser than an optical microscope that has submicrometer resolution.

Here, we describe a technology for achieving submicrometer resolution spatial barcoding, designated as Seq-Scope that is based on the solid-phase amplification of a random barcode molecule, conveniently achieved by the Illumina. We also conduct computational analysis of Seq-Scope data to reveal transcriptomic heterogeneity at the cellular and subcellular level in various tissues.

3.2 Results

Seq-Scope is initiated by generation of a single stranded oligonucleotide library that has a randomly generated spatial barcode sequence. We name this barcode a high-definition map coordinate identifier (HDMI). The HDMI oligos are amplified on a solid surface, generating clusters with unique HDMI sequences. The HDMI-array was produced with a sequenced cluster density of up to 1.5 million clusters per $\mu m2$, which is sufficient to perform single-cell and subcellular analysis of the spatial transcriptomics. Each cluster's HDMI sequence and its spatial coordinates are determined through Illumina sequencing. We name this process the 1st-Seq. Then each cluster is processed to capture RNAs released from the overlying tissue sections. Both HDMI and cDNA sequences are determined through the 2nd-Seq process.

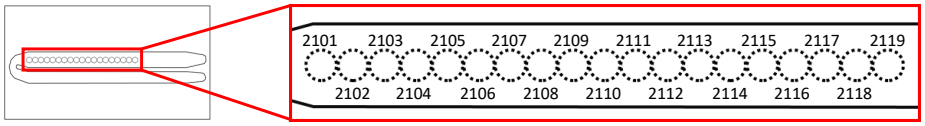

Figure 3.1: Schematic diagram depicting tile arrangement in MiSeq regular flow cell

3.2.1 Seq-Scope Has an Outstanding Transcriptome Capture Performance

Complete Seq-Scope procedure were performed on two representative gastrointestinal tissues, the liver and colon. In each 1st-Seq experiment, the HDMI-array was produced in $1mm$-wide circular areas of the MiSeq flow cell, also known as 'tiles' (Figure 3.1). The tissue sections were overlaid onto the HDMI arrays, examined by H&E staining, and subjected to 2nd-Seq. Analysis of the 1st-Seq and 2nd-Seq data demonstrated that the RNA footprints were discovered mostly from tissue-overlaid regions(Figure 3.2), confirming that Seq-Scope can indeed capture and analyze the spatial transcriptome from the tissues.

3.2.2 Seq-Scope Captures Transcriptome Information with High Efficiency

Benchmark analysis demonstrated that Seq-Scope offers a dramatic improvement in resolution and pixel density compared to previous ST solutions (Figure 3.3); center-to-center distances between HDMI pixels were measured to be $0.633 \pm 0.140 \mu m$ (liver) and $0.630 \pm 0.132 \mu m$ (colon) (mean $\pm$ SD) . Although each HDMI-barcoded cluster covers an extremely tiny area ($< 1 \mu m^2$), single HDMI pixel in tissue-covered region was able to capture 6.70 ± 5.11 (liver) and 23.4 ± 17.4 (colon) UMIs (mean $\pm$ SD) (Figure 3.3 C). The number of gene features identified per HDMI pixel was 5.88 ± 4.22 (liver) and 19.7 ± 14.3 (colon) (mean $\pm$ SD) (Figure 3.3 D). Per-pixel counts of UMIs and genes in Seq-Scope were larger than HDST but were smaller than other technologies(Figure 3.3 C and D).

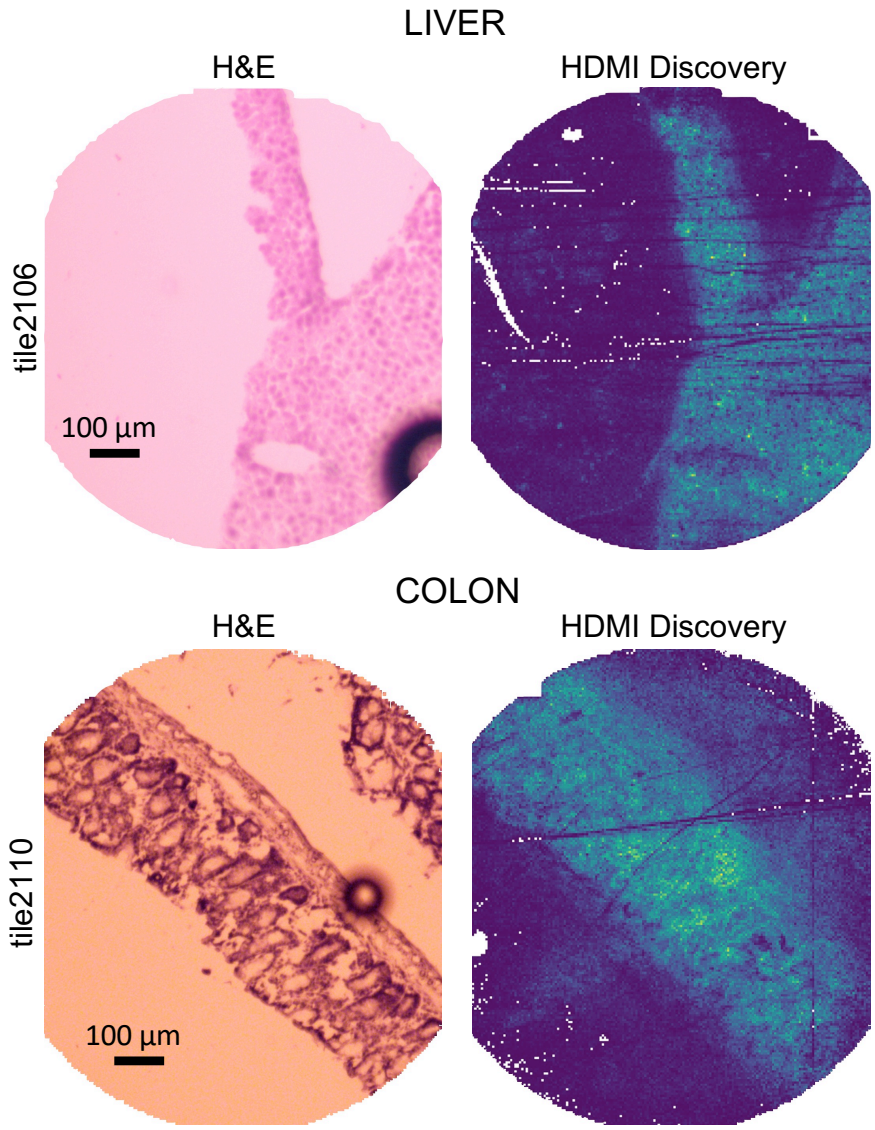

Figure 3.2: Seq-Scope Capture Performance. H&E staining and its corresponding HDMI discovery plot drawn from the analysis of 1st-Seq and 2nd-Seq outputs. Brighter color in the HDMI discovery plot indicates that more HDMI was found from 2nd-Seq in the corresponding pixel area.

However, after normalization using the pixel density, Seq-Scope showed the best transcriptome capture performance per area among the datasets we examined (Figures 3.3 E and F; colon dataset). Considering that the current data are estimated to cover only $\sim 60\%$ (liver) and $\sim 36\%$ (colon) of the total library size, the maximum possible Seq-Scope capture efficiency should be even higher than the currently presented data. Therefore, Seq-Scope provides an outstanding mRNA capture output, in addition to providing an unmatched spatial resolution output.

3.2.3 Seq-Scope Reveals Nuclear-cytoplasmic Transcriptome Architecture from Tissue Sections

mRNA is transcribed and poly-A modified in the nucleus, and transported to the cytoplasm after splicing (Figure 3.4 A). Several RNAs in the mouse liver, such as *Malat1*, *Neat1*, and *Mlxipl*, exhibit strong nuclear localization (*Halpern et al.* (2015)). On the other hand, the cytoplasmic mitochondria contain many mitochondria-encoded RNAs (mtRNA) (Figure 3.4 A).

We spatially plotted all spliced and unspliced transcripts discovered from Seq-Scope mouse liver data and it reveals nuclear-cytoplasmic transcriptome architecture from tissue sections. Unspliced transcript expression was restricted in tiny circles with a diameter of $10\mu m$ (Figure 3.4 B), which is about the size of hepatocellular nuclei (*Baratta et al.* (2009)) in liver data. Spliced mRNAs were relatively scarce in the unspliced area , whereas nuclear-targeted RNAs were more abundant in the unspliced area (3.4 B). Mt-RNAs were mostly in the spliced area (Figure 3.4 C). These observations were substantiated by correlation analysis of the single-cell images (Figures 3.4 D).

These results suggest that spliced and unspliced transcripts are useful to determine the nuclear-cytoplasmic structure from the Seq-Scope dataset. Indeed, when overlaid with H&E staining images, the unspliced RNA-enriched region generally agreed with

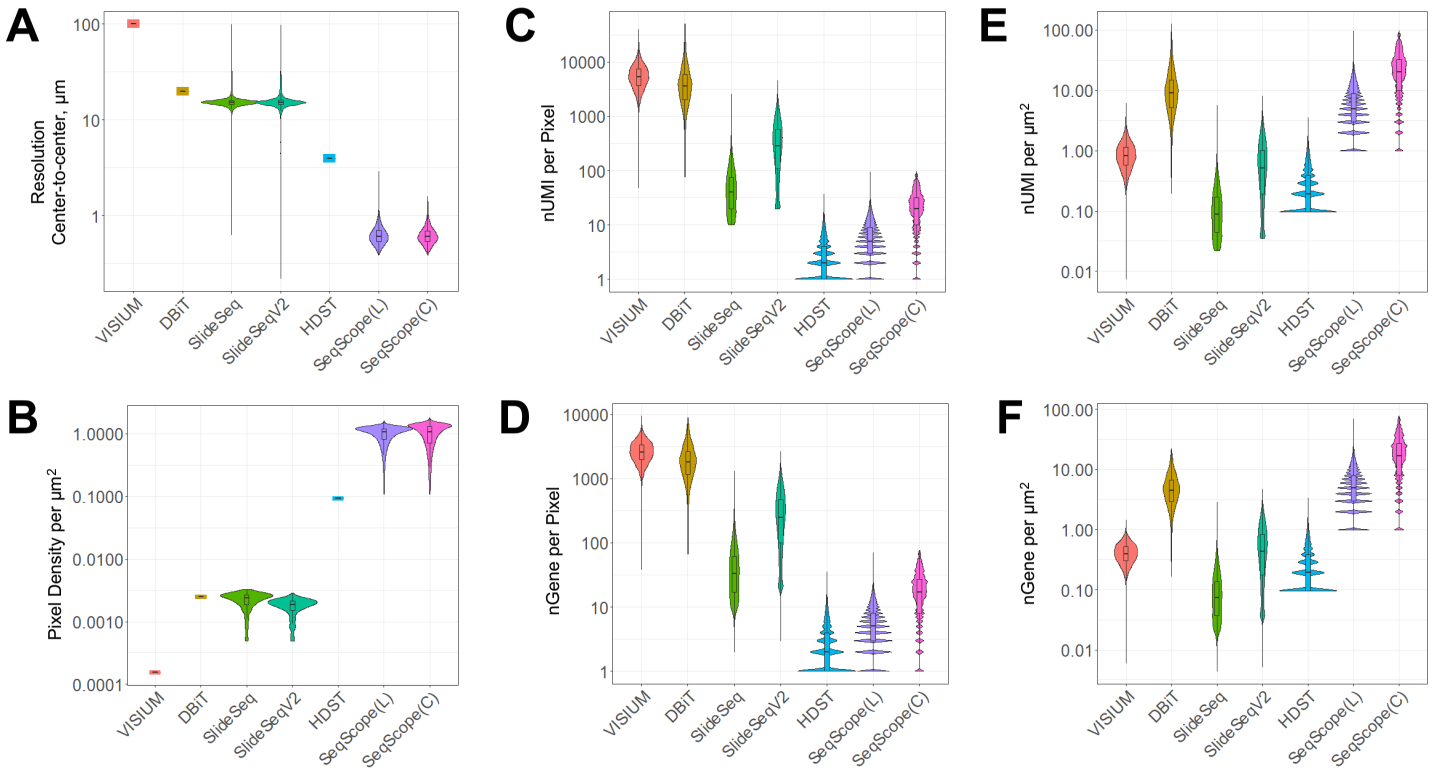

Figure 3.3: Benchmark Analysis. (A–F) Performance comparison of different ST solutions. The values were derived from each pixel (A and C–F or gridded area (B). nUMI, number of UMI; nGene, number of gene features; Seq-Scope(L) and Seq-Scope(C), liver and colon Seq-Scope data.

the nuclear position (Figure 3.4 E; note that some hepatocytes are known to be multinucleate) (*Donne et al.* (2020)). However, in some hepatocytes, the unspliced RNA-enriched regions were not observed (Figure 3.4 E), which can be explained by the absence of the cell's nucleus in the tissue slice, the inadequate positioning of the nucleus for RNA capture or the intrinsic variations in the rates of transcription, splicing, and nuclear export.

To further test the robustness of these observations, we randomly divided all genes into three independent subsets and examined the expressions of spliced and unspliced mRNAs from each subset. All three datasets similarly visualized a nuclear-cytoplasmic structure with a strong correlation (Figures 3.4 F).

Finally, we identified nuclear centers by using unspliced transcripts(Figure 3.4 G). Then, we searched for genes whose transcripts were enriched within $5\mu m$ from the nuclear centers. Consistent with previous cell fractionation and RNA in situ hybridization studies (*Halpern et al.* (2015)) and our observations described above, *Malat1*, *Neat1*, and *Mlxipl* were identified as the top 3 genes enriched in the nuclear area (Figure 3.4 H). These results demonstrate that Seq-Scope can perform subcellular transcriptome studies.

3.2.4 Seq-Scope Performs Spatial Single-cell Analysis of Hepatocytes

Using an image segmentation method (*Sage and Unser* (2003)), single hepatocellular areas were identified from the H&E image (Figures 3.4 E and 3.5 A). The single hepatocellular transcriptome from the segmented Seq-Scope data showed a substantial number of UMIs ($4,294$, median; $4,734 \pm 2,480$, mean $\pm$ SD) and genes ($1,617$, median; $1,673 \pm 631.7$, mean $\pm$ SD), which are comparable to the recent single hepatocyte transcriptome datasets obtained from MARS-Seq (*Halpern et al.* (2015)) and Drop-Seq (*Park et al.* (2021)) (Figure 3.5 B). The transcriptome content of Seq-Scope was similar to the results from the MARS-Seq, Drop-Seq, and bulk RNA-seq analyses

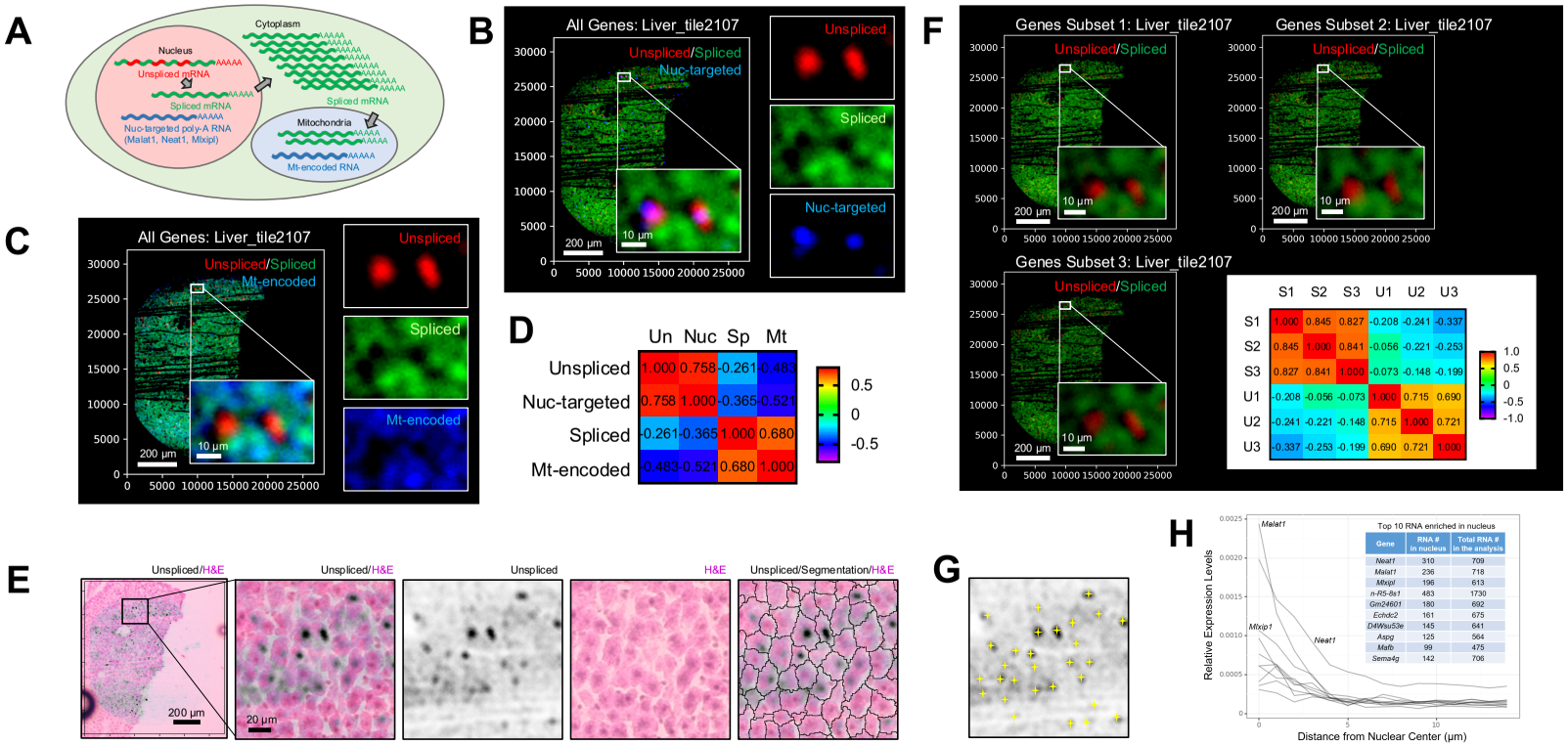

Figure 3.4: Seq-Scope visualizes subcellular spatial transcriptome. (A) Schematic diagram depicting the distribution of different RNA species in subcellular compartments.(B–D) Spatial plot of all unspliced and spliced transcripts, as well as nuclear-targeted (B) and mitochondria-encoded (C) transcripts. Pearson correlations (r) between these transcript intensities were presented in a heatmap (D). (E) Images displaying unspliced RNA discovery, H&E histology, and histology-based cell segmentation boundaries. Inset in the first panel is magnified in right panels. (F) Spatial plot of unspliced and spliced transcripts in three independent subsets of genes (gene subset 1–3). Pearson correlations (r) were presented as a heatmap. S1–3, spliced 1–3; U1–3, unspliced 1–3. (G) Identification of transcriptomic nuclear centers (yellow crosses) through local maxima detection. (H) Identification of nuclear-enriched RNA species. Top 10 nuclear-enriched RNAs are shown.

of the normal liver.

Cell type mapping analysis of the segmented single hepatocyte dataset revealed the spatial structure of hepatocellular zonation, identifying both pericentral (PC) and periportal (PP) profiles (Figure 3.5 C), which were found in their corresponding spatial locations (Figure 3.5 D). PP- and PC-specific genes isolated from Seq-Scope were also found in MARS-Seq and Drop-Seq data (see table link at `https://drive.google.com/file/d/19w55bBZwtpc7cJv7tcUtpCp3YnZT3HRi/view?usp=sharing`). The top 50 PC/PP genes from Drop-Seq and MARS-Seq were sufficient to classify PC/PP cells in the Seq-Scope dataset. Therefore, Seq-Scope single-cell analysis agreed with the former scRNA-seq results and revealed every single cell's actual spatial locations.

Cell type mapping analysis of the segmented single hepatocyte dataset revealed the spatial structure of hepatocellular zonation, identifying both PC and PP) profiles and multiple transcriptome layers ordered across the portal-central zonation axis (Figures 3.5 C and D). Many of the cluster marker genes showed a spectrum of diverse zonation patterns between the PC and PP profiles (Figure 3.6C). These gene expression patterns are consistent with the previous RNA in situ hybridization (*Aizarani et al.* (2019); *Halpern et al.* (2017)) and immunostaining results (*Park et al.* (2021)). However, previous studies using original ST (*Hildebrandt et al.* (2021)) or Slide-Seq (*Rodriques et al.* (2019)) were not able to uncover this level of detail (Figures 3.6A and B), possibly due to the limitations in resolution and RNA capture efficiency (Figure 3.3).

3.2.5 Seq-Scope Detects Non-parenchymal Cell Transcriptome from Liver Section

Although hepatocytes are the major cellular component in the liver, NPC such as macrophages (Mϕ) (blue), hepatic stellate cells (HSC) (dark green), endothelial

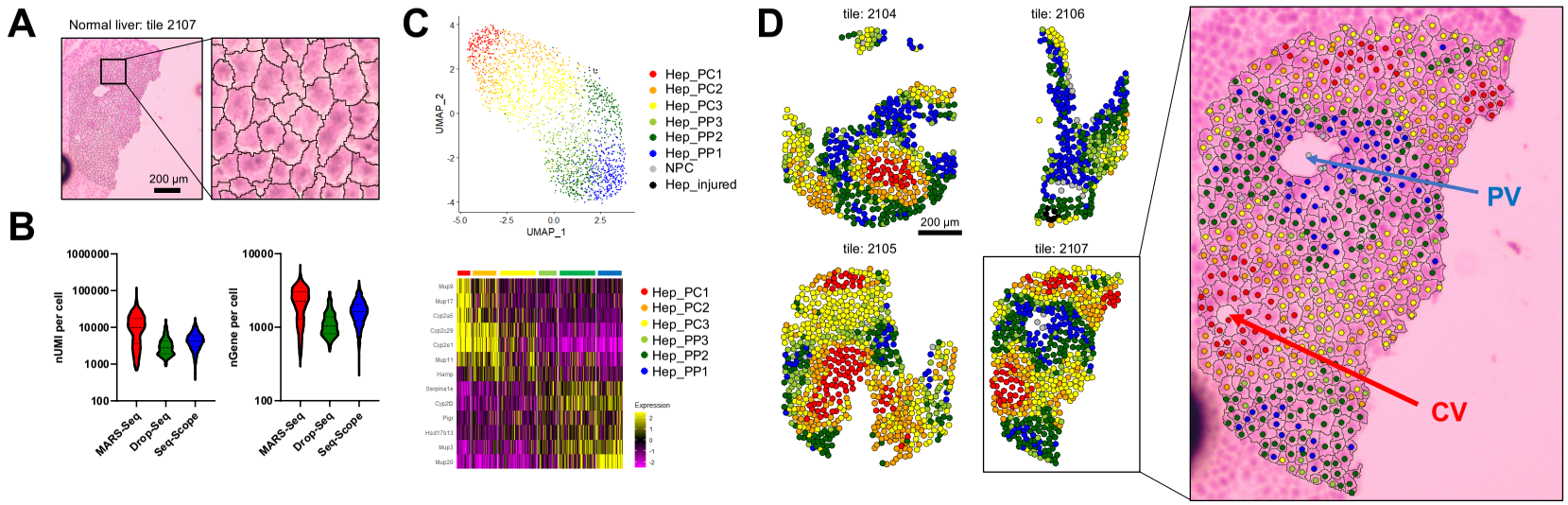

Figure 3.5: Seq-Scope performs spatial single-cell analysis in normal mouse liver. (A)Single hepatocyte segmentation based on H&E staining. (B)Comparison of Seq-Scope single-cell output with those obtained from MARS-Seq and Drop-Seq. (C)Cell-type clustering revealed multiple layers of hepatocellular zonation (Hep_PC1-3 and Hep_PP1-3), as well as a small number of non-parenchymal (NPC) and injured (Hep_injured) transcriptome phenotypes. PC, pericentral; PP, periportal. (D) Spatial map of different hepatocellular clusters (left) was overlaid with H&E staining and cell segmentation images (right). PV, portal vein; CV, central vein.

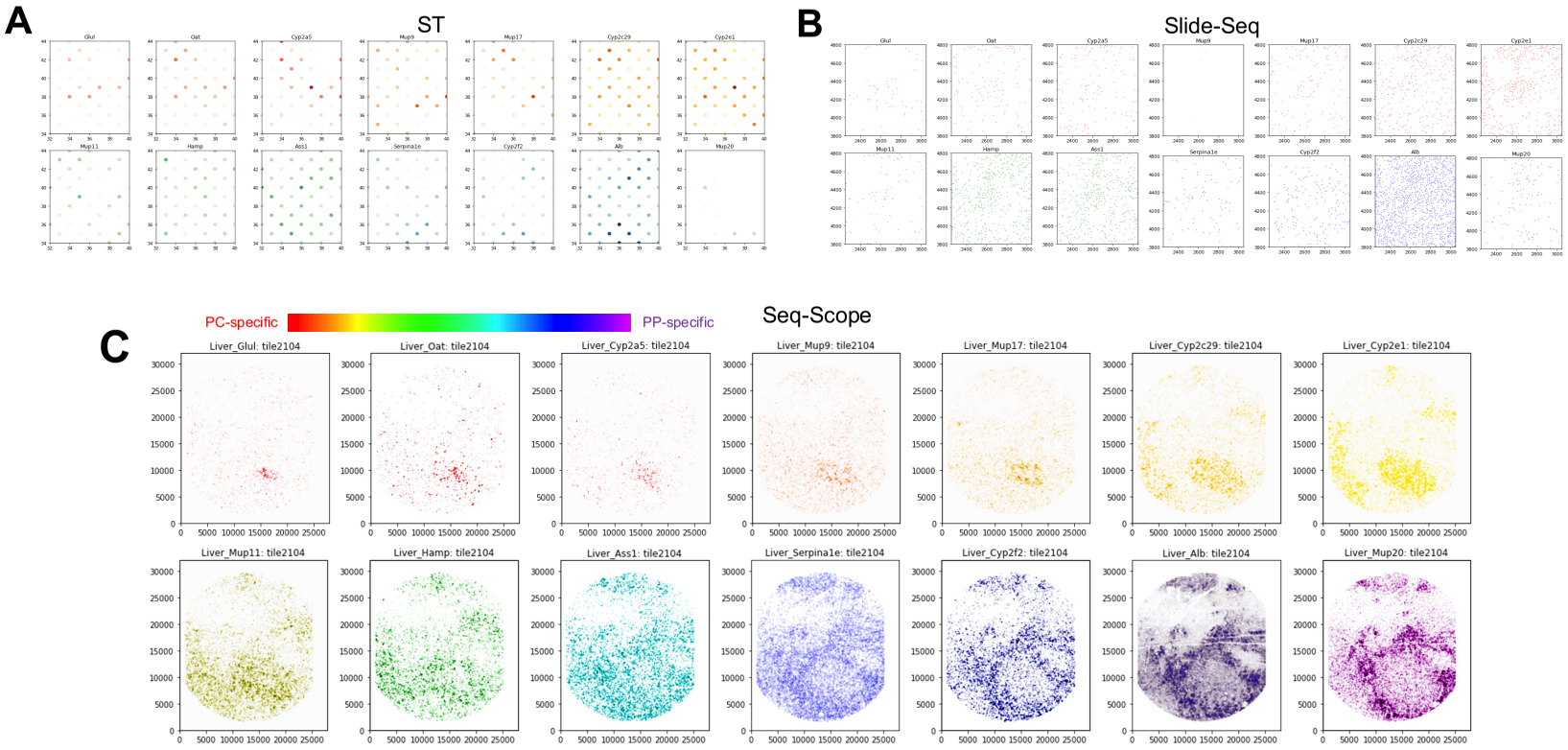

Figure 3.6: Spatial expressions of individual genes using (A)mouse liver ST, (B)Slide-Seq datasets and (C)Seq-Scope liver data. PC-specific genes are depicted in warm (red-orange-yellow) colors, whereas PP-specific genes are depicted in cold (blue-purple) colors.

cells (ENDO) (orange), and red blood cells (RBC) (red) can be found in a small portion of the histological area(*Ben-Moshe and Itzkovitz* (2019)). Due to their small sizes, these cells were not easily isolated through H&E-based image segmentation assays; H&E-based segmentation assay failed to reveal the NPC transcriptome except around the portal vein area (gray clusters in Figures 3.5 C and D), where RBCs and Mϕs often accumulate in large quantities (*Dou et al.* (2020)).

Therefore, alternatively, we segmented the Seq-Scope dataset with a uniform grid consisting of $10\mu m$-sided squares (Figures 3.7 A-D). Cell-type mapping analysis of the gridded Seq-Scope dataset identified the grids that correspond to these NPC cell types (Figure 3.7 E), based on the expression of cell-type-specific markers (Figures 3.7 E-G). Although most of the histological space was occupied by the hepatocellular area (Hep_PP and Hep_PC), the small and fragmented spaces scattered throughout the section represented the NPC area (Figure 3.8 C). The locations of the Mϕ and ENDO grids (Figure 3.8 D, first and second panels) were consistent with the spatial location of their corresponding cell-type-specific marker expression (Figure 3.8 D, arrows in the third panel) and the histologically identified Mϕ and sinusoid areas (Figure 3.8 D, arrows in the fourth panel) that are located around the segmentation boundaries (Figure 3.8 D, arrows in the fifth panel). Therefore, histology-guided cell segmentation analysis and histology-agnostic square gridding analysis complemented each other in identifying different cell types.

3.2.6 Seq-Scope Visualizes Histological Layers of Colonic Wall

The colon is another gastrointestinal organ with complex tissue layers, histological zonation structure, and diverse cellular components (*Levine and Haggitt* (1989)). Using the colon, we examined whether Seq-Scope can examine the spatial transcriptome in a non-hepatic tissue. The colonic wall is histologically divided into the colonic mucosa and the external muscle layers (*Farkas et al.* (2015)). The colonic mucosa

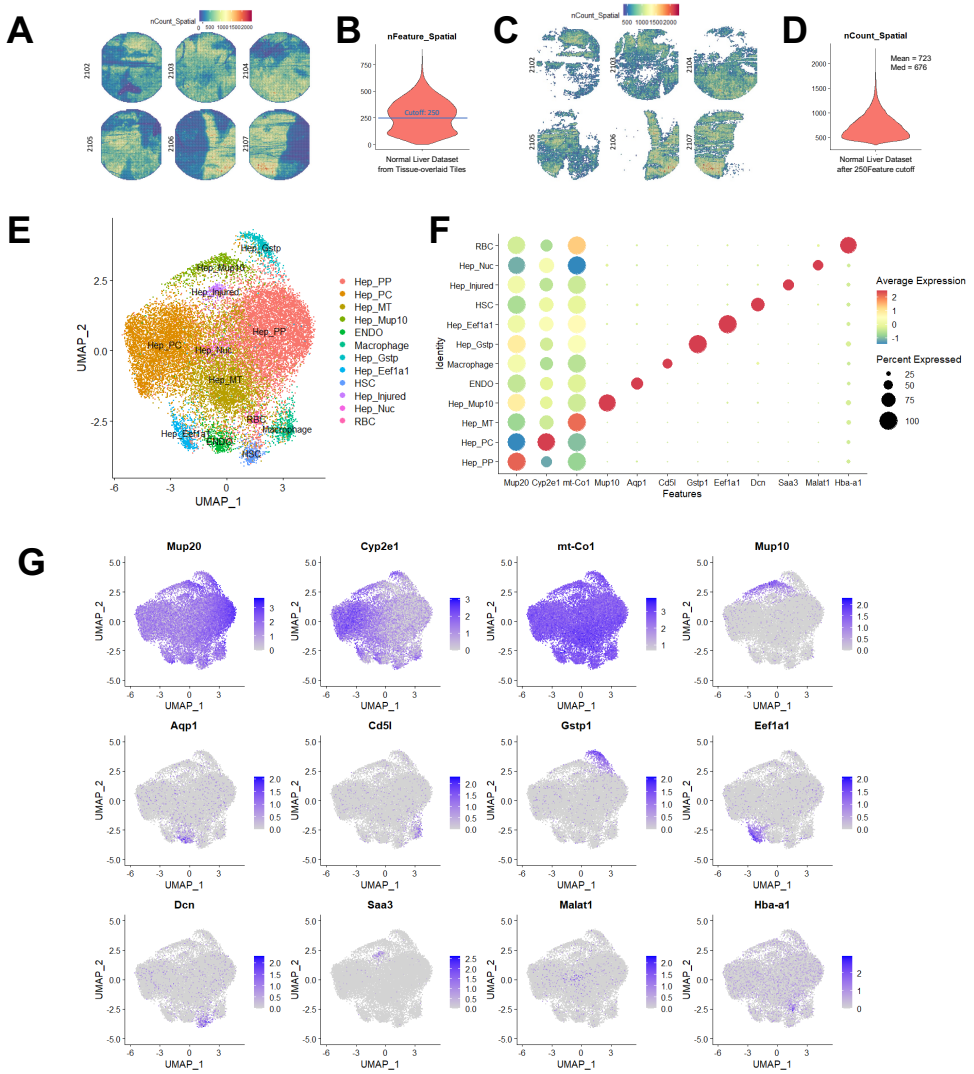

Figure 3.7: Normal liver Seq-Scope dataset analyzed by data binning with $10\mu m$-sided square grids. (A) Spatial density plot depicting the number of UMIs discovered across $10\mu m$ square grids. (B)Violin plot depicting the number of gene features (nFeature) across the $10\mu m$ square grids. Setting a 250 cutoff isolated grid units covered by the tissue area (C), each of which contains around 700 UMIs (D). A UMAP plot visualizing all clusters (E) and a dot plot (F) and UMAP plots (G) visualizing expression of cluster-specific markers are presented.

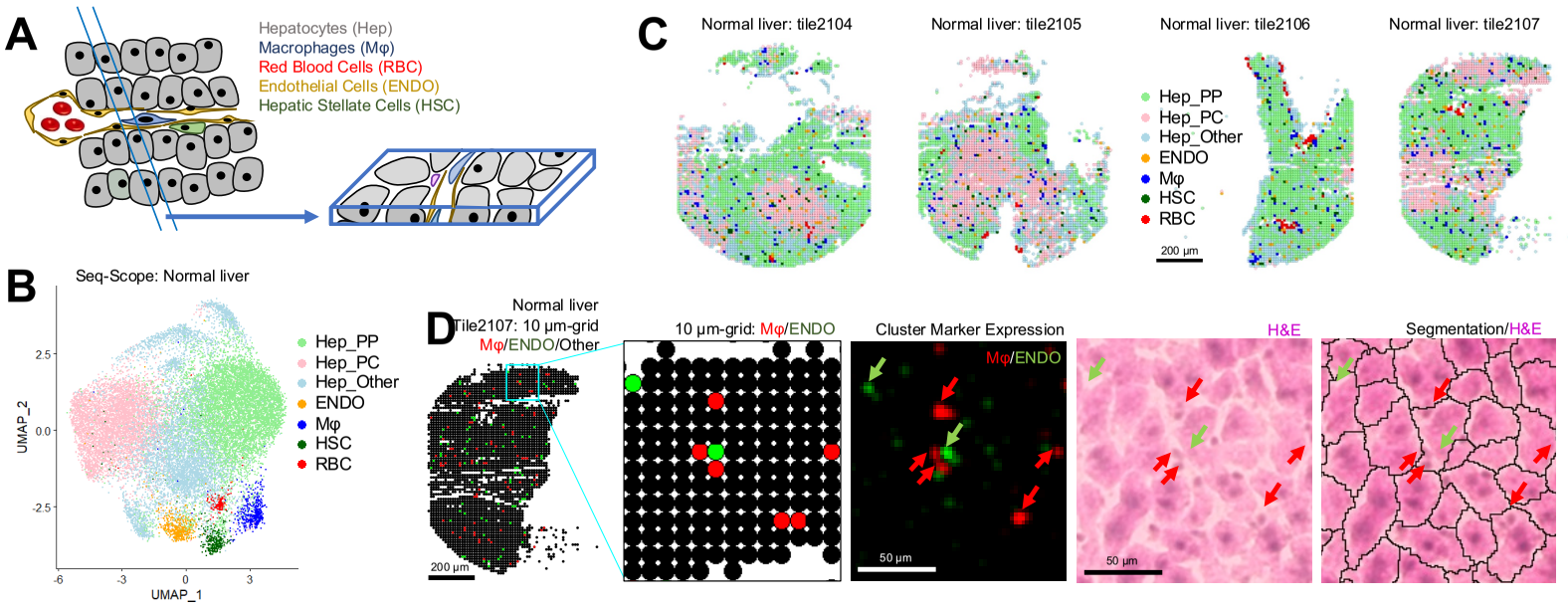

Figure 3.8: Detection of NPC transcriptome through histology-agnostic segmentation with $10\mu m$ grids. (A) Schematic diagram depicting cellular components of normal liver and their representation in a tissue section. (B and C) UMAP (B) and spatial plots (C) visualizing clusters of $10-\mu m$ grids representing indicated cell types.(D) $10-\mu m$ grid-based Mϕ and ENDO mapping data (first and second panel) are compared with spatial plot data of cluster-specific markers (third panel), H&E (fourth), and segmented H&E (fifth) data.

consists of the epithelium and lamina propria, and the epithelium is further divided into the crypt-base, transitional, and surface layers (Figure 3.9 A). Clustering analysis of the gridded Seq-Scope dataset revealed transcriptome phenotypes corresponding to these layers (Figure 3.9 B) and visualized their spatial locations (Figures 3.9 C).

3.2.7 Seq-Scope Identifies Individual Cellular Components from Colon Tissue

In addition to visualizing the layer structure, Seq-Scope also revealed the various colonic epithelial and non-epithelial cell types (Figures 3.9 D–I). In the crypt base, stem/dividing, DCSC and Paneth-like cell phenotypes (Figures 3.9 E and F) were identified. The stem/dividing cells expressed higher levels of ribosomal proteins while expressing lower levels of other epithelial cell-type markers (Figure 3.9 J;). DCSCs expressed secretory cell markers, such as *Agr2*, *Spink4*, and *Oit1* (Figure 3.9 J), whereas Paneth-like cells expressed *Mptx1*, a recently identified marker of the Paneth cell in the small intestine (*Haber et al.* (2017)).

Seq-Scope also identified distinct cell types at the surface of the colonic mucosa (Figures 3.9 D–F). The top layer of the epithelial cells expressed surface colonocyte markers, such as *Aqp8* (*Fischer et al.* (2001)), *Car4* (*Borenshtein et al.* (2009)), and *Saa1* (*Eckhardt et al.* (2010)) (Figure 3.9 J). Some of the epithelial cells expressed goblet cell-specific markers, such as *Zg16*, *Fcgbp*, and *Tff3* (*Haber et al.* (2017); *Pelaseyed et al.* (2014)) (Figure 3.9). In addition, Seq-Scope also identified EEC expressing hormones, such as glucagon, peptide YY, insulin-like peptide, and Cholecystokinin (CCK) (Figure 3.9 J).

Below the epithelium, there are connective tissue layers, including the lamina propria, submucosa, and external muscle layers. Seq-Scope identified many non-epithelial cell types from these layers, including smooth muscle, fibroblasts, enteric neurons, Mϕs, and B cells (Figures 3.9 G–I). These results indicate that Seq-Scope

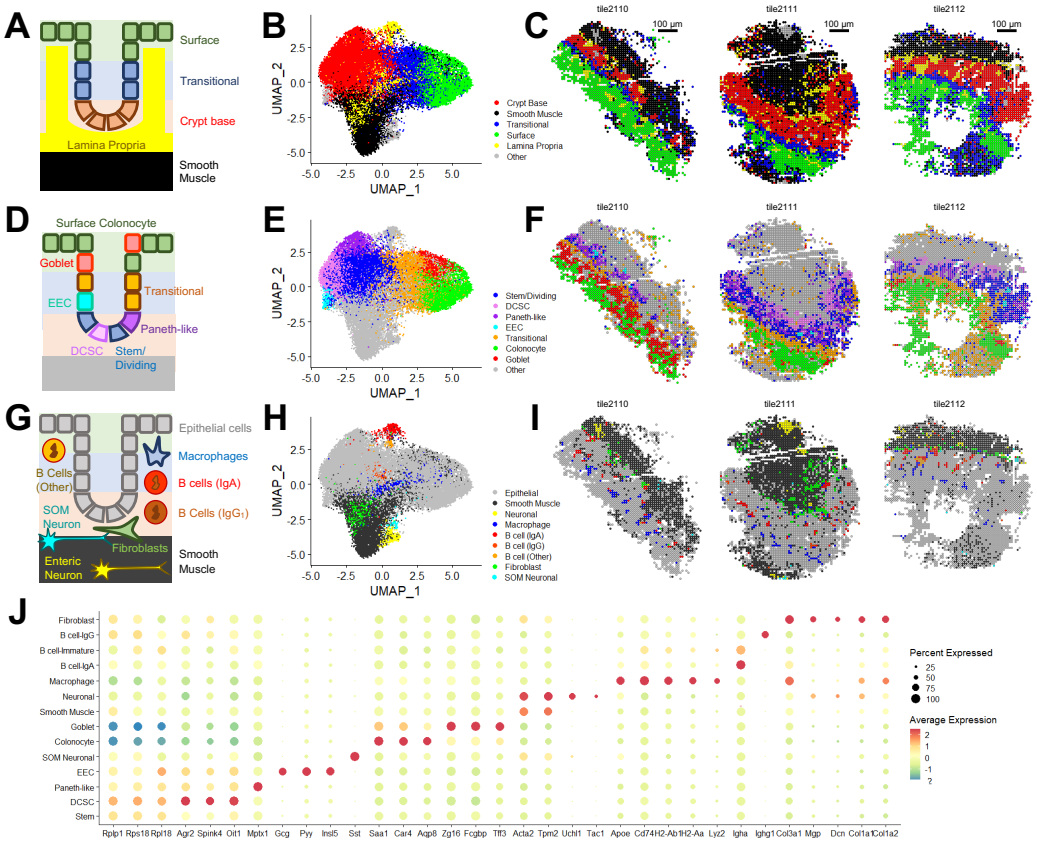

Figure 3.9: Seq-Scope identifies various cell types from colonic wall histology. (A–I) Seq-Scope reveals major histological layers (A–C), epithelial cell diversity (D–F), and non-epithelial cell diversity (G–I) through transcriptome clustering. (A, D, and G) Schematic representation of colonic wall structure. Clusters corresponding to the indicated cell types were visualized in UMAP manifold (B, E, and H) and histological space (C, F, and I).(J) Cluster-specific markers were examined in dot plot analysis. deep crypt secretory cell (DCSC), deep crypt secretory cells; enteroendocrine cells (EEC), enteroendocrine cells; SOM Neuronal, somatostatin-expressing neuronal cells.

can transcriptomically recognize most of the major cell types present in the normal colonic wall.

3.2.8 Seq-Scope Performs Microscopic Analysis of Colonic Spatial Transcriptome

To take advantage of Seq-Scope's high-resolution data, we employed Multi-scale Sliding Windows analysis(MSSW) (Figures 3.10 A–C) and spatial plotting of cluster markers (Figures 3.10 D-F), focusing on the same region of the colonic wall. MSSW analysis drew a clear line between different cellular compartments (Figures 3.10 A–C); the original gridding analysis ($10\mu m$) or analysis with smaller grids ($5\mu m$) did not reveal this level of high-resolution detail.The sliding windows cluster assignments (Figures 3.10 A-C) were congruent with the spatial plotting of the relevant cluster marker genes (Figures 3.10 D–F) and H&E histology data (Figure 3.10 G). For instance, in all of these data, B cells and Mϕs were confined to the lamina propria, whereas crypt base cell markers were confined to the epithelium (separated by dotted lines in Figures 3.10 D–G). The B cells and Mϕs are often in very close proximity (Figures 3.10 C and F), likely due to their functional interactions (*Spencer and Sollid* (2016)). Genes specifically expressed in S and G2/M cell-cycle phases (*Nestorowa et al.* (2016)) were highly expressed in the crypt base area where stem/dividing cells are located (*Levine and Haggitt* (1989)), however, their expression was lower in the surface area (Figure 3.10 H).

3.3 Materials and Methods

3.3.1 Seq-Scope Technology

The Seq-Scope experiments are divided into two rounds of sequencing steps: 1st-Seq and 2nd-Seq (Figure 3.11) . 1st-Seq generates a physical array of spatially bar-

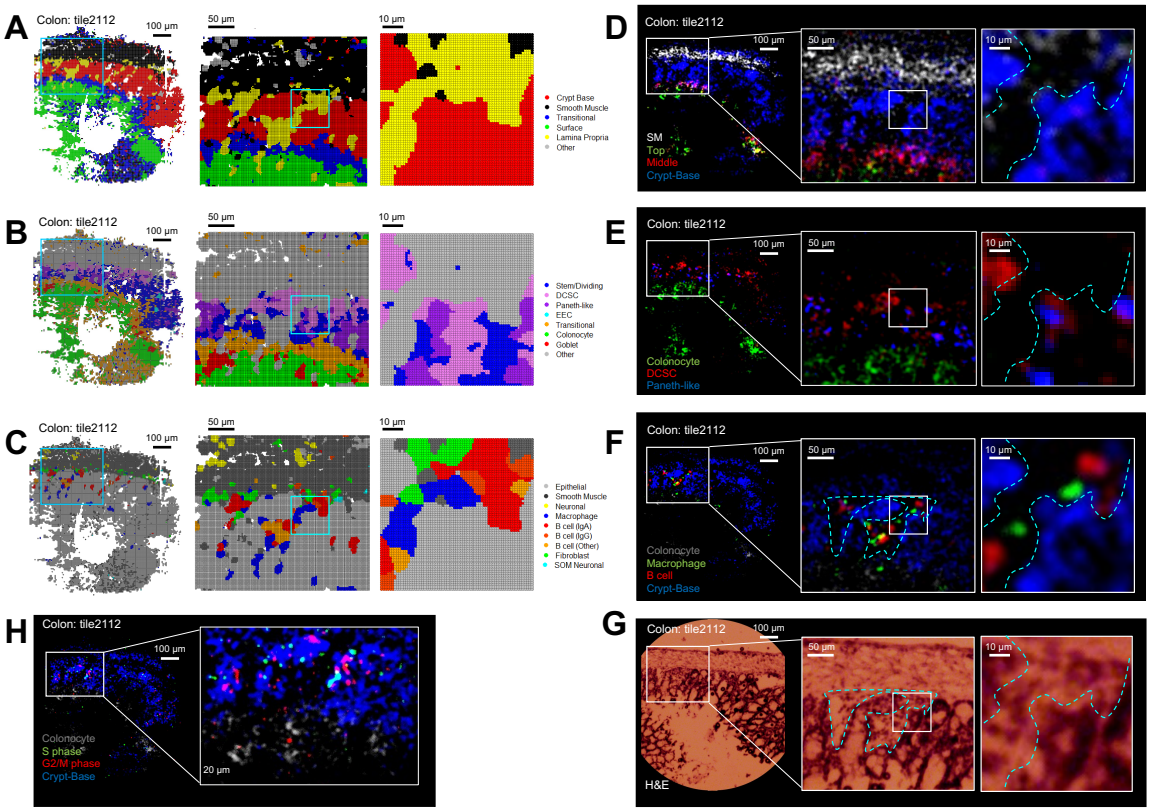

Figure 3.10: Seq-Scope enables microscopic analysis of colon spatial transcriptome. (A–C) Spatial cell-type mapping shown in Figure 3.9 was refined using Multi-scale Sliding Windows analysis with $5\mu m$ (left), $2\mu m$ (center), or $1\mu m$ (right) intervals. (D–H) Original Seq-Scope dataset was analyzed by spatial gene expression plotting, using indicated layer-specific (D), cell-type-specific (E and F), or cell-cycle-specific (H) marker genes. These spatial transcriptome features were consistent with underlying H&E histology (G).

coded RNA-capture molecules and a spatial map of barcodes where each barcoded sequence is associated with a spatial coordinate in the physical array. 2nd-Seq captures mRNAs released from the tissue placed on the physical array from the 1st-Seq and sequences the captured molecules containing both cDNA and spatial barcode information.

1st-Seq of Seq-Scope starts with the solid-phase amplification of a single-stranded synthetic oligonucleotide library using an Illumina sequencing platform. The oligonucleotide "seed" molecule contains the PCR/read adaptor sequences, the restriction enzyme-cleavable RNA-capture domain (oligo-dT), and the high-definition map coordinate identifier (HDMI), a spatial barcode composed of a 20–32 random nucleotide sequence. The library is amplified on a lawn surface coated with PCR adapters, generating a number of clusters, each of which is derived from a single "seed" molecule. Each cluster has thousands of oligonucleotides that are identical clones of the initial oligonucleotide "seed" (*Bentley et al.* (2008)). The HDMI sequence and spatial coordinate of each cluster are determined through a sequencing-by-synthesis (SBS) procedure using the realtime analysis (RTA) software, without requiring any in-house custom image analysis. After SBS, the oligonucleotides in each cluster are processed to expose the nucleotide-capture domain, producing an HDMI-encoded RNA-capturing array (HDMI-array), the physical array produced by 1st-Seq of Seq-Scope.

2nd-Seq of Seq-Scope begins with overlaying the tissue slice onto the HDMI-array (Figure 3.11 E). The mRNAs from the tissue are used as a template to generate cDNA footprints on the HDMI-barcoded RNA capture molecule. Then, the secondary strand is synthesized on the cDNA footprint. Because each cDNA footprint is paired with a single random primer after washing, the random priming sequence is used as a UMI. The secondary strand, which is a chimeric molecule of HDMI and cDNA sequences, is then collected and prepared as a library through PCR. The paired-end sequencing of this library reveals the cDNA footprint sequence, as well as its corresponding HDMI

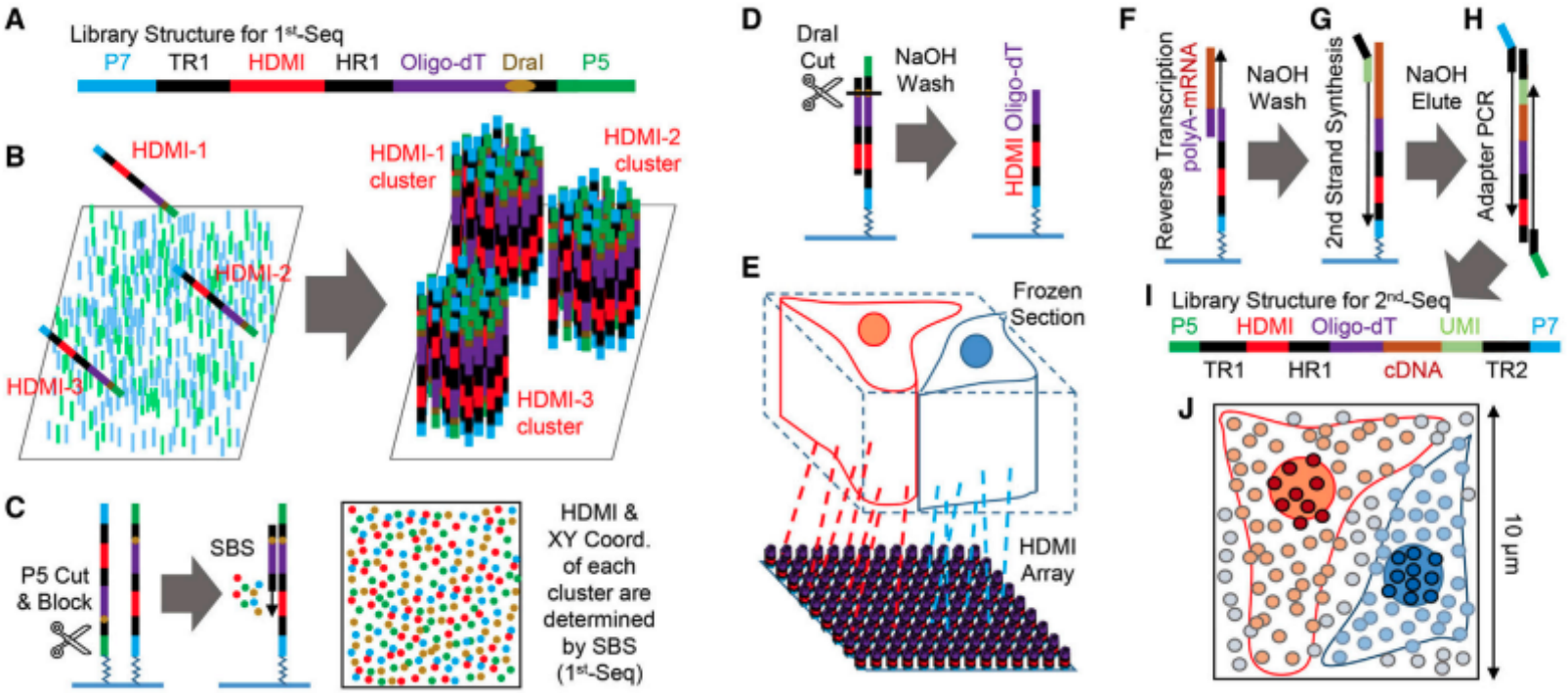

Figure 3.11: Seq-Scope Overview. (A and B) Chemistry workflow for generating HDMI-array in 1st-Seq (A), and using the HDMI-array for constructing library for 2nd-Seq (B). The 2nd-Seq library is subjected to the standard next-generation sequencing workflow in Illumina and BGI platforms.(C-E) Bioinformatics workflow for estimating tissue boundaries (C), visualizing and analyzing spatial gene expression patterns (D), and determining nuclear and cytoplasmic areas (E).(F) Chemistry workflow for generating UMI-encoded HDMI-array in 1st-Seq.(G) Evaluation of UMI-encoding methods based on either random priming (UMI_Randomer) or array encoding (UMI_Array). The number of HDMI with multiple read counts was efficiently reduced by either UMI_Randomer- or UMI_Array-based collapsing methods.

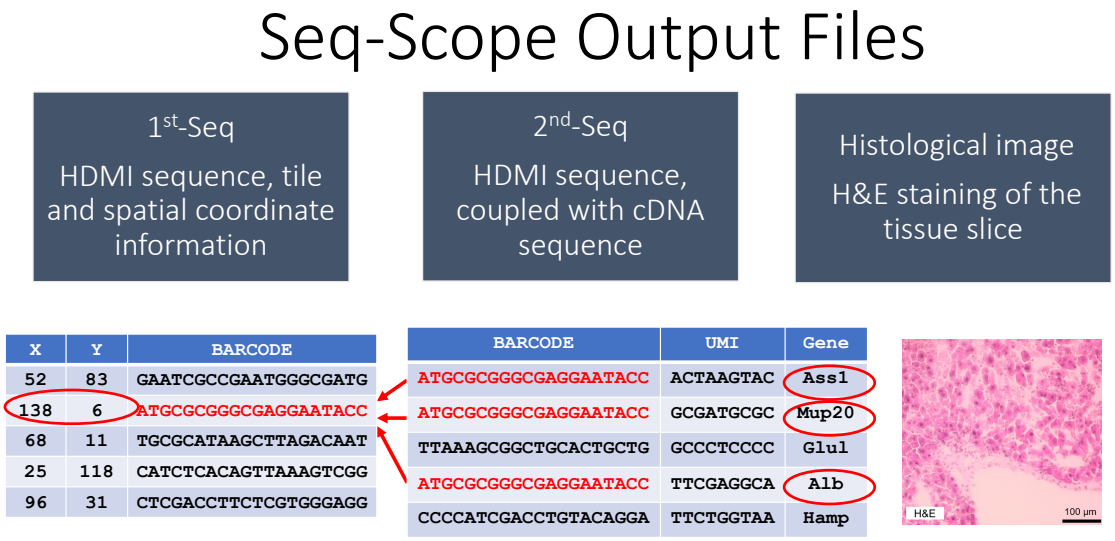

Figure 3.12: Seq-Scope data structure. There are three outputs from Seq-Scope: 1st-Seq, 2nd-Seq and H&E image. 1st-Seq provides a data table that has a spot's XY coordinate and its HDMI sequence. 2nd-Seq generates another data table where each HDMI is associated with a cDNA sequence. Using HDMI sequence as an index, each gene's spatial coordinates can be quickly identified.

sequence.

For each HDMI sequence, 1st-Seq provides spatial coordinate information whereas 2nd-Seq provides captured cDNA information. Correspondingly, the spatial gene expression matrix is constructed by combining the 1st-Seq and 2nd-Seq data, which is used for various analyses.

There are three experimental outputs from Seq-Scope(Figure 3.12), which will serve as input data for downstream computational analysis. (1) HDMI sequence, tile and spatial coordinate information from 1st-Seq, (2) HDMI sequence, coupled with cDNA sequence from 2nd-Seq, and (3) Histological image obtained from H&E staining of the tissue slice.

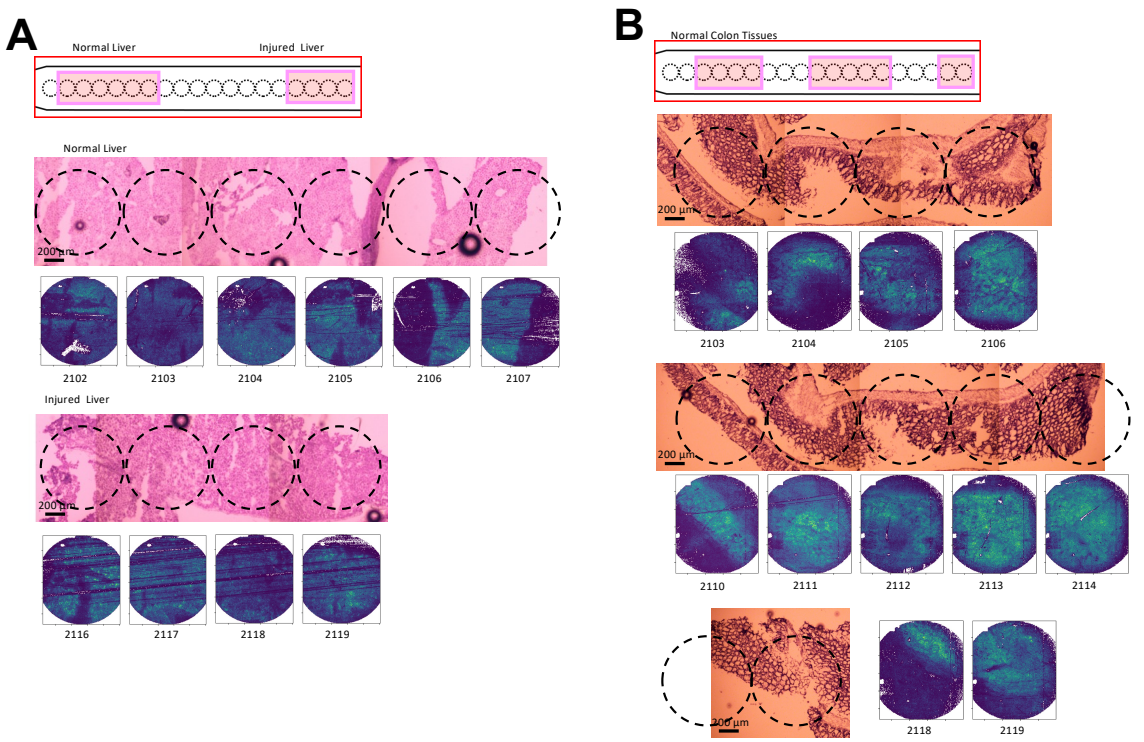

Figure 3.13: HDMI discovery plot for Seq-Scope liver and colon Data. The schematic diagram visualizes the tiles which were attached to the indicated liver (A, top) or colon (B, top) tissues. On the bottom, H&E staining images and their corresponding HDMI discovery plots were presented.

3.3.2 Tissue Boundary Estimation

To estimate the tissue boundary, the 2nd-Seq data were joined into 1st-Seq data according to their HDMI sequence. As a result, for each of the 2nd-Seq data whose HDMI was found from 1st-Seq, the tile number and XY coordinates were assigned. Finally, an HDMI discovery plot was generated to visualize the density of HiSeq HDMI in a given XY space of each tile. The density plots were manually assigned to the corresponding H&E images for quality control examination (Figures 3.13).

3.3.3 Read Alignment and Generation of Digital Gene Expression Matrix

Read alignment was performed using STAR/STARsolo 2.7.5c (*Dobin et al.* (2013)), from which the digital gene expression (DGE) matrix was generated. From MiSeq

data, HDMI sequences of clusters located on the bottom tile were extracted and used as a "whitelist" for the cell (HDMI) barcode after reverse complement conversion. The first 20 (HDMI-DraI version) or 30 (HDMI32-DraI) basepairs of HiSeq data Read 1 were considered as the cell (HDMI) barcode. HDMI assignments were performed using the default error correction method implemented in STARsolo (1MM_multi).

Due to the extensive washing steps after secondary strand synthesis, it was expected that each single molecule of HDMI-cDNA hybrid would lead to one secondary strand in the library. Therefore, the first 9-mer of Read 2 sequence, which is derived from the Randomer sequence, could serve as a proxy of the unique molecular identifier (UMI). Accordingly, the first 9 basepairs of HiSeq Read 2 data were copied to Read 1 and used as the unique molecular identifier (UMI). Read 2 was trimmed at the 3' end to remove polyA tails of length 10 or greater and was then aligned to the mouse genome (mm10) using the GeneFull option with no length threshold and no cell filtering. For the genes whose expression couldn't be monitored by the GeneFull option, the Gene option was used to generate the gene expression discovery plots. UMIs were deduplicated using the default error correction method implemented in STARsolo (1MM_All), in which all UMIs with 1 mismatch distance to each other are collapsed (i.e., counted once).

For saturation analysis, multiple read alignments were performed using 25%, 50% and 75% random subsets of the 2nd-Seq results. The alignment output values were plotted in a graph (Figure 3.14) to generate a saturation curve. Hyperbolic regression was used to estimate the total unique transcript number in the liver (60,292,407 to 96,899,822; 95% confidence interval) and colon (308,586,493 to 510,224,639; 95% confidence interval) Seq-Scope libraries.

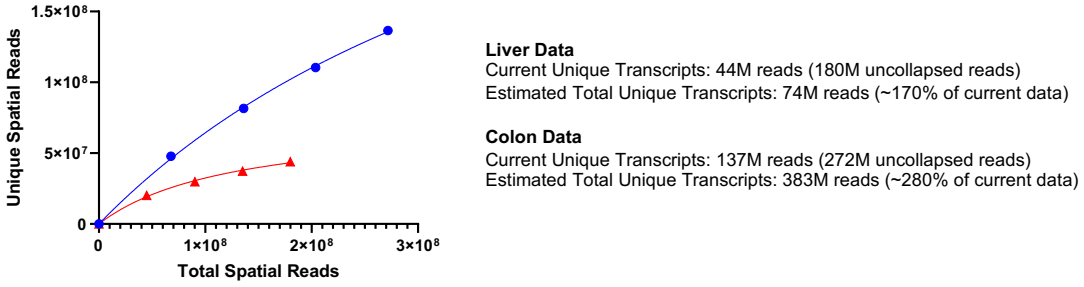

Figure 3.14: Saturation analysis of liver (red) and colon (blue) Seq-Scope dataset. Hyperbolic regression was used to estimate the total unique transcript number in the liver and colon Seq-Scope libraries.

3.3.4 Error Correction Methods for Spatial Barcodes

Although the possibility of per-base error is very low, Seq-Scope involves a multi-step processing of sequences and DNA samples, so we expect that a small but non-negligible fraction of HDMI barcodes will contain errors. In current study, error correction and demultiplexing of HDMI barcodes were performed in STARsolo using the 2nd-Seq result as a FASTQ input, and the 1st-Seq result as a barcode whitelist. We used the STARsolo's default option (1MM_multi), which implements a robust statistical error correction method similar to 10X CellRanger 2.2.0. In this method, HDMIs are allowed to have one mismatch, and the posterior probability calculation is used to choose the barcode when multiple mismatched sequences are present.

3.3.5 Analysis of Spliced and Unspliced Gene Expression

To obtain separate read counts for spliced and unspliced transcripts, we used the Velocyto (*La Manno et al.* (2018)) option in the STARsolo software. Unspliced and spliced mRNA read counts were plotted onto the histological coordinate plane to identify the nuclear-cytoplasmic structure. To test the reproducibility of the image

analysis, all genes were randomly divided into three groups, and spliced and unspliced read counts were obtained independently. Nuclear-specific (*Malat1*, *Neat1* and *Mlx-ipl*) and mitochondria-encoded (all genes whose name start with '*mt-*') transcripts were also plotted and analyzed.

3.3.6 H&E Based Image Segmentation for Spatial Single Cell Analysis

To perform cell segmentation using H&E histology images, the watershed algorithm was implemented on H&E histology images(*Kornilov and Safonov* (2018)). The cell segmentation results isolated the single hepatocyte areas, which are consistent with the visual inspection of the H&E images (Figure 3.5 A). Cell boundary images and cell center coordinates were exported to aggregate Seq-Scope data so that the transcriptome information from all HDMI pixels within each segmented area were collapsed into their corresponding cell center coordinate barcodes, generating a single cell-indexed DGE matrix. The DGE matrix was then used for clustering analysis.

3.3.7 Simple Aggregation

Simple aggregation generated square bins by dividing the imaging space into $100\mu m^2$ ($10\mu m$-sided) square grids and collapsing all HDMI-UMI information into one barcode per grid. Alternatively, data binning was also performed with $25\mu m^2$ ($5\mu m$-sided) square grids. After data binning, gene types were filtered to only contain protein-coding genes, lncRNA genes, and immunoglobulin/T cell receptor genes. When multiple genes share the same gene symbol, we retained only first-appearing gene. We also exclude any hypothetical gene models (genes designated as *Gm*-number).

3.3.8 Clustering Analysis

The binned and processed DGE matrix was analyzed in the Seurat v4 package (*Butler et al.* (2018)). Feature number threshold was applied to remove the grids that corresponded to the area that was not overlaid by the tissue or was extensively damaged through scratches. Data were normalized using regularized negative binomial regression implemented in Seurat's SCTransform function. Clustering was performed using the shared nearest neighbor modularity optimization implemented in Seurat's FindClusters function. Clusters with mixed cell types were subjected to an additional round of clustering to get separation between the different cell types, while similar cell types were grouped together. UMAP (*Becht et al.* (2019)) manifold, also built in the Seurat package, was used to assess the clustering performance. Top markers from each cluster, identified through the FindAllMarkers function, were used to infer and annotate cell types. Then the clusters were visualized in the UMAP manifold or the histological space using DimPlot and SpatialDimPlot functions, respectively. Raw and normalized transcript abundance in each tile, cluster and spatial grid was visualized through the VlnPlot, DotPlot, FeaturePlot and SpatialFeaturePlot functions built in the Seurat package.

3.3.9 Analysis of Transcripts Discovered Outside of Tissue-overlaid Region

Some RNAs were discovered in an area where the tissue was not overlaid. It is possible that a trace of tissue fluid or debris, as well as ambient RNAs released from the tissues, may have generated this pattern. Although the RNA discovery in these regions was scarce, the compositions of RNA discovered in tissue-overlaid (nFeature > 250 in liver dataset) and non-overlaid regions (nFeature ≤ 250 in liver dataset) were very similar to each other ($r = 0.98$ in Spearman coefficients). The minor differences between these two regions could be obviously explained by the different

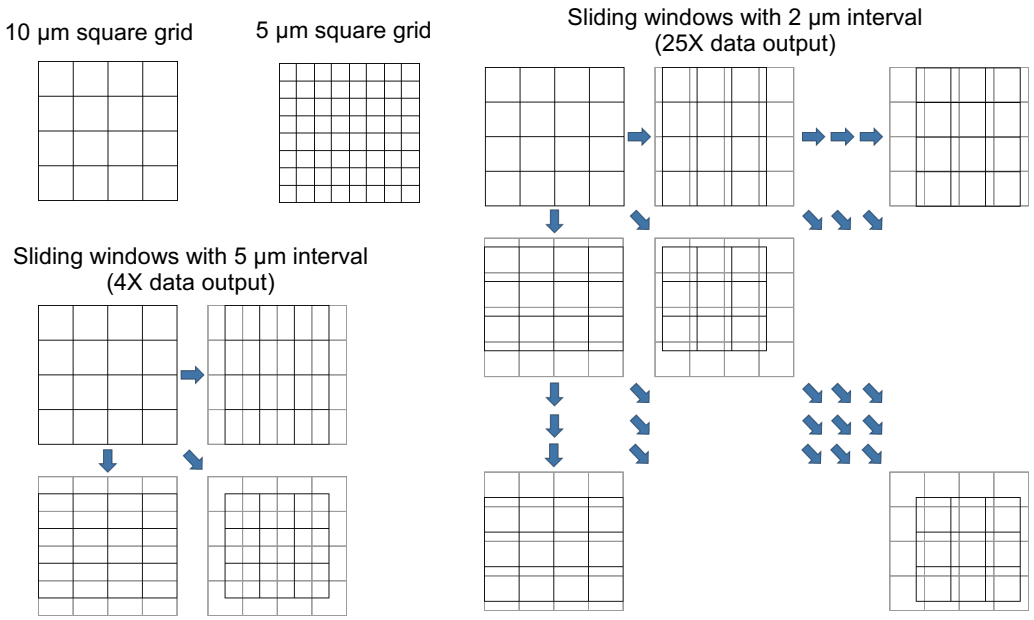

Figure 3.15: Schematic diagrams depicting the sliding windows analysis methodology. Compared to the $10\mu m$ grid dataset, $5\mu m$ grid dataset produces higher resolution; however, the transcriptome information revealed by $5\mu m$ grid area is only 25% of what was recovered from $10\mu m$ grid area. Correspondingly, $5\mu m$ dataset produced substantial noises in cell type assignment. To overcome this, sliding windows analysis was performed to maintain transcriptome information per pixel while achieving higher resolution of cell type mapping by oversampling the data 4 times ($5\mu m$-interval), 25 times ($2\mu m$-interval) or 100 times ($1\mu m$-interval; scheme not shown).

rates of ambient RNA release/capture and the different composition of cell types in the tissue debris. Therefore, it is plausible that ambient and debris-derived RNAs generated the pattern of RNA discovery in the tissue non-overlaid region.

3.3.10 Multi-scale Sliding Window Analysis

Multi-scale Sliding Window (MSSW)analysis (see Chapter IV for details) was employed to fine tune the annotation using FindTransferAnchors and TransferData functions implemented in Seurat. The anchors provided by the $10\mu m$ grid dataset were used to guide other datasets produced from the same Seq-Scope result. Com-

pared to the $10\mu m$ grid dataset, the $5\mu m$ grid dataset was much noisier in spatial (Figure 3.16 A) analyses even after multi-scale fine tuning. To circumvent this problem, we employed the sliding windows analysis; after the initial $10\mu m$ grid sampling, the grid was shifted both horizontally and vertically with $5\mu m$, $2\mu m$ or $1\mu m$ intervals, producing 4, 25 and 100 times more data, respectively (see Figure 3.15 for a schematic illustration). Then, the original $10\mu m$ grid dataset was used to guide these sliding windows datasets to perform high-resolution cell type annotation. Sliding windows analysis with $5\mu m$ intervals (Figure 3.16 C, right) performed much better when compared to the $5\mu m$ grid datasets (Figure 3.16 C, center), and showed the UMAP pattern (Figure 3.16 B) whose shape is more similar to the original $10\mu m$ grid dataset. Sliding windows analyses with $5\mu m$ intervals were used to produce left panels in Figures 3.10 A–C. Sliding windows analyses with $2\mu m$ intervals were used to produce middle panels in Figures 3.10 A–C. Sliding windows analyses with $1\mu m$ intervals were used to produce the right panels in Figures 3.10 A–C.

3.3.11 Visualization of Spatial Gene Expression

Spatial gene expression was visualized using a custom python code. Raw digital expression data of the queried gene (or gene list) were plotted onto the coordinate plane according to their HDMI spatial index. Considering the lateral RNA diffusion distance of $1.7 \pm 2\mu m$(mean $\pm$ SD) measured from the original ST study (*Ståhl et al.* (2016)), gene expression densities were plotted as a $\sim 3\mu m$-radius circle at a transparency alpha level between 0.005 and 0.5. In spatial gene expression images with a white background, the intensity of the colored spot indicates the abundance of transcripts around the spot location. Spatial gene expression images with a black background were created for genes or gene lists of high expression values, to make it easy to adjust the linear range of gene expression density and to overlay gene expression densities of different queries with different pseudo-color encoding. The inverse image of the greyscale

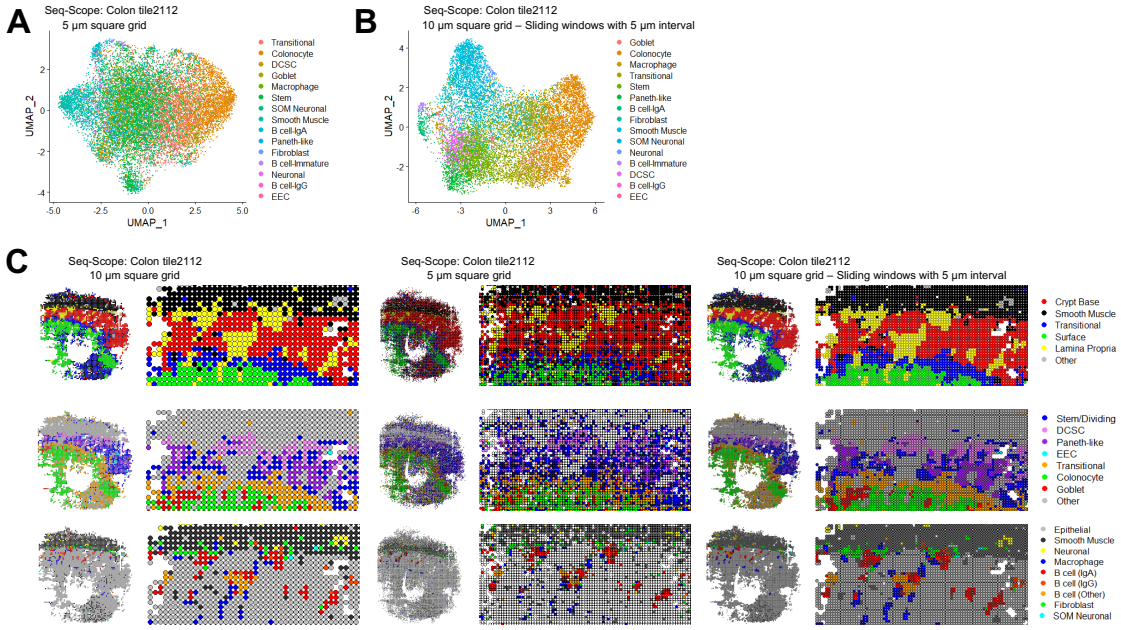

Figure 3.16: Cell type mapping by Multi-scale Sliding Window analysis. (A and B) UMAP plots constructed from $5\mu m$ grid dataset (A) and sliding windows dataset of $10\mu m$ grids with $5\mu m$ intervals (B). (C) Multi-scale cell type mapping combined with sliding window analysis identifies clear boundaries between different cell types with high resolution. Colon Seq-Scope dataset was analyzed using simple gridding with $10\mu m$-sided squares (left). Using the $10\mu m$ dataset as an anchor, multi-scale cell type mapping was performed in $5\mu m$ gridding dataset (center). Even though $5\mu m$ gridding improved the resolution, the image was very noisy due to scarce genetic information in each grid. To overcome this, we performed the same analysis using a dataset produced by sliding windows analysis of $10\mu m$ gridding dataset with $5\mu m$ intervals. The output images (right) clearly visualize the boundaries between different cell types with high resolution. Cell type annotations depict major histological layers (upper), epithelial cell diversity (middle), and non-epithelial cell diversity (lower).

plot was pseudo-colored with red, blue, green or gray, and the image contrast was linearly adjusted to highlight the biologically relevant spatial features. Finally, different pseudo-colored images were overlaid together to compare the gene expression patterns in the same histological coordinate plane. Cell cycle-specific genes, such as S phase- and G2/M phase-specific gene lists (*Nestorowa et al.* (2016)), were retrieved from the Seurat package, and their mouse homologs were identified using the biomaRt package (*Durinck et al.* (2009)). The list of cell type markers used in spatial plots can be found at `https://docs.google.com/spreadsheets/d/1jb1QpDisTEGAy6EXJtSNXMQfv0jsUrkA/edit?usp=sharing&ouid=104769725873530382604&rtpof=true&sd=true`

3.3.12 Benchmark Analysis

The performance of Seq-Scope in liver and colon experiments were benchmarked against publicly available datasets produced by 10X VISIUM (`https://support.10xgenomics.com/spatial-geneexpression/datasets/1.1.0/V1_Human_Brain_Section_1`). DBiT-Seq (GEO: GSM4096261 in GSE137986) (*Liu et al.* (2020)), SlideSeq (Single Cell Portal: 180819_11 in SCP354) (*Rodriques et al.* (2019)), Slide-SeqV2 (Single Cell Portal: 190921_19 in SCP815) (*Stickels et al.* (2021)), and HDST (GEO: GSM4067523 in GSE130682) (*Vickovic et al.* (2019)). Liver Seq-Scope dataset was separately benchmarked against former liver datasets produced using original ST (Zenodo: 10.5281/zenodo.4399655) (*Hildebrandt et al.* (2021)) and Slide-Seq (Single Cell Portal: 1808038_8 in SCP354) (*Rodriques et al.* (2019)).

The center-to-center resolution was calculated per each pixel as the distance from the closest tissue-overlaid pixel. For the technologies that have a defined pixel area (VISIUM, DBiT-Seq and HDST), pixel density was calculated as the inverse of the pixel area. For Slide-Seq, Slide-SeqV2 and Seq-Scope, pixel density was calculated in $150\mu m$ grids (Slide-Seq and Slide-SeqV2) and $10\mu m$ grids (Seq-Scope) of the final dataset. Grids that contained less than 10 pixels were excluded from the analysis.

nUMI corresponds to the number of unique transcripts mapped to the transcriptome, and nGene corresponds to the number of gene features discovered per each pixel. nUMI/pixel and nGene/pixel values were multiplied by the average pixel density (pixel/$\mu m2$) to obtain the area-normalized nUMI and nGene (nUMI/$\mu m2$ and nGene/$\mu m2$, respectively) for each pixel.

3.4 Summary

In this chapter we described Seq-Scope, a novel ST technology achieving submicrometer resolution ($\sim 0.6\mu m$) and efficient transcriptome capture rate. Seq-Scope repurposes the Illumina sequencing platform for ST. Seq-Scope reveals the variation of spatial transcriptome at various resolutions, including tissue zonation according to the portal-central (liver) and crypt-surface (colon) , cellular components including single-cell types and subtypes, and subcellular architectures of nucleus and cytoplasm. Seq-Scope is quick, straightforward, precise, and easy-to-implement and makes spatial single-cell analysis accessible to a wide group of biomedical researchers.

CHAPTER IV

STtools: Comprehensive Software Pipeline for Ultra-high-resolution Spatial Transcriptomics Data

4.1 Background

Recent developments in single-cell and spatial RNA-sequencing (RNA-seq) technologies enabled fine-scale exploration of cell-type-specific expressions and tissue compositions. Technologies such as VISIUM (*Ståhl et al.* (2016)), Slide-Seq (*Rodriques et al.* (2019); *Stickels et al.* (2021)) and Seq-Scope (*Cho et al.* (2021)) associates specific barcode sequences with spatial coordinates and attaches these spatial barcodes to individual cDNA fragments to resolve transcriptomic profiles with spatial resolution.

Current software tools analyzing spatially resolved transcriptomes (*10X Genomics*, 2022; *Palla et al.* (2022); *Stickels et al.* (2021)) are primarily designed for relatively coarse resolution technologies such as VISIUM ($100\mu m$) or Slide-Seq ($10\mu m$), where each spatial barcode typically represents more than a single cell. However, when analyzing transcriptome spatially resolved at a micrometer or a submicrometer resolution, most current tools perform poorly due to various computational challenges. First, the number of spatial barcodes per mm^2 rapidly increases as resolution increases ($\sim$ 120 for VISIUM, $\sim 3K$ for Slide-Seq and $> 1M$ for Seq-Scope), and few tools seamlessly scale to handle millions of spatial barcodes. Second, even though higher-

resolution technologies may contain UMI counts per given area, the UMI count per spatial barcode is typically much lower due to the limited number of mRNAs that can be captured. As a result, existing tools may perform poorly if they assume that individual spatial barcodes contain sufficient UMIs to be clustered into a cell type. Third, submicrometer-resolution technologies inform us of subcellular transcriptomic architecture within individual cells (*Cho et al.* (2021)), but existing tools do not account for subcellular components in their analysis and visualization to accommodate the ultra-high resolution from recent technologies.

To address these challenges, we developed STtools, which is capable of handling various ST platforms, including submicrometer-resolution ST technology such as Seq-Scope. STtools provides a comprehensive framework for analyzing ST datasets, enabling both super-cellular, cellular and sub-cellular resolution analysis and visualization.

4.2 Results

4.2.1 STtools Enables High Resolution Cell Type Mapping

We illustrated example results from Seq-Scope mouse liver dataset (Figure 4.1 A–D) and Slide-Seq mouse cerebellum dataset (Figure 4.1 E–H), which have $\sim 0.8\mu m$ and $\sim 10\mu m$ distance between adjacent spatial barcodes, respectively. We first applied simple square barcodes aggregation ($100\mu m^2$ for Seq-Scope, $2500\mu m^2$ for Slide-Seq) and then estimated their cell types and UMAP manifolds (Figure 4.1 A and E). STtools is featured at Multi-scale Sliding Window (MSSW, See Materials and Methods for details) analysis by accumulating reads counts in a smaller square grid, to enhance resolution via sliding grid strategy. Using MSSW, we produced 25-fold finer resolution spatial map ($4\mu m^2$ for Seq-Scope and $100\mu m^2$ for Slide-Seq) and performed high-resolution cell-type identification by high-dimensional projection implemented in

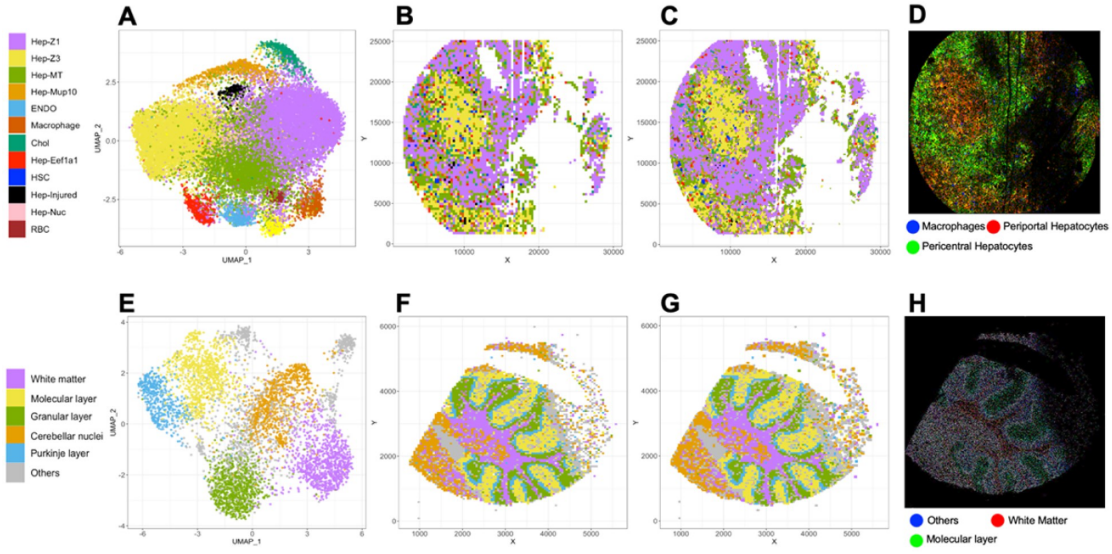

Figure 4.1: Visualization of spatial transcriptomics data with STtools. (A–D) Visualize Seq-Scope mouse liver dataset, and (E–H) visualize Slide-Seq mouse cerebellum dataset. (A, E) Visualize UMAP coordinates and clustered cell types for each squared grid from simple aggregation of 10 μm grids. (B, F) Visualize the clustered cell types for each simple grid ($10\mu m$ for Seq-Scope, $50\mu m$ for Slide-Seq). (C, G) Visualize the cell types from MSSW with higher resolution ($2\mu m$ for Seq-Scope and $10\mu m$ for Slide-Seq). (D, H) Visualize selected marker genes in red, green, and blue (RGB) color at ultra-high resolution ($1\mu m/pixel$ for Seq-Scope and $10\mu m/pixel$ for Slide-Seq)

Seurat (Figure 4.1 C). As a result, the spatial cluster map from MSSW algorithm provides finer cell-type boundaries than simple barcode aggregation (Figure 4.1 B) for Seq-Scope dataset. On the other hand, for Slide-Seq, the benefit of MSSW was not visually pronounced primarily due to the low resolution of the technology. (Figure 4.1 F and G).

To further demonstrate that STtools enables micrometer-resolution cell-type mapping, we produced 25− and 100-fold finer resolution spatial map in Seq-Scope colon datasets(Figure 4.2). 100-fold ($1\mu m$ window size) finer resolution spatial map(Figure 4.2 C) clearly visualize the boundaries between different cell types compartments compared with simple aggregation(Figure 4.2 A) and 25-fold spatial cell-type mapping(Figure 4.2 B).

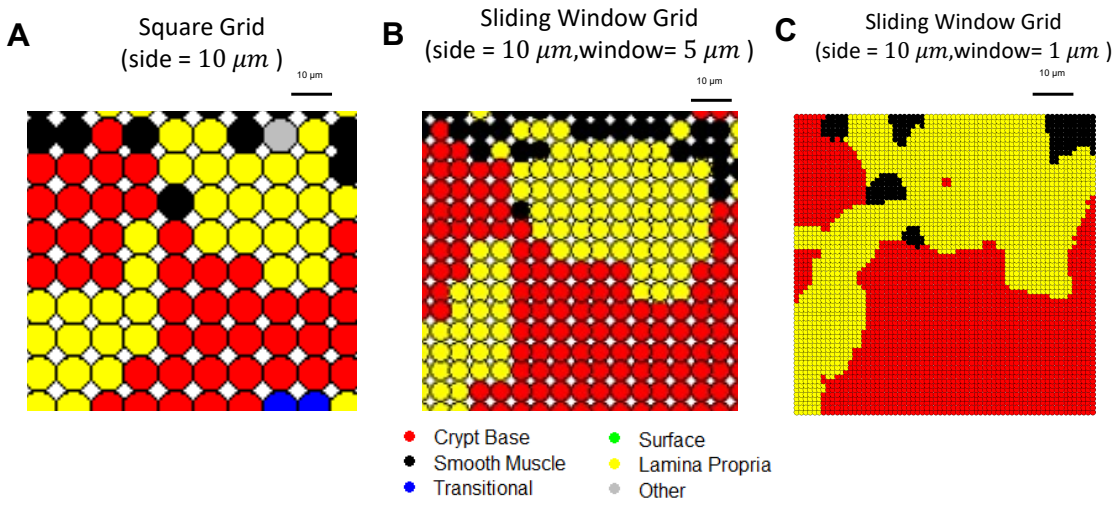

Figure 4.2: Multi-scale Sliding Window (MSSW) analysis enables micrometer-resolution cell-type mapping. (A)Zoomed-in image of spatial map of cell types with simple aggregation of $10\mu m$ sided grids in Seq-Scope colon data. Spatial cell-type mapping in is refined using MSSW analysis with $5\mu m$ (B) and $1\mu m$ (C) intervals.

4.2.2 STtools Visualizes Spatial Gene Expression at Various Scales

We also produced ultra-high-resolution spatial RGB geneset plots that visualize the expressions of selected marker gene sets with STtools to visualize customized spatial maps based on user-defined genes. This RGB plotting tool is capable of separating spliced and unspliced reads, and we were able to visualize both cell-type differences (periportal versus pericentral hepatocytes versus macrophages; Figure 4.1 D, Figure 4.3 A) as well as subcellular differences (e.g. nucleus versus mitochondria versus macrophages; Figure 4.3 C) at a resolution of $1\mu m^2$/pixel. These plots can help investigators interpret ST data at an ultra-high resolution to understand subcellular architecture or infiltration of non-parenchymal cell types.

STtools also generates additional visualization of ST data such as the distribution of UMIs across spatial coordinates (Figure 4.4 A) or violin plots of gene counts, UMI counts or fraction of mitochondrial genes (Figure 4.4B–D) by seamless connection to other single-cell or ST software such as STARsolo (*Kaminow et al.* (2021)), Seurat (*Hao et al.* (2021)) and seqtk (*Li*, 2021). The digital gene expression matrix generated

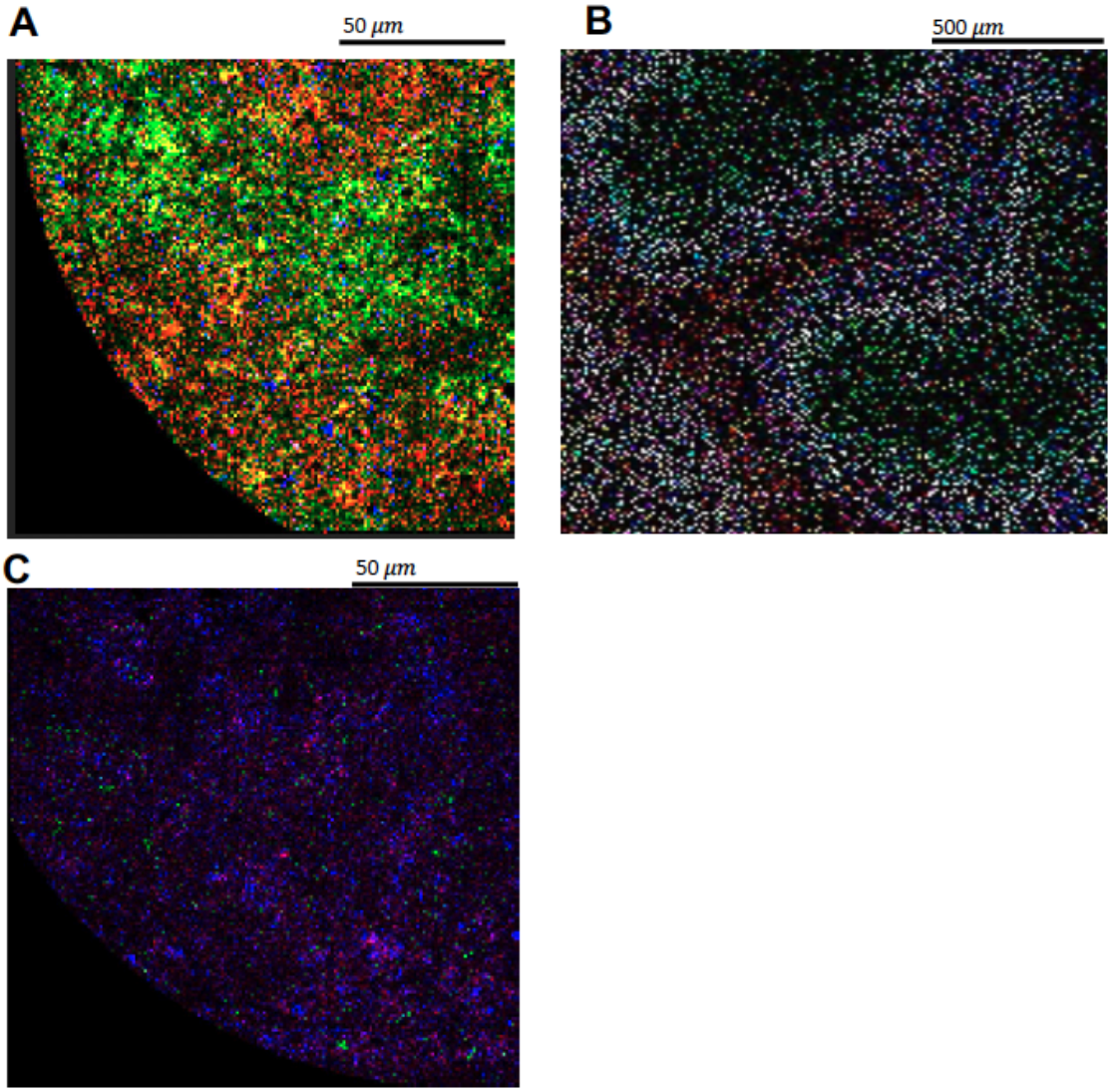

Figure 4.3: Spatial RGB visualization of marker gene sets by STtools. Zoomed-in from Figure 4.1D and H. (A) magnifies the spatial map of Seq-Scope mouse liver from Figure 4.1D, visualizing unspliced reads(blue), periportal hepatocyte(red), and pericentral hepatocyte(green). (B) magnifies the spatial map of Slide-Seq mouse cerebellum from Figure 4.1H, visualizing white matter (red), molecular layer (green) and other cell types (blue). (C) magnifies the spatial map of Seq-Scope mouse liver from Figure 4.1D (same region to A), visualizing mitochondrial RNAs (red), macrophages (green – same to blue in A), unspliced reads (blue).

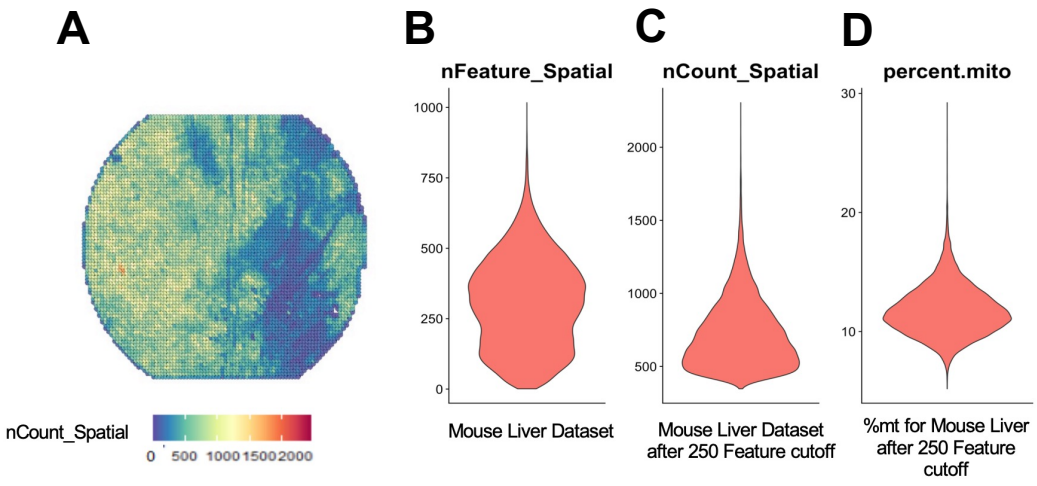

Figure 4.4: Additional visualization of spatial transcriptomics data produced by STtools.(A) –(D) are generated with Seq-Scope mouse liver data(ref). (A) visualizes the spatial distribution of total UMIs per simple square grid ($10\mu m$). (B)-(D) visualizes the the distribution of (B) gene counts (C) total UMIs, and (D) percentage of mitochondrial genes per each grid. (C) and (D) only visualizes grids with more than 250 genes expressed.

by STtools follows the widely used format from 10× Genomics can be directly read from other software tools such as Seurat (*Hao et al.* (2021)) or squidpy (*Palla et al.* (2022)). STtools also offers a functionality to run Bayespace (*Zhao et al.* (2021)) for VISIUM data to enhance its resolution.

4.2.3 STtools Can Efficiently Process Spatial Transcriptomic Data Scaling with Millions of Spatially Resolved Barcodes

STtools is designed to efficiently process ST data scaling with millions of spatially resolved barcodes. The total computational cost to process the Seq-Scope data consisting of 15M spatial barcodes and 1.9 billion raw sequence reads across all stages

was modest, taking $\sim 16h$ in an high performance computing (HPC) cluster with six 3.0 GHz Intel Xeon Gold 6154 CPUs with 30 GB of RAMs for the mouse liver dataset. The cost was orders of magnitude smaller for lower-resolution datasets such as Slide-Seq or VISIUM. To compare the computational efficiency of STtools with spacemake, we ran both tools with the same mouse liver Seq-Scope data (GSM5212844). Spacemake could not handle the full 2nd-seq data with 625.1G bases, so we ran experiments for a subset (SRR14082757) with 104.7G bases. As shown in Table 4.1, spacemake approximately corresponds to STtools steps A1 to A3, which produces the spatial gene expression matrix after alignments. Spacemake does not have the functionality to extract spatial coordinates from 1st-seq data of Seq-Scope, so we used the coordinates generated by STtools step A1. It took 254 min to run STtools steps A1–A3, while spacemake took 1100 min. Both pipelines were executed with —cores 8 option and ran locally on an HP DL380 server with Dual Intel Xeon-G 5118 processor (24 physical cores). The usage of STARsolo instead of STAR and efficiencies in intermediate file generation resulted in significant runtime differences between two pipelines for high-resolution Seq-Scope data.

Table 4.1: Comparison between STtools and other related tools (spacemake and squidpy).

Functionality	**Spacemake**	**Squidpy**	**STtools**
Preprocess 1st-Seq FASTQ to prepare alignment (Step A1)	X	X	O
Quality control of spatial coordinates and tissue boundary detection (Step A2)	X	X	O
Aligns the transcriptomic sequence reads and produces spatial expression matrix (Step A3)	O	X	O
Grid-based simple spatial segmentation (Step C1)	O	X	O
MSSW segmentation (Step C2)	X	X	O
Clustering of each segment (Step C3)	X	O	O
High-resolution visualization of selected genes (Step V1)	X	X	O
Compatible with SlideSeq	O	O	O
Compatible with Seq-Scope	Δ	Δ	O
Provides an end-end solution (including alignment, clustering and visualization)	X	X	O
Allows running individual steps separately	X	O	O
Quantifies both spliced and unspliced reads for subcellular analysis	X	X	O

4.3 Materials and Methods

STtools is able to process ST data from various platforms including, but not limited to, Seq-Scope, Slide-Seq and VISIUM. STtools provides a complete solution from raw FASTQ file preprocessing to automated downstream analysis with the flexibility to run the pipeline end-to-end automatically. It also allows users to run a specified set of consecutive steps, or to run individual steps separately. For example, users can skip the FASTQ processing steps and instead start from a spatial gene expression matrix for downstream analysis such as clustering and visualization using STtools. STtools workflow currently performs three major tasks—alignment, clustering and visualization—consisting of eight individual steps (Figure 4.5). The alignment step performs quality control (QC), alignment and spatial digital gene expression matrix generation from raw sequence data. The clustering steps perform cell-type clustering in Multi-scale resolution. The visualization steps visualize the ST data from multiple different perspectives as illustrated (Figure 4.1; Figure 4.5).

4.3.1 Alignment

The full STtools workflow starts with taking two sets of raw sequence reads in FASTQ format. The first FASTQ file (1st-seq) contains spatial barcode sequences associated with spatial coordinates that are encoded in their Illumina sequence identifiers (Line 1 of FASTQ reads). The second FASTQ file (2nd-seq) contains cDNA sequences from transcripts, attached with the spatial barcodes. To estimate the tissue boundary, the 2nd-Seq were joined into 1st-Seq data according to their HDMI sequence. For each of HDMIs, the tile number and XY coordinates were extracted and assigned. Finally, an HDMI discovery plot was generated to visualize the density of HDMIs in a given XY space of each tiles and were manually assigned to the corresponding H&E images.

After performing initial QC to inspect the distribution of spatial coordinates of

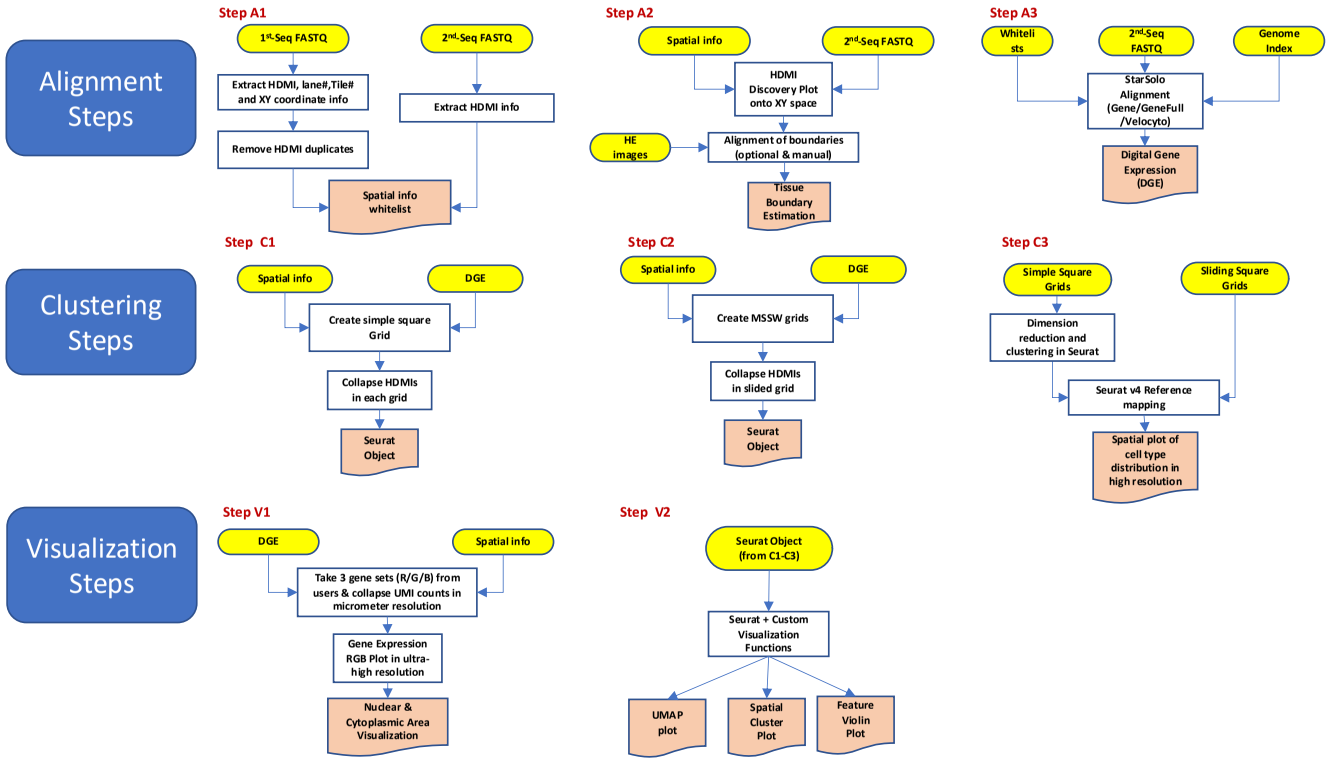

Figure 4.5: STtools workflow. STtools workflow consists of three major steps – (A) Alignment steps, (C) Clustering steps, and (V) Visualization steps The alignment steps starts with two sets of raw FASTQ files to extract spatial coordinates to be used for alignment (A1), visualizes the distribution of spatial barcodes (A2), and perform alignment with STARsolo (A3). The clustering step starts with simple square grid approach to perform initial clustering (C1), followed by Multi-scale Sliding Window (MSSW) to generate high-resolution grids (C2), whose cell types are projected from the initial clusters onto high-resolution grids with Seurat (C3). The visualization steps generates ultra-highresolution RGB plots for user-specified marker gene sets (V1). It also produces UMAP plots, Cluster plots along spatial coordinates, and feature plots for basic QC metrics or user-specified genes by leveraging Seurat (V2)

barcodes, using these two sets of FASTQ files, STtools aligns each cDNA sequence to the reference genome using STARsolo. Each aligned 2nd-seq read in the BAM file is annotated with error-corrected spatial barcodes based on 1st-seq. After the alignment, multiple sets of digital gene expression matrices are generated focusing on exonic reads only (Gene), exonic and intronic reads together (GeneFull), or by distinguishing spliced and unspliced reads (Velocyto).

4.3.2 Two-track Approach for Clustering

STtools takes digital expression matrices annotated with spatial coordinates, either from the steps above or from external sources, to aid interpretation of the data through barcode aggregation, clustering and visualization. Aggregation across nearby spatial barcodes is particularly important for submicrometer resolution ST technologies and will help infer cell types accurately. However, it may compromise the subcellular resolution attainable by the technology. To support clustering at cellular/subcellular level while keeping the details of high spatial resolution, STtools employs two different spatial aggregation (i.e. binning) algorithms: simple aggregation and Multi-scale Sliding Window (MSSW) aggregation.

Due to the extremely high number of HDMI and relatively low number of UMI per HDMI, HDMI-UMI information needs to be aggregated. The simple aggregation method generates a set of non-overlapping, equal-sized bins to capture enough transcripts to be used for cell-type clustering.

MSSW (Figure 4.6) generates a set of overlapping bins for finer resolution cell-type identification and visualization . This two-track approach seamlessly and efficiently integrates with Seurat (*Hao et al.* (2021)), so that simple aggregation is used for clustering cell types and MSSW aggregated bins are used to assign cluster types at a finer resolution.

Algorithm: Multi-scale Sliding Window (MSSW) algorithm

Data: $D[x, y, z]$: gene expression counts of pixel at (x, y) for gene z, $1 \leq x \leq w, 1 \leq y \leq h, 1 \leq z \leq g$.
m_x, m_y : the width and height of a grid
s_x, s_y : the unit of sliding window step.
Result: $(w/s_x) \times (h/s_y)$ image I with each pixel having a cluster id

```
for y ← m_y/2 to h − m_y/2 by s_y do
    for x ← m_x/2 to w − m_x/2 by s_x do
        d ← {0}^g;
        for k_y ← y − m_y/2 + 1 to y + m_y/2 do
            for k_x ← x − m_x/2 + 1 to x + m_x/2 do
                d ← d + D[k_x, k_y, :];
            end
        end
        I[x/s_x, y/s_y] ← refmap(d);
    end
end
/* refmap():N^g → N, maps a Seurat object (with g genes) to
   one of reference clusters identified by simple grids.
   This function can be easily replaced with other mapping
   functions.                                              */
/* All unmapped pixels at the outside boundary region will
   be padded with the nearest mapped                       */
```

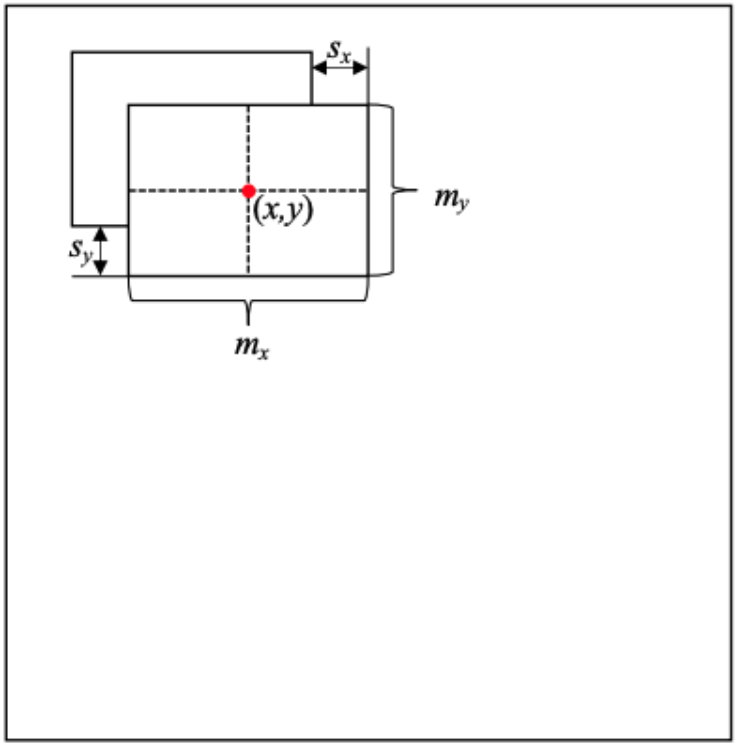

Figure 4.6: Multi Scale Sliding Window (MSSW) Algorithm

4.3.3 Visualization

STtools also generates high-resolution ($< 1\mu m^2$/pixel) images where RGB colors quantify specific arbitrary marker gene sets to help investigators understand the raw spatial gene expression without sacrificing the resolution.

STtools also generates additional visualization of ST data such as the distribution of UMIs across spatial coordinates or violin plots of gene counts, UMI counts or fraction of mitochondrial genes by seamless connection to other single-cell or ST software such as STARsolo (*Kaminow et al.* (2021)), Seurat (*Hao et al.* (2021)) and seqtk (*Li*, 2021).

4.4 Summary

In Chapter IV, we developed STtools, a comprehensive pipeline capable of processing ST data from various platforms including, but not limited to, Seq-Scope, Slide-Seq and VISIUM. STtools provides a complete solution from raw FASTQ file preprocessing to automated downstream analysis with the flexibility to run the pipeline end-to-end automatically. It also allows users to run a specified set of consecutive steps, or to run individual steps separately.

CHAPTER V

Discussion

5.1 Summary

In this dissertation, we presented a robust framework *SiftCell* framework of upstream quality control focusing on the challenges of contaminations from ambient RNAs in single-cell RNA-seq (scRNA-seq) and single-nucleus RNA-seq (snRNA-seq) experiments, introduced Seq-Scope spatial transcriptomics(ST) technologies with comprehensive computational analysis to reveal the histological organization of the transcriptome architecture at multiple scales and proposed STtools pipeline to provide a versatile framework to handle ST datasets with various resolution from different platforms. Here, we review these works, discuss their limitations and suggest possible directions for future research.

5.2 Upstream Quality Control in Single-cell RNA Sequencing with Ambient RNAs

In Chapter II, we described *SiftCell* framework, a suite of software tools implementing methods including *SiftCell-Shuffle*, *SiftCell-Boost* and *SiftCell-Mix*, focusing on the challenges of contaminations from ambient RNAs in single-cell and single-nucelus RNA-seq experiments. *SiftCell-Shuffle* works with digital gene expression

matrix and aids the investigators to visually distinguish cell-free from cell-containing droplets by contrasting with a randomized digital gene expression matrix. *SiftCell-Boost* takes the output of *SiftCell-Shuffle* as input and applies a semi-supervised machine learning method to classify cell-containing droplets and cell-free droplets. *SiftCell-Mix* is a model-based tool that allows quantitative estimation of the contribution of "ambient RNAs" in each droplet.

Compared to existing methods (*Lun et al.* (2019),*Alvarez et al.* (2020),*Fleming et al.* (2019), *Yang et al.* (2020)) *SiftCell-Boost* and *SiftCell-Mix* consistently performed better than or comparably with the best-performing methods across the three datasets, posed by different types of challenges. Most methods performed well for the PBMC dataset, which is the most recognized single-cell dataset, but many methods struggled with snRNA-seq(brain nuclei), or scRNA-seq generated with Drop-Seq (colon cell line mixture). We also noticed that existing tools do not provide effective visualization to understand the quality of scRNA-seq data in terms of ambient RNA contamination, and we believe that *SiftCell-Shuffle* is a unique tool that allows users to visually interpret the spectrum of all barcoded droplets. It should be noted, however, that *SiftCell-Shuffle* offers only quasi-ground truth under the assumption that randomized droplets are good representatives of cell-free droplets.

The *SiftCell* framework can be easily adopted for other quality control methods for single-cell genomic data. Existing methods for identifying cell-containing droplets may be improved by incorporating the results from *SiftCell-Shuffle*. The key idea underlying *SiftCell-Shuffle*, *SiftCell-Boost*, *SiftCell-Mix* is not necessarily limited to scRNA-seq or snRNA-seq, so it should be possible to apply the same principle to scATAC-seq data or single cell multiome datasets, even though some tweaks may be required to optimize its performance.

There are rooms for further improvements in *SiftCell-Shuffle*. For example, it may be better to assume that ambient RNAs are not a totally random sample of the

pseudo-bulk scRNA-seq reads. In fact, there are studies demonstrating that ambient RNAs are enriched for specific features, such as mitochondrial genes or necrosis marker genes(*Muskovic and Powell* (2021)). Our current approach to randomly shuffle barcodes droplets for *SiftCell-Shuffle*, but provides an option to remove specific genes that are determined to be enriched or depleted in cell-free droplets. Our method can be further extended to a non-random permutation or bootstrapping, and how to define a better "null" distribution of ambient RNAs is a subject of further research.

Binary classification of droplets into cell-containing/cell-free droplets with *SiftCell-Boost* may make the downstream analysis simpler, but a more sophisticated procedure is needed to handle datasets with heavy contamination from ambient RNAs. In such cases, estimated from *SiftCell-Mix* can inform the quality of the classification results. For example, in brain nuclei dataset, *SiftCell-Mix* estimates that 27.0% of cell-containing droplets (inferred by *SiftCell-Boost*) have $> 10\%$ of ambient RNAs present. This is substantially larger than 4.7% for PBMC, and 9.9% for colon cell line mixture, suggesting the importance of accounting for ambient RNAs when analyzing snRNA-seq data. In the colon cell line mixture dataset, we observed that droplets containing multiple cells (multiplets) tended to be classified as cell-free droplets more often than true single cells because mixture of multiple cell types tend to be more similar to ambient RNAs.

Although *SiftCell-Mix* provides quantitative estimation of the contribution from ambient RNAs, when the number of reads per cell is limited, we noticed that the estimates can be quite unstable under our maximum-likelihood framework. Imposing a stronger prior under Bayesian framework may make the estimation of more stable for sparse data.

5.3 Spatial Transcriptomics Technique with High Resolution

In Chapter III, we introduced Seq-Scope ST technique that achieves submicrometer resolution. Through comprehensive computational analysis, we have demonstrated that Seq-Scope can reveal spatial single cell and subcellular analysis of liver and colon tissues. Equipped with an ultra-high-resolution output and an outstanding transcriptome capture output, Seq-Scope drew a clear boundary between different tissue zones, cell types, and subcellular components. Previously existing technologies could not provide this level of clarity due to their low-resolution output and/or inefficiency in transcriptome capture.

There are more improvement that can be made to Seq-Scope technology. In the current study, we used the MiSeq platform to generate the HDMI arrays; however, virtually any sequencing platforms that use spatially localized amplification, such as Illumina GAIIx, HiSeq, NextSeq, and NovaSeq, could be used to generate the HDMI-arrays. Although MiSeq has small imaging areas, HiSeq2500 and NovaSeq can provide $\sim 90mm^2$ and $\sim 800mm^2$ of the uninterrupted imaging area, respectively, providing a larger field of view. Newer sequencing methods, such as NextSeq and NovaSeq, are based on a patterned flow cell technology, which could provide more defined spatial information for the HDMI-encoded clusters.

5.4 Tools for High Resolution Spatial Transcriptomics

In Chapter IV, we developed STtools, a software pipeline that allows users to align, cluster and visualize ST sequence data generated at submicrometer resolution. In particular, STtools improves the resolution of spatial inference compared to typical segmentation-based approach by leveraging MSSW algorithm. The spatial expression matrix, spatial segmentation and clustering results produced by STtools can be easily fed into other software tools widely used for downstream analysis, such as Seurat (*Hao*

et al. (2021)) and squidpy (*Palla et al.* (2022)).

While STtools offers all-in-one analysis to translate raw sequence reads into spatial expression matrix and clustering, it also provides options to perform step-by-step analysis so that the investigators can perform sanity checks at each step and adjust the parameters as needed. Users can customize many parameters during the alignment and clustering, including adapter sequences to trim, reference genomes to align and the thresholds to filter genes and spatial segments before clustering. Users can always load the spatial expression matrix generated by STtools in a standard format to perform more tailored analysis on their own using Seurat, squidpy or other downstream software tools.

Although higher-resolution spatial inference can be made by the MSSW algorithm, compared to other standard spatial transcriptomic analysis tools, it still has room for improvement. Due to the limited number of UMIs per region, each spatial segment still needs to be larger than subcellular compartments (e.g. $\sim 10\mu m$), so subcellular analysis with MSSW is not feasible. Spatial smoothing algorithms that deliver robust inference for extremely sparse expression profiles per spatial unit will be needed to enable truly subcellular inference beyond visualization of subcellular compartments.

There are many more improvements that can be made to STtools in the future. For example, methods to impute spatial expression profiles (*Shengquan et al.* (2021)), methods to jointly cluster cellular and subcellular components together, or methods to automatically overlay histological images and spatial expressions are useful features that can be added in the next major updates of STtools.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aizarani, N., A. Saviano, L. Mailly, S. Durand, J. S. Herman, P. Pessaux, T. F. Baumert, and D. Grün (2019), A human liver cell atlas reveals heterogeneity and epithelial progenitors, *Nature*, *572*(7768), 199–204.

Alvarez, M., et al. (2020), Enhancing droplet-based single-nucleus rna-seq resolution using the semi-supervised machine learning classifier diem, *Scientific reports*, *10*(1), 11,019.

Asp, M., J. Bergenstråhle, and J. Lundeberg (2020), Spatially resolved transcriptomes—next generation tools for tissue exploration, *BioEssays*, *42*(10), 1900,221.

Baratta, J. L., A. Ngo, B. Lopez, N. Kasabwalla, K. J. Longmuir, and R. T. Robertson (2009), Cellular organization of normal mouse liver: a histological, quantitative immunocytochemical, and fine structural analysis, *Histochemistry and cell biology*, *131*, 713–726.

Becht, E., L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell (2019), Dimensionality reduction for visualizing single-cell data using umap, *Nature biotechnology*, *37*(1), 38–44.

Ben-Moshe, S., and S. Itzkovitz (2019), Spatial heterogeneity in the mammalian liver, *Nature reviews Gastroenterology & hepatology*, *16*(7), 395–410.

Bentley, D. R., et al. (2008), Accurate whole human genome sequencing using reversible terminator chemistry, *nature*, *456*(7218), 53–59.

Bergenstråhle, J., L. Larsson, and J. Lundeberg (2020), Seamless integration of image and molecular analysis for spatial transcriptomics workflows, *BMC genomics*, *21*(1), 1–7.

Borenshtein, D., K. A. Schlieper, B. H. Rickman, J. M. Chapman, C. W. Schweinfest, J. G. Fox, and D. B. Schauer (2009), Decreased expression of colonic slc26a3 and carbonic anhydrase iv as a cause of fatal infectious diarrhea in mice, *Infection and immunity*, *77*(9), 3639–3650.

Borenstein, M., L. V. Hedges, J. P. Higgins, and H. R. Rothstein (2021), *Introduction to meta-analysis*, John Wiley & Sons.

Buenrostro, J. D., B. Wu, H. Y. Chang, and W. J. Greenleaf (2015), Atac-seq: a method for assaying chromatin accessibility genome-wide, *Current protocols in molecular biology*, *109*(1), 21–29.

Butler, A., P. Hoffman, P. Smibert, E. Papalexi, and R. Satija (2018), Integrating single-cell transcriptomic data across different conditions, technologies, and species, *Nature biotechnology*, *36*(5), 411–420.

Cao, J., et al. (2017), Comprehensive single-cell transcriptional profiling of a multicellular organism, *Science*, *357*(6352), 661–667.

Chen, T., and C. Guestrin (2016), Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Cho, C.-S., J. Xi, Y. Si, S.-R. Park, J.-E. Hsu, M. Kim, G. Jun, H. M. Kang, and J. H. Lee (2021), Microscopic examination of spatial transcriptome using seq-scope, *Cell*, *184*(13), 3559–3572.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the em algorithm, *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2013), Star: ultrafast universal rna-seq aligner, *Bioinformatics*, *29*(1), 15–21.

Donne, R., M. Saroul-Aïnama, P. Cordier, S. Celton-Morizur, and C. Desdouets (2020), Polyploidy in liver development, homeostasis and disease, *Nature Reviews Gastroenterology & Hepatology*, *17*(7), 391–405.

Dou, L., X. Shi, X. He, and Y. Gao (2020), Macrophage phenotype and function in liver disorder, *Frontiers in immunology*, *10*, 3112.

Durinck, S., P. T. Spellman, E. Birney, and W. Huber (2009), Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart, *Nature protocols*, *4*(8), 1184–1191.

Eckhardt, E. R., et al. (2010), Intestinal epithelial serum amyloid a modulates bacterial growth in vitro and pro-inflammatory responses in mouse experimental colitis, *BMC gastroenterology*, *10*(1), 1–9.

Farkas, A. E., C. Gerner-Smidt, L. Lili, A. Nusrat, and C. T. Capaldo (2015), Cryosectioning method for microdissection of murine colonic mucosa, *Journal of visualized experiments: JoVE*, (101).

Fischer, H., R. Stenling, C. Rubio, and A. Lindblom (2001), Differential expression of aquaporin 8 in human colonic epithelial cells and colorectal tumors, *BMC physiology*, *1*, 1–5.

Fisher, R. A. (1925), *Statistical methods for research workers*, Oliver and Boyd.

Fleiss, J. L. (1993), Review papers: The statistical basis of meta-analysis, *Statistical methods in medical research*, *2*(2), 121–145.

Fleming, S. J., J. C. Marioni, and M. Babadi (2019), Cellbender remove-background: a deep generative model for unsupervised removal of background noise from scrna-seq datasets, *BioRxiv*, *791699*.

Garcia-Alonso, L., et al. (2021), Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro, *Nature Genetics*, *53*(12), 1698–1711.

Griffiths, J. A., A. C. Richard, K. Bach, A. T. Lun, and J. C. Marioni (2018), Detection and removal of barcode swapping in single-cell rna-seq data, *Nature communications*, *9*(1), 2667.

Grün, D., and A. van Oudenaarden (2015), Design and analysis of single-cell sequencing experiments, *Cell*, *163*(4), 799–810.

Haber, A. L., et al. (2017), A single-cell survey of the small intestinal epithelium, *Nature*, *551*(7680), 333–339.

Habib, N., et al. (2017), Massively parallel single-nucleus rna-seq with dronc-seq, *Nature methods*, *14*(10), 955–958.

Halpern, K. B., I. Caspi, D. Lemze, M. Levy, S. Landen, E. Elinav, I. Ulitsky, and S. Itzkovitz (2015), Nuclear retention of mrna in mammalian tissues, *Cell reports*, *13*(12), 2653–2662.

Halpern, K. B., et al. (2017), Single-cell spatial reconstruction reveals global division of labour in the mammalian liver, *Nature*, *542*(7641), 352–356.

Hao, Y., et al. (2021), Integrated analysis of multimodal single-cell data, *Cell*, *184*(13), 3573–3587.

Heaton, H., A. M. Talman, A. Knights, M. Imaz, D. J. Gaffney, R. Durbin, M. Hemberg, and M. K. Lawniczak (2020), Souporcell: robust clustering of single-cell rna-seq data by genotype without reference genotypes, *Nature methods*, *17*(6), 615–620.

Hildebrandt, F., et al. (2021), Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver, *Nature communications*, *12*(1), 7046.

J. Sweeting, M., A. J. Sutton, and P. C. Lambert (2004), What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data, *Statistics in medicine*, *23*(9), 1351–1375.

Jiang, Y., et al. (2016), An expanded evaluation of protein function prediction methods shows an improvement in accuracy, *Genome biology*, *17*(1), 1–19.

Jovic, D., X. Liang, H. Zeng, L. Lin, F. Xu, and Y. Luo (2022), Single-cell rna sequencing technologies and applications: A brief overview, *Clinical and Translational Medicine*, *12*(3), e694.

Kaminow, B., D. Yunusov, and A. Dobin (2021), Starsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus rna-seq data, *Biorxiv*, pp. 2021–05.

Kim, B. S., and B. H. Margolin (1992), Testing goodness of fit of a multinomial model against overdispersed alternatives, *Biometrics*, pp. 711–719.

Kiselev, V. Y., et al. (2017), Sc3: consensus clustering of single-cell rna-seq data, *Nature methods*, *14*(5), 483–486.

Klein, A. M., L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner (2015), Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, *Cell*, *161*(5), 1187–1201.

Kornilov, A. S., and I. V. Safonov (2018), An overview of watershed algorithm implementations in open source libraries, *Journal of Imaging*, *4*(10), 123.

La Manno, G., et al. (2018), Rna velocity of single cells, *Nature*, *560*(7719), 494–498.

Levine, D. S., and R. C. Haggitt (1989), Normal histology of the colon, *The American journal of surgical pathology*, *13*(11), 966–984.

Li, X., and C.-Y. Wang (2021), From bulk, single-cell to spatial rna sequencing, *International Journal of Oral Science*, *13*(1), 36.

Lin, D.-Y., and D. Zeng (2010), On the relative efficiency of using summary statistics versus individual-level data in meta-analysis, *Biometrika*, *97*(2), 321–332.

Lipták, T. (1958), On the combination of independent tests, *Magyar Tud Akad Mat Kutato Int Kozl*, *3*, 171–197.

Liu, D., R. Y. Liu, and M.-g. Xie (2014), Exact meta-analysis approach for discrete data and its application to 2× 2 tables with rare events, *Journal of the American Statistical Association*, *109*(508), 1450–1465.

Liu, Y., et al. (2020), High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue, *Cell*, *183*(6), 1665–1681.

Lun, A. T., S. Riesenfeld, T. Andrews, T. Gomes, J. C. Marioni, et al. (2019), Emptydrops: distinguishing cells from empty droplets in droplet-based single-cell rna sequencing data, *Genome biology*, *20*(1), 1–9.

Macosko, E. Z., et al. (2015), Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, *Cell*, *161*(5), 1202–1214.

Martin, D. O., and H. Austin (2000), An exact method for meta-analysis of case-control and follow-up studies, *Epidemiology*, pp. 255–260.

Marx, V. (2021), Method of the year: spatially resolved transcriptomics, *Nature methods*, *18*(1), 9–14.

Muskovic, W., and J. E. Powell (2021), Dropletqc: improved identification of empty droplets and damaged cells in single-cell rna-seq data, *Genome Biology*, *22*, 1–9.

Nestorowa, S., F. K. Hamey, B. Pijuan Sala, E. Diamanti, M. Shepherd, E. Laurenti, N. K. Wilson, D. G. Kent, and B. Göttgens (2016), A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation, *Blood, The Journal of the American Society of Hematology*, *128*(8), e20–e31.

Ntranos, V., L. Yi, P. Melsted, and L. Pachter (2019), A discriminative learning approach to differential expression analysis for single-cell rna-seq, *Nature methods*, *16*(2), 163–166.

Ozsolak, F., and P. M. Milos (2011), Rna sequencing: advances, challenges and opportunities, *Nature reviews genetics*, *12*(2), 87–98.

O'Flanagan, C. H., et al. (2019), Dissociation of solid tumor tissues with cold active protease for single-cell rna-seq minimizes conserved collagenase-associated stress responses, *Genome biology*, *20*(1), 1–13.

Palla, G., et al. (2022), Squidpy: a scalable framework for spatial omics analysis, *Nature methods*, *19*(2), 171–178.

Park, S. R., C.-S. Cho, J. Xi, H. M. Kang, and J. H. Lee (2021), Holistic characterization of single-hepatocyte transcriptome responses to high-fat diet, *American Journal of Physiology-Endocrinology and Metabolism*, *320*(2), E244–E258.

Park, S. R., et al. (2020), Single-cell transcriptome analysis of colon cancer cell response to 5-fluorouracil-induced dna damage, *Cell reports*, *32*(8), 108,077.

Pelaseyed, T., et al. (2014), The mucus and mucins of the goblet cells and enterocytes provide the first defense line of the gastrointestinal tract and interact with the immune system, *Immunological reviews*, *260*(1), 8–20.

Petukhov, V., R. J. Xu, R. A. Soldatov, P. Cadinu, K. Khodosevich, J. R. Moffitt, and P. V. Kharchenko (2022), Cell segmentation in imaging-based spatial transcriptomics, *Nature biotechnology*, *40*(3), 345–354.

Preissl, S., et al. (2018), Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation, *Nature neuroscience*, *21*(3), 432–439.

Qiu, X., Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell (2017), Reversed graph embedding resolves complex single-cell trajectories, *Nature methods*, *14*(10), 979–982.

Richardson, D. B., S. R. Cole, R. K. Ross, C. Poole, H. Chu, and A. P. Keil (2021), Meta-analysis and sparse-data bias, *American journal of epidemiology*, *190*(2), 336–340.

RMJ, S. (1949), The american soldier, vol. 1: Adjustment during army life.

Rodriques, S. G., et al. (2019), Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution, *Science*, *363*(6434), 1463–1467.

Rubin-Delanchy, P., N. A. Heard, and D. J. Lawson (2019), Meta-analysis of mid-p-values: some new results based on the convex order, *Journal of the American Statistical Association*, *114*(527), 1105–1112.

Sage, D., and M. Unser (2003), Teaching image-processing programming in java, *IEEE Signal Processing Magazine*, *20*(6), 43–52.

Schiebinger, G., et al. (2019), Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming, *Cell*, *176*(4), 928–943.

Setty, M., et al. (2016), Wishbone identifies bifurcating developmental trajectories from single-cell data, *Nature biotechnology*, *34*(6), 637–645.

Shannon, H. (2016), A statistical note on karl pearson's 1904 meta-analysis, *Journal of the Royal Society of Medicine*, *109*(8), 310–311.

Shengquan, C., Z. Boheng, C. Xiaoyang, Z. Xuegong, and J. Rui (2021), stplus: a reference-based method for the accurate enhancement of spatial transcriptomics, *Bioinformatics*, *37*(Supplement_1), i299–i307.

Soneson, C., and M. D. Robinson (2018), Bias, robustness and scalability in single-cell differential expression analysis, *Nature methods*, *15*(4), 255–261.

Spencer, J., and L. M. Sollid (2016), The human intestinal b-cell response, *Mucosal immunology*, *9*(5), 1113–1124.

Ståhl, P. L., et al. (2016), Visualization and analysis of gene expression in tissue sections by spatial transcriptomics, *Science*, *353*(6294), 78–82.

Stark, R., M. Grzelak, and J. Hadfield (2019), Rna sequencing: the teenage years, *Nature Reviews Genetics*, *20*(11), 631–656.

Stickels, R. R., E. Murray, P. Kumar, J. Li, J. L. Marshall, D. J. Di Bella, P. Arlotta, E. Z. Macosko, and F. Chen (2021), Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2, *Nature biotechnology*, *39*(3), 313–319.

Stoeckius, M., S. Zheng, B. Houck-Loomis, S. Hao, B. Z. Yeung, W. M. Mauck, P. Smibert, and R. Satija (2018), Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics, *Genome biology*, *19*(1), 1–12.

Tarone, R. E. (1979), Testing the goodness of fit of the binomial distribution, *Biometrika*, *66*(3), 585–590.

Tsoucas, D., and G.-C. Yuan (2018), Giniclust2: a cluster-aware, weighted ensemble clustering method for cell-type detection, *Genome biology*, *19*, 1–13.

Vickovic, S., et al. (2019), High-definition spatial transcriptomics for in situ tissue profiling, *Nature methods*, *16*(10), 987–990.

Volovitz, I., et al. (2016), A non-aggressive, highly efficient, enzymatic method for dissociation of human brain-tumors and brain-tissues to viable single-cells, *BMC neuroscience*, *17*, 1–10.

Waltman, L., and N. J. Van Eck (2013), A smart local moving algorithm for large-scale modularity-based community detection, *The European physical journal B*, *86*, 1–14.

Wang, Z., M. Gerstein, and M. Snyder (2009), Rna-seq: a revolutionary tool for transcriptomics, *Nature reviews genetics*, *10*(1), 57–63.

Welch, J. D., A. J. Hartemink, and J. F. Prins (2016), Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data, *Genome biology*, *17*(1), 1–15.

Williams, C. G., H. J. Lee, T. Asatsuma, R. Vento-Tormo, and A. Haque (2022), An introduction to spatial transcriptomics for biomedical research, *Genome Medicine*, *14*(1), 1–18.

Yang, S., S. E. Corbett, Y. Koga, Z. Wang, W. E. Johnson, M. Yajima, and J. D. Campbell (2020), Decontamination of ambient rna in single-cell rna-seq with decontx, *Genome biology*, *21*, 1–15.

Young, M. D., and S. Behjati (2020), Soupx removes ambient rna contamination from droplet-based single-cell rna sequencing data, *Gigascience*, *9*(12), giaa151.

Zhao, E., et al. (2021), Spatial transcriptomics at subspot resolution with bayesspace, *Nature Biotechnology*, *39*(11), 1375–1384.

Zheng, G. X., et al. (2017), Massively parallel digital transcriptional profiling of single cells, *Nature communications*, *8*(1), 14,049.