Statistical and Computational Methods for High-Dimensional Genomics Data

by

Ying Ma

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:

Associate Professor Xiang Zhou, Chair
Associate Research Scientist Lars G. Fritsche
Assistant Professor Jean Morrison
Assistant Professor Joshua Welch

Ying Ma

yingma@umich.edu

ORCID iD: 0000-0003-3791-7018

Dedication

To my husband and my parents,

Endless respect for learning, love, and being loved.

Acknowledgements

I am deeply grateful for the opportunity to pursue my Ph.D. studies in the Department of Biostatistics at the University of Michigan. It has been a privilege to be immersed in a community of erudite, dedicated, and amicable individuals who have profoundly influenced my personal and academic growth. The uplifting and challenging experiences I have encountered throughout my doctoral studies have left indelible memories, shaped my resilience, and served as a wellspring of inspiration for the future.

First and foremost, I would like to extend my deepest gratitude to my exceptional advisor, Dr. Xiang Zhou. Under his mentorship, I was introduced to the captivating world of statistical genomics and genetics. Throughout my academic journey, research pursuits, and professional development, he has been a constant source of motivation and strength. Dr. Zhou's dedication to scientific rigor and genuine passion for this field have been truly inspiring. Whenever I encountered challenges in my research projects, his encouragement, patience, and unwavering support were invaluable. He is not only my advisor but also a guide and a role model for my future career. I am profoundly grateful for his mentorship throughout my doctorate journey, and I will forever cherish the knowledge and experience he shared with me.

I would also like to thank my dissertation committee members, Dr. Lars Fritsche, Dr. Jean Morrison, and Dr. Joshua Welch, for their invaluable time and commitment as committee members. Their guidance, feedback, and expertise have contributed to shaping my thinking and enhancing the quality of the dissertation. It has been an honor to learn from their wealth of knowledge and experience. I'm also deeply grateful for their support during my job search process. Additionally,

I am sincerely grateful to my collaborators Dr. Bhramar Mukherjee, and Dr. Lars Fritsche for their scientific guidance and valuable contributions throughout our collaborations. Their expertise in polygenic risk scores, insightful perspectives, constructive feedback during our enlightening discussions have continually inspired me to delve deeper into this field.

The experience of working and learning at Umich Biostatistics has been truly enriching. The supportive community among faculty, staff, and fellow students has been exceptional. I am deeply grateful for the support of the administrative team, whose contributions have greatly enhanced the well-being of Ph.D. students. Special thanks to Kirsten Herold for her assistance with both scientific and professional writing.

Being part of Dr. Zhou's lab has been a cherished memory. I would like to extend my gratitude to all previous and current members of the lab for their support, help, and engaging discussions that broadened my knowledge and motivated my methodological development. I am thankful for the time we have spent together, both in the office and within the SPH building. My heartfelt gratitude goes to my dear friends at U of M, Di Wang, Jiongming Wang, Irina Zhang, Shuze Wang, Xubo Yue, Hojae Lee, for their generous help and support. I am also thankful for the enduring friendship of Yuanquan Zhou and Tianjin Fang, who have always been there.

It is truly beyond words to convey my gratitude for having embarked on this journey with my husband, Chao Gao, who gave me unparalleled support, encouragement, and love. We met in Ann Arbor, became engaged, and married throughout this challenging journey of pursing our own Ph.D. degrees. Thank you for being my loving partner, best friend, and steadfast supporter. Lastly, I want to thank my brother, sister-in-law, niece, my parents for their unconditional love throughout it all.

Table of Contents

Dedication ..... **ii**

Acknowledgements ..... **iii**

List of Tables ..... **ix**

List of Figures ..... **x**

List of Appendices ..... **xvii**

Abstract ..... **xviii**

Chapter 1 Introduction **1**

1.1 Background 1

1.2 Single-cell RNA-sequencing 1

1.3 Spatially resolved transcriptomics 2

1.4 Challenges in transcriptomics data analysis 4

1.4.1 Differential expression and gene set enrichment analysis 4

1.4.2 Cell type deconvolution 5

1.4.3 Spatial domain detection 6

1.5 Dissertation outline 7

Chapter 2 Integrative Differential Expression and Gene Set Enrichment Analysis Using Summary Statistics for scRNA-seq Studies. **9**

2.1 Abstract 9

2.2 Introduction 9

2.3 Results 11

2.3.1 Methods overview and simulation design 11

2.3.2 Simulation results 13
2.3.3 Human embryonic stem cell scRNA-seq data 18
2.3.4 Mouse sensory neuron scRNA-seq data 21
2.3.5 10x Genomics PBMC scRNA-seq data 24
2.4 Discussion 28
2.5 Methods 30
2.5.1 iDEA overview 30
2.5.2 Summary statistics and gene annotations 33
2.5.3 Compared methods 34
2.5.4 Simulations 36
2.5.5 scRNA-seq datasets 39
2.5.6 Sensitivity analysis 42
2.5.7 Data and code availability 43
2.6 Supplementary Figures 43
2.7 Supplementary Tables 59
Chapter 3 Spatially Informed Cell Type Deconvolution for Spatial Transcriptomics 64
3.1 Abstract 64
3.2 Introduction 64
3.3 Results 67
3.3.1 Simulations 67
3.3.2 Mouse olfactory bulb data 71
3.3.3 Human pancreatic ductal adenocarcinomas data 75
3.3.4 Mouse hippocampus data from multiple sources 79
3.3.5 Extension of CARD for reference-free deconvolution 83
3.4 Discussion 84

3.5 Methods 86

3.5.1 CARD method overview 86

3.5.2 Imputation and construction of high-resolution spatial maps for cell type composition and gene expression 90

3.5.3 Basis matrix construction 91

3.5.4 Simulations and deconvolution analysis evaluation 92

3.5.5 Compared methods 92

3.5.6 Real data analyses 94

3.5.7 Data and code availability 96

3.6 Supplementary Figures 97

3.7 Supplementary Tables 134

Chapter 4 Accurate and Efficient Integrative Reference-Informed Spatial Domain Detection for Spatial Transcriptomics 136

4.1 Abstract 136

4.2 Introduction 137

4.3 Results 140

4.3.1 Method overview 140

4.3.2 Human dorsolateral prefrontal cortex 10x Visium data 142

4.3.3 Mouse spermatogenesis Slide-seq data 146

4.3.4 High resolution mouse olfactory bulb Stereo-seq data 152

4.3.5 High resolution human breast cancer 10x Xenium data 155

4.4 Discussion 160

4.5 Methods 164

4.5.1 IRIS method overview 164

4.5.2 Compared methods for spatial domain detection 168

4.5.3 Real data analysis 169

4.5.4 Data and code availability 174
4.6 Supplementary Figures 175
Chapter 5 Conclusion 193
Appendices 198
Appendix A. Chapter 2 (iDEA) Supplementary Text. 199
Appendix B. Chapter 3 (CARD) Supplementary Text 210
Appendix C. Chapter 4 (IRIS) Supplementary Text 231
Bibliography 237

List of Tables

Table S2.1 Top 10 enriched gene sets identified by iDEA on the human embryonic stem cell scRNA-seq dataset. 59

Table S2.2 Top 10 enriched gene sets identified by iDEA on the mouse neuronal cell scRNA-seq dataset. 60

Table S2.3 Top 10 enriched gene sets identified by iDEA on the 10x Genomics PBMC scRNA-seq dataset. 60

Table S2.4 Top 10 gene sets identified by fGSEA, CAMERA, PAGE, GSEA respectively. 62

Table S2.5 Results for the top gene set GO:0001944 with the combinations of the top 5 gene sets in the human embryonic stem cell scRNA-seq dataset. 63

Table S2.6 Results for the top gene set GO:0044425 with the combinations of the top 5 gene sets in the mouse neuronal cell scRNA-seq dataset. 63

Table S3.1 Results of Moran's I and Geary's C spatial statistical tests in real data applications. 134

Table S3.2 List of 5 spatially resolved transcriptomics datasets and 10 scRNA-seq datasets we used in our analysis. 135

List of Figures

Figure 2.1 Schematic overview of iDEA 14

Figure 2.2 iDEA produces well-calibrated p-values for gene set enrichment analysis under null simulations. 15

Figure 2.3 iDEA is more powerful for both GSE and DE analyses than existing approaches in power simulations. 18

Figure 2.4 Analysis results in the embryonic stem cell scRNA-seq data. 21

Figure 2.5 Analysis results in the mouse neuronal cell scRNA-seq data. 24

Figure 2.6 Analysis results in the 10X Genomics scRNA-seq data. 27

Figure S2.1 Characteristics of simulated data. 44

Figure S2.2 iDEA produces well-calibrated p-values for gene set enrichment analysis under null simulations. 44

Figure S2.3 iDEA is more powerful than GSE methods for identifying enriched gene sets under alternative simulations. 45

Figure S2.4 iDEA is more powerful for both GSE and DE analyses than existing approaches in power simulations. 46

Figure S2.5 iDEA is more powerful than DE methods for identifying DE genes under alternative simulations when gene set enrichment parameter is larger. 47

Figure S2.6 iDEA is more powerful in DE analysis than zingeR, when varying $\tau 1$ and CR. 47

Figure S2.7 Distribution of marginal DE p-values from common DE methods. 48

Figure S2.8 iDEA produces calibrated (or slightly conservative) FDR estimates. 48

Figure S2.9 iDEA displays high consistency in detecting DE genes in simulations. 49

Figure S2.10 GSE Analysis including hypergeometric test results in human embryonic stem cell scRNA-seq dataset. 49

Figure S2.11 iDEA displays high consistency in detecting DE genes in human embryonic stem cell scRNA-seq data. 50

Figure S2.12 iDEA displays high consistency in detecting DE genes in human embryonic stem cell scRNA-seq data. 51

Figure S2.13 GSE Analysis including hypergeometric test results in mouse neuronal cell scRNA-seq dataset. 52

Figure S2.14 iDEA displays high consistency in detecting DE genes in mouse neuronal cell scRNA-seq data. 52

Figure S2.15 iDEA displays high consistency in detecting DE genes in mouse neuronal cell scRNA-seq data. 53

Figure S2.16 The scatterplot of first two t-SNE principal components for 10x Genomics data set. 53

Figure S2.17 GSE Analysis including hypergeometric test results in 10x Genomics data set. 54

Figure S2.18 iDEA displays high consistency in detecting DE genes in 10x Genomics scRNA-seq data. 55

Figure S2.19 iDEA displays high consistency in detecting DE genes in 10x Genomics scRNA-seq data. 55

Figure S2.20 DE analysis results in the 10X Genomics scRNA-seq data. 56

Figure S2.21 Sensitivity analysis of hyperparameters in prior distribution of $\sigma\beta 2$. 56

Figure S2.22 Type I error rate in real datasets. 57

Figure S2.23 Analysis results in the bulk RNAseq data. 57

Figure S2.24 Posterior inclusion probabilities (PIPs) calculated by iDEA when adding specific gene set is highly correlated with averaging PIPs across all gene sets in all three scRNA-seq datasets. 58

Figure S2.25 iDEA produces calibrated p-values in scRNA-seq based null simulations when using Louis Method to correct the observed information matrix. 59

Figure 3.1 Schematic overview of CARD 68

Figure 3.2 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenarios I-V. 70

Figure 3.3 Analyzing the mouse olfactory bulb data. 74

Figure 3.4 Analyzing the pancreatic ductal adenocarcinoma (PDAC) data. 78

Figure 3.5 Analyzing the hippocampus region in the Slide-seq V2 and 10x Visium Mouse Brain (Coronal) data. 83

Figure S3.1 Simulated cell type proportions..... 97

Figure S3.2 Simulated data are realistic, preserving data features observed in the published spatial transcriptomics data. 98

Figure S3.3 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario I. 98

Figure S3.4 Comparison of deconvolution accuracy of different methods in simulations under all simulation scenarios. 99

Figure S3.5 Deconvolution accuracy on detecting the dominant cell type at each spatial location for each method at Simulation Scenario I. 100

Figure S3.6 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario II. 101

Figure S3.7 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario II. 102

Figure S3.8 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario IV 103

Figure S3.9 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario IV across all possible combinations of the merged cell types 104

Figure S3.10 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario V. 105

Figure S3.11 Comparison of deconvolution accuracy of different methods at the major cell type level (Scenario I) and at the sub-cell type level. 106

Figure S3.12 Scatterplot of the first principal component of the estimated cell type compositions matrix of mouse olfactory bulb ST data. 107

Figure S3.13 Scatter plot of cell type proportion distributions across spatial locations in the mouse olfactory bulb Spatial Transcriptomics data. 108

Figure S3.14 Accuracy of CARD imputation in the masking analysis in the mouse olfactory bulb data. 108

Figure S3.15 The refined spatial map of cell type composition constructed by CARD. 109

Figure S3.16 The refined spatial map of gene expression constructed by CARD. 110

Figure S3.17 The refined spatial map of gene expression constructed by CARD helps to reveal spatial patterns of genes. 111

Figure S3.18 Clustering results on the original mouse olfactory bulb ST data (n = 282), CARD and BayesSpace imputed data at a higher resolution (n = 2538). 112

Figure S3.19 Scatterplot of the first principal component of the estimated cell type compositions matrix. 112

Figure S3.20 Boxplot of the first principal component score (PC1) of the estimated cell type proportions by CARD and other methods at different regions in the human PDAC dataset respectively. 113

Figure S3.21 Spatial distribution of dominant cell type on each location based on the cell type proportions from each method. 114

Figure S3.22 Comparisons of cell type proportions in cancer region versus non-cancer region. 115

Figure S3.23 Scatter plot of cell type proportion distributions across spatial locations in the human pancreatic ductal adenocarcinomas data. 116

Figure S3.24 CARD generated consistent deconvolution results across different scRNASeq references. 117

Figure S3.25 Accuracy of CARD imputation in the masking analysis in the human pancreatic ductal adenocarcinoma (PDAC) data. 117

Figure S3.26 Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10) when using different scRNA-seq as references. 118

Figure S3.27 The refined spatial map of cell type composition constructed by CARD in the human pancreatic ductal adenocarcinoma (PDAC) tissue. 119

Figure S3.28 The refined spatial map of gene expression constructed by CARD in the human pancreatic ductal adenocarcinoma (PDAC) tissue. 119

Figure S 3.29 The refined spatial map of gene expression constructed by CARD in the human pancreatic ductal adenocarcinoma (PDAC) tissue. 120

Figure S3.30 Clustering results on the original human PDAC ST data (n = 428), CARD and BayesSpace imputed data at a higher resolution (n = 3852). 121

Figure S3.31 Scatterplot of the first principal component of the estimated cell type compositions matrix. 121

Figure S3.32 Scatter plot of cell type proportion distributions across spatial locations in the mouse hippocampus Slide-seq V2 data. 122

Figure S3.33 RCTD, spatialDWLS, stereoscope, and SPOTlight incorrectly locate CA3 cells into CA1 regions more so than CARD. 123

Figure S3.34 Comparisons of the specificity of inferred major regions in the Slide-seq V2 mouse hippocampus data by different deconvolution methods. 124

Figure S3.35 Correlations in cell type proportion across spatial locations between pairs of cell types inferred by CARD in the mouse hippocampus Slide-seq V2 data. 124

Figure S3.36 Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10) in the mouse hippocampus Slide-seq V2 data. 125

Figure S3.37 The refined spatial map of cell type composition constructed by CARD in the mouse hippocampus tissue. 126

Figure S3.38 The refined spatial map of gene expression constructed by CARD in the mouse hippocampus tissue. 127

Figure S3.39 The refined spatial map of gene expression constructed by CARD in the mouse hippocampus tissue. 128

Figure S3.40 Scatter plot of cell type proportion distributions across spatial locations in the mouse hippocampus 10x Visium data when overlayed on top of H&E staining. 129

Figure S3.41 Specificity in the cell type proportion in each region compared with its corresponding boundary for all methods. 130

Figure S3.42 CARDfree generates comparable deconvolution results with CARD. 131

Figure S3.43 Computation time (minutes top panel) and peak memory usage (MB, bottom panel) for each deconvolution method on four real spatial transcriptomics datasets. 132

Figure S3.44 Computation time (seconds) and peak memory usage (MB) for CARD on constructing a refined spatial map. 133

Figure S3.45 Computation time (seconds) and peak memory usage (MB) for CARD on constructing a refined spatial map on different number of new locations (new grid). 133

Figure 4.1 Schematic overview of IRIS. 141

Figure 4.2 Analyzing the human DLPFC 10x Visium data. 145

Figure 4.3 Analyzing the mouse spermatogenesis Slide-seq data. 150

Figure 4.4 Analyzing the mouse olfactory bulb stereo-seq subcellular data. 155

Figure 4.5 Analyzing the human breast cancer 10x Xenium data. 159

Figure S4.1 Evaluation on the spatial domain detection methods in the human dorsolateral prefrontal cortex (DLPFC) 10x Visium data for challenging settings. 176

Figure S4.2 Evaluation on the spatial domain detection methods in the human dorsolateral prefrontal cortex (DLPFC) 10x Visium data for challenging settings. 176

Figure S4.3 Heatmap of expression pattern of the domain specific DE genes. 177

Figure S4.4 Gene set enrichment analysis on the domain-specific DE genes in the human DLPFC data. 178

Figure S4.5 Evaluation on the robustness of different methods in different analysis settings. .. 179

Figure S4.6 The spatial pattern of the domain #1 enriched mitochondrial genes and Sertoli cell marker genes in the WT mouse (WT3_Puck7) of the main analysis. 180

Figure S4.7 Heatmap of expression pattern of the domain specific DE genes. 181

Figure S4.8 GSEA results on each spatial domain detected by IRIS in the mouse WT slice (WT3_Puck7) in the main analysis of the Slide-seq data. 182

Figure S4.9 Comparison of the spatial pattern of ES marker genes in the WT mice (WT3_Puck7) and ob/ob (diabetic, Diabetes2_Puck10) mouse. 182

Figure S4.10 Spatial domains identified by IRIS in the slice S1 in the main analysis of the mouse olfactory bulb stereo-seq data. 183

Figure S4.11 Heatmap of expression pattern of the domain specific DE genes. 184

Figure S4.12 GSEA analysis results of the slice S1 in the main anlaysis of the mouse olfactory bulb data. 185

Figure S4.13 Spatial scatter plot displays the spatial distribution of IRIS estimated cell type proportion across spatial locations in the slice S1 in the main analysis of the mouse olfactory bulb stereo-seq data. 186

Figure S4.14 Spatial domains identified by IRIS in the slice Rep1 in the main analysis of the human breast cancer 10x Xenium data. 187

Figure S4.15 Heatmap of expression pattern of the domain specific DE genes in the human breast cancer 10x Xenium data. 188

Figure S4.16 GSEA analysis results of the slice Rep1 in the human breast cancer 10x Xenium data. 189

Figure S4.17 Spatial scatter plot displays the spatial distribution of IRIS estimated cell type proportion across spatial locations in the slice Rep1 of the human breast cancer 10x Xenium data. 190

Figure S4.18 Breast cancer subtypes classified by the expression of hormonal receptors in each spatial location. 191

Figure S4.19 Computation time (minutes top panel) and (B) peak memory usage (MB, bottom panel) for spatial domain detection method on moderate-sized real spatial transcriptomics datasets. 192

Figure S4.20 Selection of the penalty parameters according to the performance of IRIS in baseline analysis of the human DLPFC dataset. 192

List of Appendices

Appendix A. Chapter 2 (iDEA) Supplementary Text..... 199

Appendix B. Chapter 3 (CARD) Supplementary Text..... 210

Appendix C. Chapter 4 (IRIS) Supplementary Text..... 231

Abstract

Advancements in transcriptomic technologies have enabled the measurement of gene expression at single cell resolution and provided spatial localization information on tissues. The increasing accessibility of these single-cell RNA sequencing (scRNA-seq) or spatially resolved transcriptomic (SRT) datasets provides a comprehensive cell atlas. It enables the thorough characterization of transcriptomic landscapes of tissues for a mechanistic understanding of many biological processes. In the meantime, improvements in transcriptomic technologies have increased both the volume and complexity of data, introducing new computational and statistical challenges for data analysis, including differential expression analysis, gene set enrichment analysis, cell type deconvolution analysis, and spatial domain clustering. In this dissertation, I propose three statistical and computational methods to address these challenges for capturing and dissecting cellular and tissue heterogeneity with high statistical power and accuracy, while providing new insight into biological systems.

In Chapter 2, I develop a method, iDEA, that performs joint DE and GSE analysis in scRNA-seq studies. By integrating DE and GSE analyses, iDEA can improve the power and consistency of DE analysis, produce effective control of type I errors, thus yielding high statistical power and accuracy of GSE analysis. Importantly, iDEA uses only DE summary statistics as input, enabling effective data modeling through complementing and pairing with various existing DE methods. I illustrate the benefits of iDEA with extensive simulations, and three scRNA-seq data sets, where iDEA achieves up to five-fold power gain over existing GSE methods and up to 64% power gain over existing DE methods.

In Chapter 3, I develop a method CARD to perform spatially informed cell type deconvolution for SRT data. CARD builds upon a non-negative matrix factorization (NMF) model that leverages the cell-type-specific gene expression from scRNA-seq data. A unique feature of CARD is its ability to accommodate the spatial correlation structure in cell-type composition across tissue locations by a conditional autoregressive (CAR) modeling assumption. This enables accurate and robust deconvolution of SRT data across technologies and in the presence of mismatched scRNA-seq references. Furthermore, modeling spatial correlation allows CARD to impute cell-type compositions and gene expression levels on new locations of the tissue, facilitating the reconstruction of high-resolution map. Importantly, CARD is computationally scalable and efficient to datasets with tens of thousands of genes measured on tens of thousands of samples. With extensive simulations and comprehensive applications to four real datasets, CARD outperforms other methods, provide novel biological insight underlines the tissue heterogeneity.

In Chapter 4, I develop a method that simultaneously characterize the transcriptomic landscapes on multiple tissues. While SRT datasets can be generated from multiple tissue sections with high resolution, existing methods primarily focus on a single tissue section and fail to utilize information from scRNA-seq datasets for spatial domain detection. Additionally, many published methods lack computational scalability for high-resolution large-scale SRT datasets being collected today. To fill these gaps, I developed IRIS, which leverages cell type specific gene expression information from scRNA-seq to detect spatial domains on multiple tissue sections. By iteratively updating spatial domain labels while considering within-slice and between-slice compositional similarities, IRIS ensures optimal clustering performance. Through in-depth analysis of six spatial transcriptomics datasets, IRIS demonstrates significant advantages,

achieving up to 1083% clustering accuracy improvement over existing methods. This enables the identification of transcriptomic landscapes in complex tissues, including the human prefrontal cortex, spermatogenesis, olfactory bulb, and human breast cancer.

Chapter 1 Introduction

1.1 Background

The transcriptomics technologies have revolutionized our understanding of the complete set of RNA transcripts present within cells, tissues, or organisms (Wang, Gerstein and Snyder 2009). Through the comprehensive analysis of gene expression patterns, these technologies provide crucial insights into regulatory networks, cellular heterogeneity, and disease mechanisms (Angerer et al. 2017, Ramachandran et al. 2020, Fiers et al. 2018, Lowe et al. 2017). In the last decade, transcriptomics technologies have also driven a paradigm shift in genomics, leading to remarkable advancements in our comprehension of biological systems. These technical breakthroughs have significantly expanded our knowledge in areas such as cellular heterogeneity, developmental processes, and disease progression, thereby paving the way for the development of novel therapeutic strategies. Over the years, a number of innovative sequencing technologies have emerged to investigate the transcriptome, with two notable advancements standing out: single-cell RNA sequencing (scRNA-seq) and spatially resolved transcriptomics (SRT). These cutting-edge approaches have propelled the field of transcriptomics further by enabling the study of gene expression at unprecedented levels of resolution and spatial context.

1.2 Single-cell RNA-sequencing

Prior to single-cell RNA sequencing (RNA-seq), bulk RNA-seq served as a popular tool to measure the average gene expression levels in a population of cells or tissues. Nonetheless, this technique lacks the ability to discern variations in gene expression between individual cells,

thereby masking rare cell populations, especially for tumor cells. Tumors exhibit high heterogeneity both between individual tumor cells and within tumor microenvironment. Tumor microenvironment is a complex ecosystem comprising stromal cells, immune cells, and other non-neoplastic components (Baghban et al. 2020). The infiltration of these diverse cell types and the crosstalk among them influence the tumor's behavior and response to therapy (Li and Wang 2021). The averaging effect of bulk RNA-seq can obscure the true signals driving tumorigenesis or therapeutic resistance that may originate from rare cell populations or specific cell types (Li and Wang 2021). Understanding the precise molecular mechanisms driving tumorigenesis and therapeutic resistance requires the ability to capture and dissect the gene expression profiles of individual cells within the heterogeneous tumor ecosystem. By overcoming the limitations of bulk RNA-seq, scRNA-seq technology has revolutionized the field of transcriptomics. It enables gene expression profiling at a single-cell resolution, facilitating the identification of rare cell populations and uncovering previously unseen cellular heterogeneity (Saliba et al. 2014, Zhu, Preissl and Ren 2020, Eberwine et al. 2014). Moreover, scRNA-seq allows for the reconstruction of cellular trajectories during disease progression, unraveling the dynamic changes in gene expression patterns that underlie the development and progression of complex disease (Qiu et al. 2022, Papalexi and Satija 2018, Farrell et al. 2018, Hwang, Lee and Bang 2018).

1.3 Spatially resolved transcriptomics

Despite its ability to characterize cell populations within a tissue, scRNA-seq is not able to capture spatial information due to the tissue dissociation step involved. However, spatial information is important for understanding the tissue organization and interactions between different cell types within a tissue. It provides crucial insights into the spatial arrangement of cells, cell-to-cell communication, and the influence of microenvironments on cellular function. To

overcome this limitation, spatially resolved transcriptomics (SRT) technologies have emerged as a groundbreaking technique that perform gene expression profiling on many tissue locations while preserving spatial localization information (Tian, Chen and Macosko 2022, Rao et al. 2021, Williams et al. 2022, Moses and Pachter 2022, Burgess 2019). Broadly, SRT technologies can be classified into two categories based on how they profile transcriptomes (Williams et al. 2022, Asp, Bergenstråhle and Lundeberg 2020). Firstly, there are imaging based SRT technologies, which utilize microscopy to image mRNAs *in situ*, thereby enabling transcriptome profiling. In imaging based SRT technologies, two widely used methods for distinguishing different mRNA species are *in situ* hybridization (ISH) and *in situ* sequencing (ISS). ISH-based methods, such as MERFISH (Chen et al. 2015, Vizgen 2021), seqFISH (Lubeck et al. 2014), seqFISH+ (Lubeck et al. 2014), 10x Xenium (Janesick et al. 2022), involve the use of fluorescently labeled probes that specifically bind to target mRNAs within the tissue. These methods allow for the visualization and differentiation of specific mRNA species within the sample. On the other hand, ISS-based methods, such as STARmap (Wang et al. 2018a) and FISSEQ (Lee et al. 2015a), use a barcoded nucleotide sequence to record imaging location. This enables the reading of the captured mRNAs' sequences and subsequent identification of different mRNA species. Secondly, sequencing-based SRT technologies, such as Spatial Transcriptomics (ST) (Ståhl et al. 2016), 10x Visium (10XGenomics), Slide-seq (Rodriques et al. 2019), HDST (Vickovic et al. 2019), Slide-seq V2 (Stickels et al. 2021), Seq-Scope (Cho et al. 2021), and Stereo-seq (Chen et al. 2022), are employed to extract mRNAs while preserving their spatial location, followed by mRNA profiling using next-generation sequencing (NGS) techniques. These sequencing-based methods provide a comprehensive transcriptomic profile of the tissue while retaining information about the spatial organization of the mRNAs. Nevertheless, both categories of methods have their unique data

features and limitations. For example, current imaging-based SRT technologies can localize only a limited number of genes, typically ranging from hundreds to thousands, within intact tissue samples (Choe et al. 2023). Conversely, sequencing-based technologies offer the advantage of genome-wide transcript coverage; however, the transcript depth achieved by these methods may not be sufficiently high. Additionally, most of the sequencing-based techniques have a resolution larger than a typical single cell (Williams et al. 2022, Asp et al. 2020). Consequently, these aspects highlight the necessity for sophisticated computational and statistical methods to harness the full potential of large-scale SRT data, enabling the extraction of spatial information to derive novel biological insights.

1.4 Challenges in transcriptomics data analysis

The rapid progress in transcriptomic technologies has broadened our comprehension of gene expression dynamics, but it has also introduced significant challenges in transcriptomics data analysis. These challenges stem from multiple factors, including the increasing volume and complexity of transcriptomics data, which often exhibit high-dimensional features, inherent noise, biases, batch effects, and the inclusion of spatial information. Consequently, it is imperative to develop efficient computational algorithms and robust statistical methods that can effectively handle these complexities and extract meaningful biological insights from the data. Next, I will discuss the major analytical challenges in transcriptomics data analysis.

1.4.1 Differential expression and gene set enrichment analysis

In scRNA-seq studies, differential expression (DE) and gene set enrichment (GSE) analysis are the most common analytical tasks (Lähnemann et al. 2020, Yu et al. 2021b). DE analysis aims to identify genes that are significantly up-regulated or down-regulated between different cell type

populations, experimental conditions, or time points. While bulk RNA-seq methods have been traditionally used for DE analysis, scRNA-seq data possesses unique characteristics, including generally low library sizes, high noise levels, and a large fraction of zeros, so-called "dropout" events. Consequently, several methods specifically tailored for scRNA-seq DE analysis have been developed (Das, Rai and Rai 2022, Soneson and Robinson 2018). These methods can be categorized based on their data input requirements (counts or transformed data), data distribution (parametric or non-parametric), and statistical models (e.g., generalized linear models, generalized additive models, mixture models). However, most existing DE methods analyze genes individually, potentially leading to power loss and inconsistencies among different approaches. Previous studies have highlighted the lack of agreement in identifying DE genes across different methods (Squair et al. 2021, Mou et al. 2020).

GSE analysis is another crucial tool that aggregates gene-level evidence to the gene set level, providing a robust and interpretable biological context for DE results. However, the application of GSE analysis to single-cell data remains challenging, with only a few methods currently available but lack efficiency (Maleki et al. 2020, Noureen et al. 2022, Aibar et al. 2017, Pont, Tosolini and Fournié 2019). Moreover, most GSE methods treat GSE analysis as a separate sequential step following DE analysis, overlooking their statistically interconnected relationship. Hence, there is an urgent need for novel computational methods capable of efficiently detecting biologically relevant gene sets and pathways while considering the relationship between DE and GSE analyses.

1.4.2 Cell type deconvolution

scRNA-seq has revealed the transcriptional heterogeneity of cell types whereas the information on tissue organization is still missing. In contrast, SRT allows for dissecting spatial

heterogeneity of complex tissues, but most sequencing-based SRT (i.e., 10x Visium) do not provide single-cell resolution. Consequently, gene expression measurements on each tissue location represent a mixture of cells belonging to potentially distinct cell types, with the proportion of each being unknown. Therefore, an essential task is to estimate the cell type compositions for each spatial location, commonly referred to as cell type deconvolution. Knowledge of the cell type composition and spatial distribution of diverse cell types is critical for identifying cellular targets of diseases and investigating the tumor microenvironment. In recent years, several deconvolution methods have been proposed (Elosua-Bayes et al. 2021, Song and Su 2021, Lopez et al. 2022, Biancalani et al. 2021, Danaher et al. 2022, Gayoso et al. 2022, Andersson et al. 2020, Kleshchevnikov et al. 2022, Dong and Yuan 2021, Cable et al. 2021, Li et al. 2022, Li et al. 2023), falling into two categories: 1) reference-based deconvolution methods that leverage cell type specific gene signatures from scRNA-seq datasets, and 2) reference-free deconvolution methods that do not rely on predefined scRNA-seq datasets. However, most of current methods have the following limitations: 1) they do not fully utilize the available spatial information 2) their choice of scRNA-seq references is not robust 3) they cannot enhance the resolution of the original dataset at both cell type and gene expression level, and 4) they lack computational scalability for large-scale datasets.

1.4.3 Spatial domain detection

In SRT studies, another critical task is detecting distinct spatial domains on the tissue, which allows for the transcriptomic characterization of tissue structures and microenvironments. While conventional clustering algorithms designed for scRNA-seq primarily rely on gene expression measurements, incorporating spatial information into clustering models becomes essential for spatial domain detection. Several computational and statistical methods (Hu et al.

2021, Zhao et al. 2021, Fu et al. 2021, Dries et al. 2021, Moses and Pachter 2022, Palla et al. 2022, Zhu et al. 2018, Tian et al. 2022, Rao et al. 2021, Li and Zhou 2022, Shang and Zhou 2022) have been developed to address the need, but most of them focused on analyzing single tissue slice. However, with the recent advancements in SRT techniques, it is now possible to generate large-scale datasets from multiple tissue sections at high resolution. As a result, detecting spatial domains efficiently and accurately through integrative analysis of multiple tissue slices has emerged as a new challenge.

1.5 Dissertation outline

In this dissertation, I present a series of statistical and computational methods to tackle the aforementioned challenges in the field. Specifically, in Chapter 2, I focus on scRNA-seq data and develop a statistical method, iDEA, to perform integrative DE and GSE analysis based on summary statistics through a hierarchical Bayesian modeling framework. iDEA models all genes together by borrowing information across genes in terms of DE effect size distributional properties. Moreover, iDEA utilizes summary statistics output from existing DE tools, making it scalable to large-scale scRNA-seq data sets. As evaluated by extensive simulations and comprehensive real data applications, iDEA shows a higher consistency in detected DE genes, produces calibrated type I error control, and a large power gain over existing GSE methods. Notably, iDEA is the only DE method to date that takes summary statistics as input, which can take advantage of flexible data modeling.

In chapter 3, I shift the focus to SRT data and develop a method, CARD for spatially informed cell type deconvolution for spatial transcriptomics. CARD is built on a non-negative matrix factorization (NMF) framework with conditional autoregressive (CAR) modeling assumption, which is widely applied in graph network analysis. By leveraging information from

scRNA-seq studies and information from the spatial correlation structure among tissue locations, CARD performs accurate and robust cell type deconvolution, and can reconstruct a refined spatial map with enhanced resolution. Additionally, I have extended CARD to include a reference-free version that allows flexibility in scRNA-seq references and a single-cell mapping version that constructs single-cell level gene expression from a reference dataset for each spatial location. Through extensive simulations studies and comprehensive analyses on four real datasets, I demonstrate that CARD is more accurate and computationally efficient than other methods and can provide novel biological insights into cell types and molecular markers that define the tissue heterogeneity.

In chapter 4, I focus on SRT data across multiple tissue slices and develop a method, IRIS for reference informed spatial domain detection on multiple tissue sections. IRIS accounts for both within-slice and between-slice compositional similarities using a graph Laplacian matrix and a Euclidean distance matrix to jointly detect spatial domains across multiple tissue sections. Through an iterative optimization framework, IRIS updates the cell type composition matrix and spatial domain labels across tissue slices. I demonstrate the advantages of IRIS through extensive analyses on diverse spatial transcriptomics datasets, including different technologies, species, and tissues. IRIS substantially improves domain detection accuracy when compared to the existing methods, enabling a more accurate depiction of the transcriptomic landscape in complex tissues such as the human prefrontal cortex, mouse spermatogenesis, mouse olfactory bulb, and human breast cancer.

Chapter 2 Integrative Differential Expression and Gene Set Enrichment Analysis Using Summary Statistics for scRNA-seq Studies.

2.1 Abstract

Differential expression (DE) analysis and gene set enrichment (GSE) analysis are commonly applied in single cell RNA sequencing (scRNA-seq) studies. Here, we develop an integrative and scalable computational method, iDEA, to perform joint DE and GSE analysis through a hierarchical Bayesian framework. By integrating DE and GSE analyses, iDEA can improve the power and consistency of DE analysis and the accuracy of GSE analysis. Importantly, iDEA uses only DE summary statistics as input, enabling effective data modeling through complementing and pairing with various existing DE methods. We illustrate the benefits of iDEA with extensive simulations. We also apply iDEA to analyze three scRNA-seq data sets, where iDEA achieves up to five-fold power gain over existing GSE methods and up to 64% power gain over existing DE methods. The power gained by iDEA allows us to identify many pathways that would not be identified by existing approaches in these data.

2.2 Introduction

DE analysis is a routine association analysis task in scRNA-seq studies for identifying genes that are differentially expressed between cell subpopulations, between experimental conditions, or between case control status. Commonly applied DE methods in scRNA-seq include MAST (Finak et al. 2015), SCDE (Kharchenko, Silberstein and Scadden 2014) and zinger (Van den Berge et al. 2018), to name a few. While different DE methods make various modelling

assumptions to capture diverse aspects of scRNA-seq data (Soneson and Robinson 2018), almost all of them analyze one gene at a time. Analyzing one gene at a time can lead to potential power loss, as this approach fails to exploit consistent DE evidence across similar genes that could otherwise be used to enhance DE analysis power. It is plausible that due to low statistical power, different scRNA-seq DE methods would tend to prioritize a different set of DE genes in real data applications, leading to sub-optimal performance and inconsistency of results among different methods. In many other types of association analysis such as genome-wide association studies, it has been well recognized that Bayesian approaches that model multiple predictor variables together, even with the simple composite likelihood strategy where information is borrowed across multiple predictor variables each treated independently, can substantially increase power over univariate approaches (Zhou, Carbonetto and Stephens 2013).

GSE analysis is also a routine task that aims to aggregate gene-level DE evidence to the gene set or pathway level. By aggregating gene-level DE evidence, GSE analysis can facilitate the robust biological interpretation of DE results. Many different GSE analysis approaches have been developed, but almost all of them are developed in the bulk RNA-seq analysis setting (Khatri, Sirota and Butte 2012). These existing GSE approaches include over-representation analysis methods such as DAVID (Huang et al. 2007) and Fisher's exact test (Camp et al. 2015); self-contained test methods such as t-test (Oron, Jiang and Gentleman 2008), Chi-square test (Goeman et al. 2004) and others; and competitive test methods such as PAGE (Kim and Volsky 2005), GSEA (Subramanian et al. 2005) and CAMERA (Wu and Smyth 2012). Despite the abundance of the existing GSE methods, their effectiveness for scRNA-seq analysis remains elusive. Indeed, no comparison studies have been performed thus far to evaluate the effectiveness of the existing GSE methods in the scRNA-seq setting. In addition, and perhaps more importantly, almost all existing

GSE methods treat GSE analysis as a separate analytic step after DE analysis. However, GSE analysis and DE analysis are interconnected with each other statistically: while DE results are certainly indispensable for performing GSE analysis to detect enriched gene sets, detected enriched or unenriched gene sets also contain invaluable information that can serve as feedback into DE analysis to enhance its statistical power. Therefore, integrating DE analysis and GSE analysis has the potential to substantially increase the power of both and ensure result reproducibility for scRNA-seq analysis.

Here, we develop a statistical method, which we refer to as the integrative Differential expression and gene set Enrichment Analysis (iDEA), that addresses the aforementioned shortcomings of previous methods for scRNA-seq data analysis. iDEA models all genes together by borrowing information across genes in terms of DE effect size distributional properties. iDEA also integrates DE analysis and GSE analysis into a joint statistical framework, providing substantial power gains for both analytic tasks. Importantly, iDEA makes use of summary statistics output from existing DE tools and does not make explicit modeling assumptions on the individual-level scRNA-seq data. Use of summary statistics not only allows iDEA to take advantage of various existing DE models for effective and flexible data modeling, but also ensures its scalable computation to large-scale scRNA-seq data sets. In addition, incorporating summary statistics from scRNA-seq DE analysis into GSE analysis under the framework of iDEA makes GSE analysis less susceptible to gene-gene correlations and other technical difficulties such as dropout events. We illustrate the benefits of iDEA with extensive simulations and applications to three scRNA-seq data.

2.3 Results

2.3.1 Methods overview and simulation design

We provided an overview of iDEA in **Methods**, with technical details provided in **Appendix A.1 - A.2** and a method schematic for shown in **Figure 2.1**. Briefly, iDEA requires gene-level summary statistics in terms of fold change/effect size estimates and their standard errors as inputs. The input summary statistics can be obtained using any existing scRNA-seq DE methods. As we will show below, given the input from any DE method, iDEA can often improve its power. Besides DE summary statistics, iDEA also requires pre-compiled gene sets. For human data, we have compiled and pruned a total of 12,033 gene sets from seven existing gene set/pathway databases including GO (Ashburner et al. 2000), KEGG (Kanehisa and Goto 2000), Reactome (Joshi-Tope et al. 2005), BioCarta (Nishimura 2001), PubChem Compound (Bolton et al. 2010), ImmuneSigDB (Godec et al. 2016), and PID (Schaefer et al. 2009). For mouse data, we have compiled and pruned a total of 2,851 gene sets from GO (Ashburner et al. 2000). With these inputs, iDEA examines one gene set at a time, performs inference through an expectation maximization algorithm, and uses Louis method (Louis 1982) to compute a calibrated *p*-value testing whether the gene set is enriched in DE genes or not. In addition, given any gene set, iDEA produces for each gene a posterior probability of DE as its DE evidence. iDEA is implemented as an open-source R package, freely available at www.xzlab.org/software.html.

We performed simulations to evaluate the effectiveness of iDEA for GSE analysis and DE analysis (details in **Methods**). Briefly, we simulated zero-inflated count data for 10,000 genes on 174 cells through a zero-truncated negative binomial distribution using parameters inferred from a real scRNA-seq data. The simulated data shared similar characteristics with the real scRNA-seq data (**Figure S2.1**). Among the simulated genes, a certain percentage of them belong to a gene set and we refer to this percentage as the gene set coverage rate (CR). In the null simulations of GSE analysis, each of the 10,000 genes is randomly assigned to be a DE gene with a probability

$\exp(\tau_0)/(1+\exp(\tau_0))$, where τ_0 determines the baseline probability of a gene being DE. Note that the null simulations of GSE analysis contain DE genes, though these DE genes are not enriched in any gene set. In the alternative simulations of GSE analysis, the *j*-th gene is randomly assigned to be a DE gene with probability $\exp(\tau_0 + a_j\tau_1)/(1+\exp(\tau_0 + a_j\tau_1))$, where a_j is a binary indicator on whether the *j*-th gene belongs to the gene set and τ_1 is the gene set enrichment coefficient that determines whether belonging to the gene set is predictive for the gene being DE. We performed our main simulations in a baseline scenario with $\tau_0 = -2.0$ and CR = 10% and explored different combinations of τ_0, τ_1 and CR to create various simulation scenarios.

2.3.2 Simulation results

For GSE analysis, we compared the performance of iDEA with the commonly used GSE analysis methods fGSEA (Sergushichev 2016), CAMERA (Wu and Smyth 2012), PAGE (Kim and Volsky 2005) and GSEA (Subramanian et al. 2005). We found that iDEA produces well-calibrated *p*-values under the null in different simulation scenarios (**Figure 2.2, Figure S2.2**). The

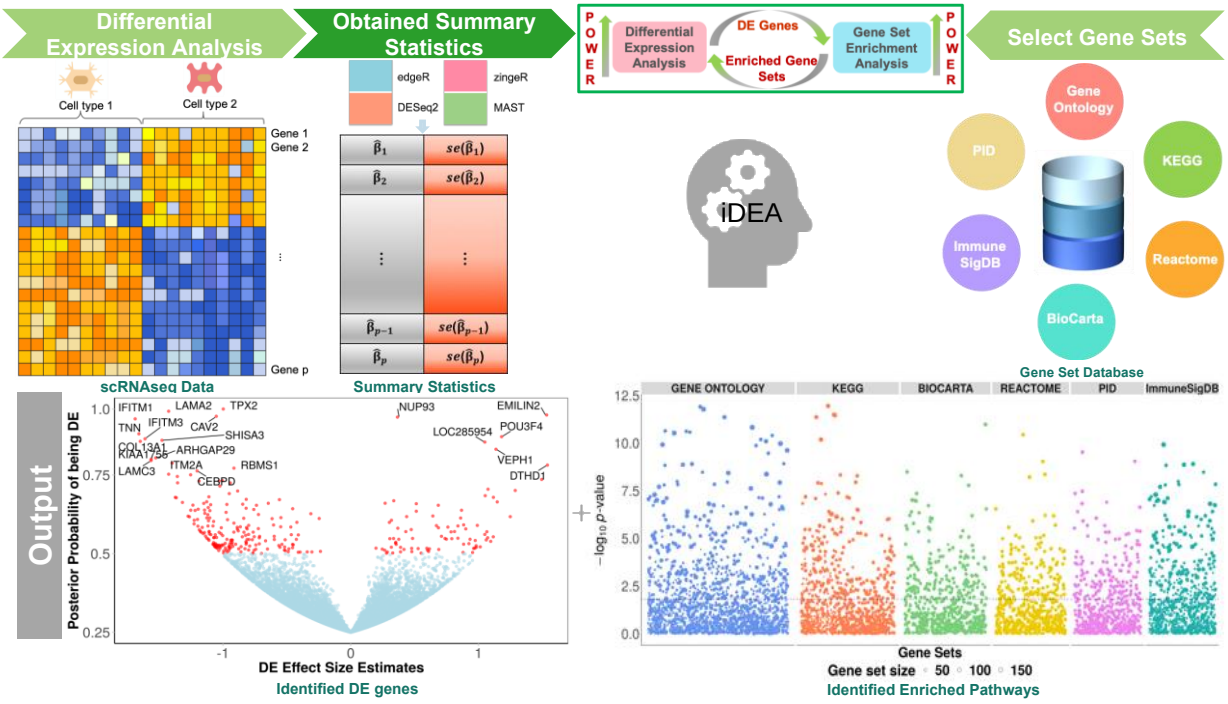


Figure 2.1 Schematic overview of iDEA.

iDEA is designed to jointly model all genes together for integrative differential expression (DE) analysis and gene set enrichment (GSE) analysis. iDEA requires input association summary statistics from existing scRNA-seq DE methods in terms of the DE effect size estimate $\hat{\beta}_j$ and its standard error se ($\hat{\beta}_j$) for every gene ($j = 1,2,\cdots,p$) (top left panels). iDEA also requires a pre-defined set of gene sets that we have compiled and pruned for use with the software (top right panels). With these two inputs, iDEA performs joint DE and GSE analysis through a Bayesian hierarchical model. For each gene set, iDEA outputs a p-value for testing whether the gene set is enriched with DE genes (bottom right panel) for GSE analysis. In addition, iDEA outputs the posterior inclusion probability of each gene being DE (bottom left panel) for DE analysis. By modeling all genes together and integrating DE and GSE analyses in a joint framework, iDEA can increase the power of both analyses.

genomic control factor (λ_{gc}), defined as the ratio between the median empirically observed test statistic and the expected median under the null, is close to one for iDEA across a range of scenarios. Among the other methods, fGSEA, PAGE, and GSEA generally produce calibrated *p*-values (**Figure S2.2**); although occasionally the *p*-values from PAGE may slightly deviate from the diagonal line (e.g., when CR = 10% and τ_0 = -2.0; **Figure 2.2D**). In contrast, the *p*-values from CAMERA are only calibrated when CR is very low (1%) and become increasingly overly conservative with increasingly large CR regardless of the DE gene percentage (e.g., **Figure 2.2B – Figure 2.2D**; the last two columns in **Figure S2.2**). The deflation of CAMERA *p*-values under large CR is presumably because the asymptotic normal approximation used in CAMERA is no longer accurate there. Certainly, we note that under settings with both extremely low τ_0 (e.g. $\tau_0 =$ -3; which corresponds to an average of 4.7% genes being DE) and extremely low CR (e.g. CR = 1%), the distribution of *p*-values from all methods would start to deviate from the expected null (e.g., **Figure S2.2**, $\tau_0 =$ -3, CR = 1%; and to a lesser extent, CR = 5%). Under these extreme parameter combinations, the suboptimal performance of iDEA in terms of type I error control is presumably due to the potential parameter identifiability issue encountered when fitting rare and imbalanced event data (Zhou et al. 2018). The suboptimal performance of the other methods is

presumably because the asymptotic normal approximation for obtaining *p*-values in these methods becomes no longer accurate.

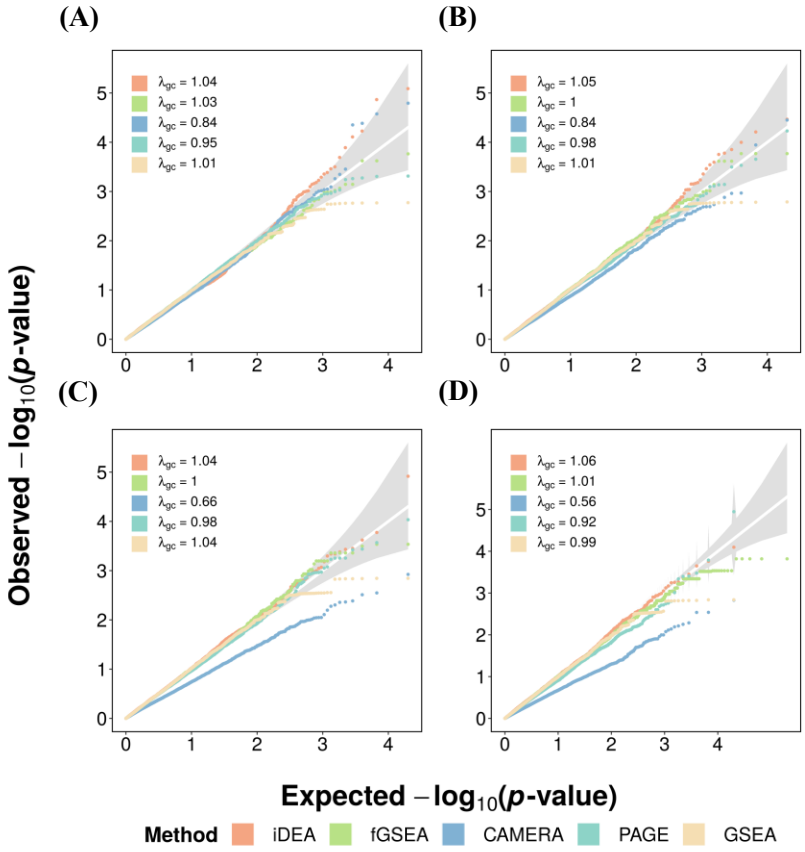


Figure 2.2 iDEA produces well-calibrated p-values for gene set enrichment analysis under null simulations.

Quantile-quantile plots of -log10(p-values) from iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown under different null scenarios with varying gene set coverage rates (CR): **(A)** CR = 1%; **(B)** CR = 2%; **(C)** CR = 5%; **(D)** CR = 10%. CR represents the percentage of genes inside the gene set. Here, the other parameters are set to be $\tau_0 = -2 \; and \; \tau_1 = 0$. λ_{gc} is genomic control factor.

Besides type I error control, we found that iDEA is more powerful than the other GSE methods across a range of alternative scenarios (**Figure 2.3A, Figure 2.3C** and **Figure S2.3**). Because different methods have different type I error control, to allow for fair comparison, we computed power at a fixed false discovery rate (FDR) of 5%. In the baseline parameter setting of $\tau_1 = 0.5$ and CR = 10%, we found that iDEA achieved a power of 98%. In contrast, fGSEA, CAMERA, PAGE and GSEA achieved a power of 26%, 0%, 8% and 26%, respectively (**Figure**

2.3A). The power of iDEA and the other methods all increase with increasing τ_1 as well as with increasing CR (**Figure S2.3**). In addition to the power versus FDR plots, the receiver operating characteristic (ROC) curves, displaying false positive rates (FPR) across a range of true positive rates (TPR), also show that iDEA achieves a higher Area Under the Curve (AUC) for GSE analysis (**Figure S2.4**). The superior performance of iDEA over existing GSE methods presumably is due to the previously known fact that methods using Kolmogorov Smirnov test (e.g. fGSEA, GSEA) are often not powerful in detecting differences between the distribution of DE test statistics in the gene set vs that outside the gene set, while methods using *t*-tests on the DE z-scores (e.g., CAMERA, PAGE) would also fail to detect gene set enrichment as there is no difference in the mean of DE test statistics in the gene set versus that outside the gene set (Kim and Whitt 2015).

For DE analysis, we found that iDEA can improve DE analysis power regardless of whether the summary statistics are from MAST (Finak et al. 2015), edgeR (Van den Berge et al. 2018, Robinson, McCarthy and Smyth 2010) or zinger (Van den Berge et al. 2018, Love, Huber and Anders 2014) (**Figure 2.3B**, **Figure 2.3D**, **Figure S2.5** and **Figure S2.6**). For example, with $\tau_1 = 5$ and CR = 10%, iDEA achieves a power of 81%, 61% and 83% at a true FDR of 5%, when it uses the input summary statistics obtained from zingeR, MAST, and edgeR, respectively. In contrast, the power of these three different DE methods is 65%, 52%, and 67%, respectively (**Figure S2.5** and **Figure S2.6**). The power improvement brought by iDEA is higher in zingeR and MAST than that in edgeR, presumably because the *p*-values from both zingeR and MAST follow approximately a uniform distribution under the null, more so than that from edgeR (**Figure S2.7**). Because the model assumption of iDEA also requires the input *p*-values from the DE methods to be well-behaved, we will mostly report results based on using zingeR as input in the main text. In the analysis, we also found that the power gain brought by iDEA is mostly due to its joint modeling

of DE and GSE analyses, rather than its joint modeling across all genes. Indeed, when the gene set enrichment parameter τ_1 is small, then the power gain brought by iDEA becomes small or negligible (**Figure S2.5** and **Figure S2.6**). Importantly, iDEA produces reasonably calibrated (or slightly conservative) FDR estimates across a range of simulation scenarios (**Figure S2.8**). The ROC curves also yielded consistent results, with iDEA achieving a higher AUC for DE analysis (**Figure S2.4**). Besides direct examination of DE analysis power, we also used the Jaccard index to examine the results consistency of different DE methods. Presumably because of the power gain brought up by iDEA, we found that iDEA can also improve the consistency of DE results in terms of the top DE gene list obtained from different methods (**Figure S2.9**). For example, when $\tau_1 = 5$ and CR = 10%, the Jaccard index for the top 1,500 genes with the strongest DE evidence obtained by each of three DE methods (MAST, edgeR, and zingeR) is 0.59. After applying iDEA to the corresponding summary statistics, the Jaccard index increases to 0.77.

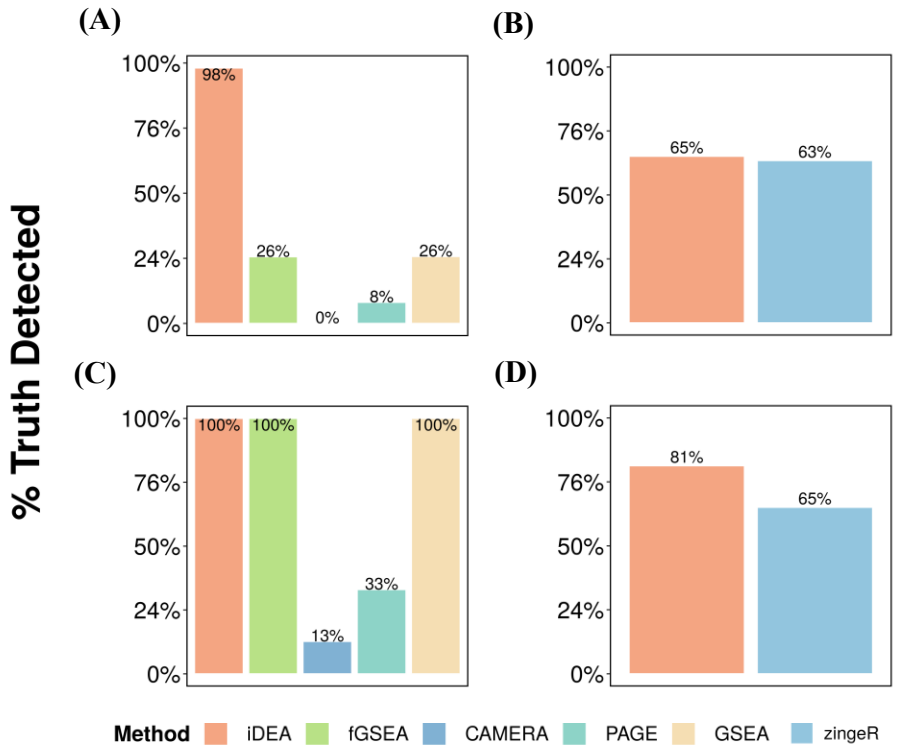


Figure 2.3 iDEA is more powerful for both GSE and DE analyses than existing approaches in power simulations.

The power of iDEA in identifying enriched pathways (y-axis; **A** and **C**) and in identifying differentially expressed genes (y-axis; **B** and **D**) are higher than that of the other methods (x-axis). The compared GSE methods (**A** and **C**) include iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow). The compared DE methods (**B** and **D**) include iDEA (orange) and zingeR (skyblue). Simulations are performed under two parameter settings: $\tau_0 = -2$, $\tau_1 = 0.5$, and CR = 10% (**A** and **B**); $\tau_0 = -2$, $\tau_1 = 5$, and CR = 10% (**C** and **D**). Here, power was calculated based on an FDR of 5%.

2.3.3 Human embryonic stem cell scRNA-seq data

We applied iDEA to analyze three publicly available scRNA-seq data sets. The first scRNA-seq dataset (Chu et al. 2016) consists of gene expression measurements for 15,280 genes on five cell types (details in **Methods**). We carried out both GSE and DE analyses on all ten pairs of the five cell types. Because results are largely consistent across different cell type pairs, we mainly report our analysis here on comparing two cell types, DECs and ECs. For the results comparing the other pairs of the cell types, please refer to (Ma et al. 2020).

We first applied iDEA and other GSE methods to detect significantly enriched gene sets across our compiled database of 11,474 human gene sets (**Figure 2.4A**). We also constructed an empirical null *p*-value distribution by permuting the gene labels for each gene set 10 times. Consistent with simulations, we found that the *p*-values in the permuted data from iDEA ($\lambda_{\text{gc}} = 1.13$), fGSEA ($\lambda_{\text{gc}} = 1.02$), PAGE ($\lambda_{\text{gc}} = 0.96$), and GSEA ($\lambda_{\text{gc}} = 1.01$) are well behaved, while that from CAMERA show severe deflation ($\lambda_{\text{gc}} = 0.29$) (**Figure 2.4B**). For each method, we relied on the empirical null distribution of *p*-values to compute power in detecting enriched gene sets based on a fixed empirical FDR. Consistent with simulations, iDEA identified more significantly enriched gene sets compared to the other GSE methods (**Figure 2.4C**). For example, at an empirical FDR of 5%, iDEA identified 2,106 significantly enriched gene sets, which is 20.9% higher than the next best GSE method (fGSEA, 1,742 significant gene sets). In contrast,

CAMERA, PAGE and GSEA identified 537, 1,328, and 1,079 gene sets, respectively. Besides these GSE methods, iDEA is also more powerful than the hypergeometric test (**Figure S2.10**). Notably, besides the statistical power, many of the top gene sets identified by iDEA are closely related to embryonic development (**Figure 2.4E**). Examples include the *Wnt* signaling pathway, the transforming growth factor beta (TGF-beta) receptor signaling pathway(Gadue et al. 2006), and relevant GO terms such as GO:0048514 (blood vessel morphogenesis), GO:0001944 (vasculature development) and GO:0007492 (endoderm development) (Vokes and Krieg 2002). To quantify the biological significance of gene sets identified by different GSE methods, we quantified the relevance between gene sets and embryonic cell development in an unbiased way by searching the related literatures in PubMed (details in **Methods**). Indeed, in the top 50 enriched gene sets identified by different methods, iDEA identified more gene sets relevant to embryonic cell development (25; **Table S2.1**) than fGSEA (20), CAMERA (23), PAGE (10), and GSEA (12). The higher number of detected enriched gene sets relevant to embryonic cell development by iDEA provides convergent support for the higher power of iDEA for GSE analysis.

We next applied iDEA for DE analysis where we treated the biologically meaningful gene set GO:0001944 (vasculature development) as the annotation. Consistent with simulations, iDEA identified more DE genes than zingeR. For example, at an empirical FDR of 1%, iDEA identified 2,753 DE genes, which is 64.0% higher than zingeR (which identified 1673; **Figure 2.4D**). The 50 selected important DE genes identified by iDEA clearly distinguishes the two examined cell types, DECs and ECs (**Figure 2.4F**). Importantly, based on (Chu et al. 2016), iDEA identified 1,119 genes directly related to definitive endoderm cell differentiation, a process one would expect to be detected by comparing DECs versus ECs, while zingeR only identified 706. The higher number of DE genes relevant to definitive endoderm cell differentiation detected by iDEA

provides convergent support for its higher power for DE analysis. Important DE genes involved in the cell differentiation process that are detected by iDEA but missed by zingeR include *SMAD3* (Teo et al. 2011), *GATA3* (Song et al. 2009), *TGFBR1* (Mullen and Wrana 2017), *WNT7B* (Wang et al. 2018b), *HAND1* (Barnes et al. 2010), *CCND1* (Pauklin et al. 2016) and *HEY2* (Weber et al. 2015). Among them, *SMAD3* is essential for activating the necessary transcriptional network for directing definitive endoderm (DE) formation (Teo et al. 2011); *GATA3* is indispensable for the signaling pathways in large vessel endothelial cells (Song et al. 2009); *TGFBR1* plays an important role in activating *SMAD2* and *SMAD3* (Mullen and Wrana 2017); *WNT7B* is necessary for the redundant ligand–receptor systems which helps activating activate β-catenin signaling in vascular endothelial cells during endoderm development (Wang et al. 2018b). Finally, as in simulations, iDEA improves the consistency of DE analysis results: the Jaccard index for the top DE genes obtained by each of the three DE methods (MAST, edgeR, or zingeR) at an FDR of 1% is only 0.10; after applying iDEA to the corresponding summary statistics, the Jaccard index increases substantially to 0.25 (**Figure S2.11 - Figure S2.12**).

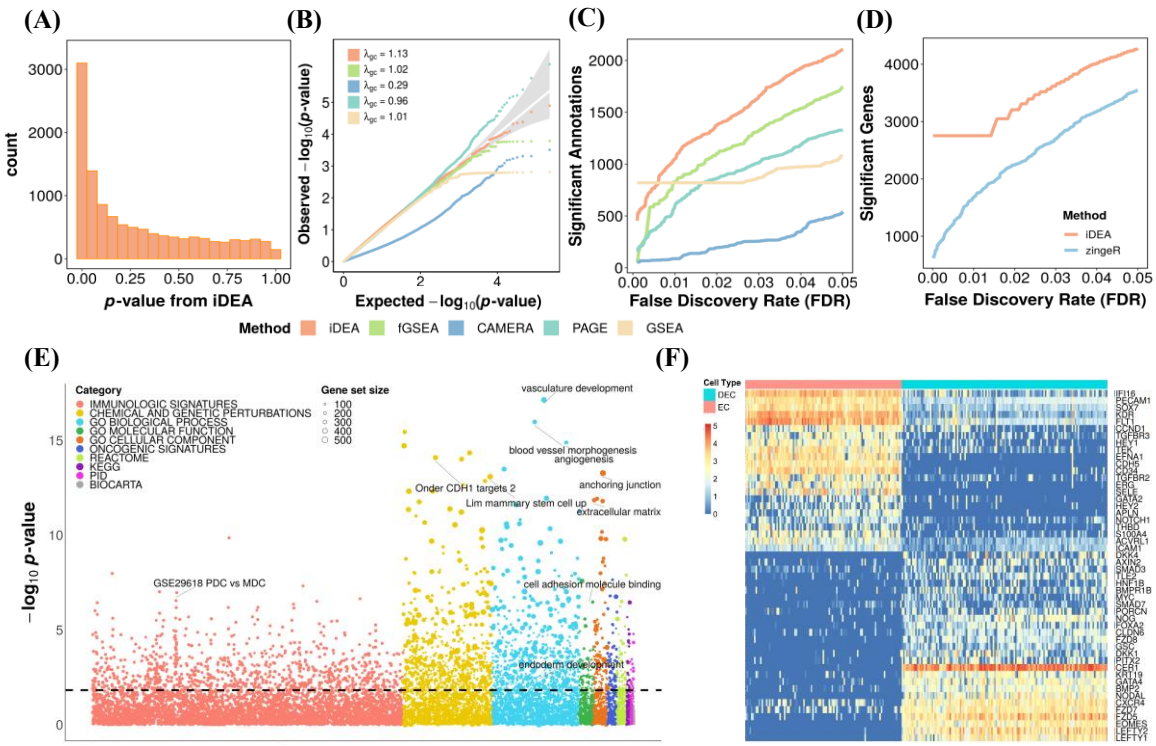


Figure 2.4 Analysis results in the embryonic stem cell scRNA-seq data.

Results are shown for comparing two cell types, endothelial cell (EC) and definitive endoderm derivatives cell (DEC). (**A**) *p*-values from iDEA for GSE analysis display expected enrichment of small *p*-values (for true signals) and a long flat tail towards large *p*-values. (**B**) Quantile-quantile plots of -log10(*p*-values) from GSE methods including iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown under permuted null. The *p*-values from iDEA, fGSEA, PAGE and GSEA are reasonably well calibrated, while that from CAEMRA are overly conservative. Here λ_{gc} is the genomic control factor. (**C**) Number of identified enriched gene sets by iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are plotted against different empirical false discovery rates (FDR). iDEA is more powerful than other methods for GSE analysis. (**D**) Number of identified DE genes by iDEA (orange) and zingeR (blue) are plotted against different empirical FDR values. iDEA is more powerful than zingeR for DE analysis. (**E**) Heatmap shows the normalized expression level (log10-transformation with pseudo-count 0.1) for selected 50 DE genes (rows) identified by iDEA for cells in the two cell types (columns). Genes are sorted by Hierarchical clustering; cells are ordered by cell types (EC: red; DEC: blue). These DE genes clearly distinguish two compared cell types. (**F**) Bubble plot shows –log10 *p*-values for GSE analysis from iDEA (y-axis) for different gene sets. Gene sets are colored by ten categories: immunologic signatures (red), chemical and genetic perturbations (yellow), GO biological process (blue), GO molecular function (green), GO cellular component (orange), oncogenic signatures (deep blue), Reactome (grass-green), KEGG (purple), PID (rose), and Biocarta (grey). The size of the dot represents the number of genes contained in the gene set. Names for ten of the gene sets that are closely related to embryonic cell development are highlighted in the panel.

2.3.4 Mouse sensory neuron scRNA-seq data

The second scRNA-seq data set (Usoskin et al. 2015) consists of 13,598 genes and 622 mouse neuronal cells from eleven different cell types. Following the original paper (Usoskin et al. 2015), we carried our analysis on comparing the nonpeptidergic nociceptor type I (NP1) neurons with each of the other ten cell-types (details in **Methods**). Because the results are again largely consistent across different cell type pairs, we mainly report our analysis here on comparing NP1 versus the remaining ten cell types together. For the corresponding results comparing NP1 versus each of the ten cell types, please refer to ref (Ma et al. 2020).

We first applied iDEA for GSE analysis on a pre-compiled set of 2,851 mouse gene sets (**Figure 2.5A**). Consistent with simulations, the GSE *p*-values in the permuted data from iDEA ($\lambda_{\mathrm{gc}} = 1.07$), fGSEA ($\lambda_{\mathrm{gc}} = 1.08$), PAGE ($\lambda_{\mathrm{gc}} = 0.99$), and GSEA ($\lambda_{\mathrm{gc}} = 0.94$) are all well-behaved, while the *p*-values from CAMERA show severe deflation ($\lambda_{\mathrm{gc}} = 0.11$) (**Figure 2.5B**). Also consistent with simulations, iDEA identified more significantly enriched gene sets compared to the other methods (**Figure 2.5C**). For example, at an FDR of 5%, iDEA identified 1,268 enriched gene sets, which is five times higher than the second-best method (GSEA, 246). In contrast, fGSEA, CAMERA and PAGE identified 236, 134, and 205 enriched gene sets, respectively. Besides these GSE methods, iDEA is also more powerful than the hypergeometric test (**Figure S2.13**). Notably, the significant gene sets identified by iDEA are biologically relevant to the compared cell type pair (**Figure 2.5E** and **Table S2.2**). Most of the top 1% enriched terms were associated with the nervous system, neuronal response, and neuronal functions. Such examples include neuron projection (GO:0043005), neuron part (GO:0097458), and somatodendritic compartment (GO:0036477) (Guo et al. 2019). Other identified gene sets such as axon (GO:0030424) synapse (GO:0045202) and ion transport (GO:0006811) (Hubel 1985) also play important roles in neuronal functions and activities. None of these gene sets were detected by

fGSEA and CAMERA. PAGE and GSEA can also detect these gene sets but do not rank them highly: the rank of these gene sets ranges from top 5% to 61% by PAGE and from top 15% to 71% by GSEA. In addition, use of iDEA recovered 102 out of the 237 gene sets known to be involved in inflammatory itch (Usoskin et al. 2015). In contrast, fGSEA, CAMERA, PAGE, and fGSEA identified 31, 20, 19, and 29 gene sets among them, respectively.

We next applied iDEA for DE analysis where we treated the gene set GO:0097458 (neuron part) as the annotation. Again, iDEA identified more DE genes than zingeR (**Figure 2.5D**). At an FDR of 1%, iDEA detected 1,103 DE genes, which is 11.0% higher than zingeR (993). We illustrate 50 selected DE genes identified by iDEA in **Figure 2.5F**, which clearly distinguish the two compared cell types. Many NP1 neuron marker genes are identified by iDEA but missed by zingeR even at an FDR of 5%. These marker gene examples include *MRGPRB5*, *STX1B*, *FAM167A*, *KLK8*, and *STK32A*. Among these genes, *MRGPRB5* is a Mas-related gene expressed in primary nociceptive sensory neurons (Zylka et al. 2003). *KLK8* mediates signals in the PAR1-dependent signaling responses in the nociceptive neurons (Oikonomopoulou, Diamandis and Hollenberg 2010). Importantly, iDEA detected 79 DE genes out of top 100 previously known NP1 DE genes listed in the original study, while zingeR detected 75 DE genes, again supporting the high power of iDEA. Finally, consistent with simulations, iDEA also improves the consistency of DE analysis results; namely, the Jaccard index for the top DE genes obtained by each of the three DE methods (MAST, edgeR, or zingeR) at an FDR of 1% is 0.14; after applying iDEA to the corresponding summary statistics, the Jaccard index increases to 0.17 (**Figure S2.14 - Figure S2.15**).

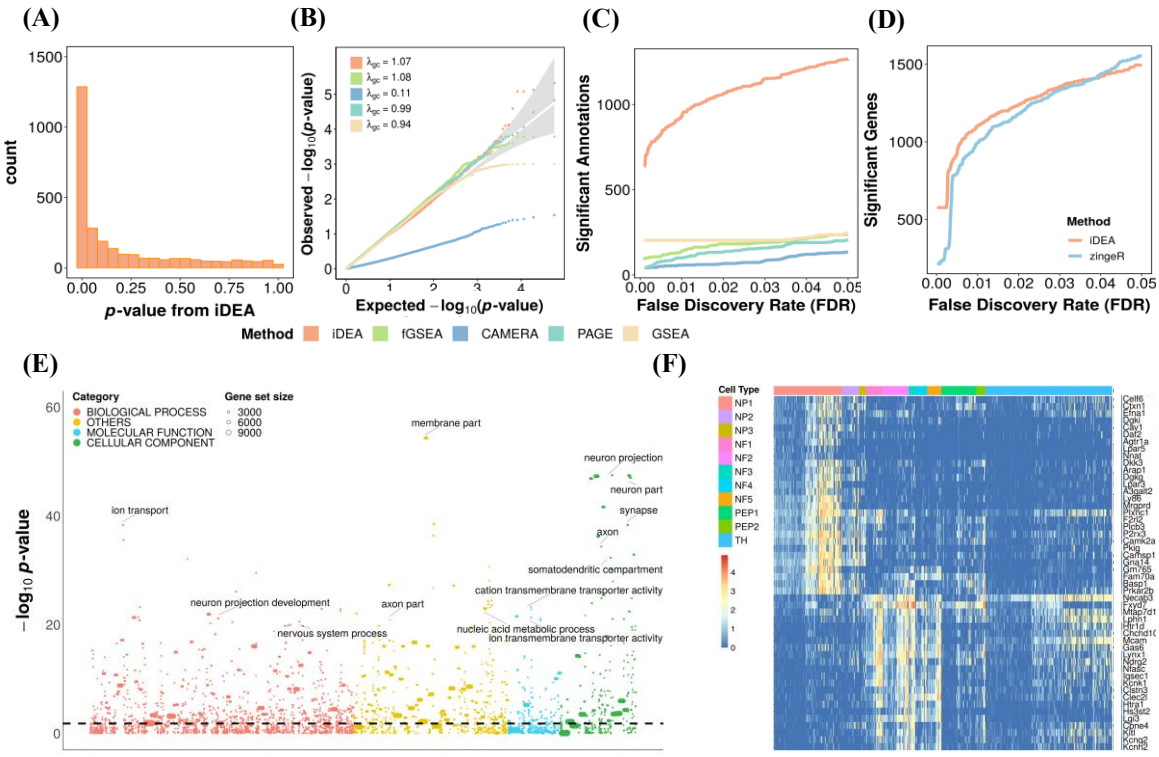


Figure 2.5 Analysis results in the mouse neuronal cell scRNA-seq data.

Results are shown for comparing nonpeptidergic nociceptors 1 (NP1) versus all the other cell types. (**A**) *p*-values from iDEA for GSE analysis display expected enrichment of small *p*-values (for true signals) and a long flat tail towards large *p*-values. (**B**) Quantile-quantile plots of -log10(*p*-values) from GSE methods including iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown under permuted null. The *p*-values from iDEA, fGSEA, PAGE and GSEA are reasonably well calibrated, while those from CAEMRA are overly conservative. Here λ_{gc} is the genomic control factor. (**C**) Number of identified enriched gene sets by iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are plotted against different empirical false discovery rates (FDR). iDEA is more powerful than other methods for GSE analysis. (**D**) Number of identified DE genes by iDEA (orange) and zingeR (blue) are plotted against different empirical FDR values. iDEA is more powerful than zingeR for DE analysis. (**E**) Heatmap shows the normalized expression level (log10-transformation with pseudo-count 0.1) for selected 50 DE genes (rows) identified by iDEA for cells in the two cell types (columns). Genes are sorted by Hierarchical clustering; cells are ordered by cell types (NP1: blue; Others: red). These DE genes clearly distinguish two compared cell types. (**F**) Bubble plot shows –log10 *p*-values for GSE analysis from iDEA (y-axis) for different gene sets. Gene sets are colored by four categories: GO biological process (orange), GO molecular function (blue), GO cellular component (green) and other gene ontology terms with only GO numbers (yellow). The size of the dot represents the number of genes contained in the gene set. Names for ten of the gene sets that are closely related to nociceptive sensory neurons' activities are highlighted in the panel.

2.3.5 10x Genomics PBMC scRNA-seq data

The third scRNA-seq data set consists of 13,713 genes and 2,638 cells collected from peripheral blood mononuclear cells (PBMCs) (Zheng et al. 2017b). We focused on comparing CD4+ T-cells with CD8+ T-cells to examine the performance of various methods in the challenging setting where the two examined cell types are similar (**Figure S2.16**). We also focused on examining a small set of 144 gene sets that contain important gene signatures of immune and stroma cell types (Aran, Hu and Butte 2017). These gene sets contain CD4+ and CD8+ cell type signatures and thus can be treated as true positives for method comparison in this data.

We first applied iDEA to identify enriched gene signatures among these true positives (**Figure 2.6A**). Due to the small number of gene sets examined here, the *p*-values from all methods in the permuted data are not discernable from the null expectation (**Figure 2.6B**). Likely due to the low read depth in 10x genomics data and the subsequent high gene expression measurement noise, all GSE methods have similar power in terms of detecting enriched gene sets based on a fixed FDR threshold (**Figure 2.6C**). However, almost all top enriched gene sets identified by iDEA are relevant to CD4 or CD8 cell functions (**Figure 2.6E**). For example, in the top 25 gene sets identified by iDEA, 22 of them are relevant to CD4 or CD8 cells (**Table S2.3**). In contrast, 13 from fGSEA, 13 from CAMERA, 14 from PAGE, and 13 from GSEA are relevant to CD4 or CD8 cells (**Table S2.4**). Besides these commonly used GSE methods, iDEA is also more powerful than the hypergeometric test, which only identified one significant gene set (**Figure S2.17**).

We next applied iDEA to perform DE analysis where we treated the gene set CD8+ T-effector memory as the annotation. Consistent with simulations, iDEA identified more DE genes than zingeR (**Figure 2.6D**). At an FDR of 1%, iDEA detected 255 significant DE genes, which is 15.3% higher than that detected by zingeR (221). We illustrate 30 selected DE genes identified by iDEA (**Figure 2.6F**), which clearly distinguish the two cell types. The significant DE genes

identified by iDEA include *CD8A, KLRG1*, *GNLY*, and *PRF1* that are all relevant to CD T cell differentiation (Palmer et al. 2006). Indeed, iDEA identified many T cell activation and differentiation related genes that are missed by zingeR. Among the genes missed by zingeR, *BTG2* is important for T-cell activation marker expression, T cell proliferation and migration (Terra et al. 2008); *KLF2* is involved in both the activation of CD4+ T cell trafficking (through regulation of *S1PR1*) and T helper cells differentiation (Lee et al. 2015b); *CD247* of the Ctex region is essential for the TCR-mediated activation of T cells (Lundholm et al. 2010); and *LSP1* is found to be down regulated in human T-cell lines and plays an important role in the process of T-cell transformation (Huang et al. 1997). iDEA also improved the consistency of DE analysis results. Specifically, the Jaccard index for the top DE genes obtained by each of three DE methods (MAST, edgeR or zingeR) at an FDR of 1% was only 0.06; after applying iDEA to the corresponding summary statistics, the Jaccard index increased substantially to 0.15 (**Figure S2.18 - Figure S2.19**).

Finally, while the DE analysis relies on a pre-selected gene set, we found that the number of DE genes identified by iDEA without the pre-selected gene set (=252, at an FDR of 1%) is similar to the results using the cell type defined gene set (=255), both are larger than that identified by zingeR (=221) (**Figure S2.20**). In the three real data applications, we also found that the results from iDEA are largely insensitive to the choice of hyperparameters in the prior distribution for the variance parameters (analysis details in **Methods; Figure S2.21**). Performing analysis on two groups of cells randomly selected form the same cell type also demonstrates the proper type I error control by iDEA (**Figure S2.22**).

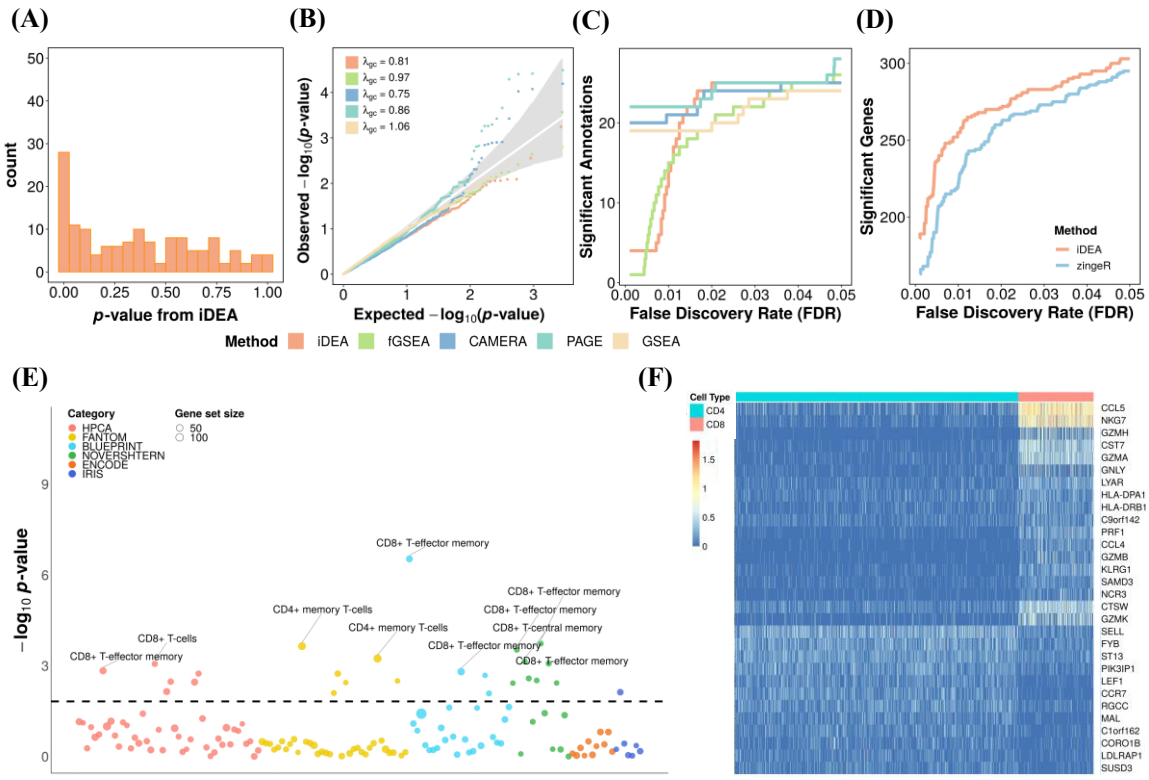


Figure 2.6 Analysis results in the 10X Genomics scRNA-seq data.

Results are shown for comparing CD4+ T cells versus CD8+ T cells. (**A**) *p*-values from iDEA for GSE analysis display expected enrichment of small *p*-values (for true signals) and a long flat tail towards large *p*-values. (**B**) Quantile-quantile plots of -log10(*p*-values) from GSE methods including iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown under permuted null. The *p*-values from all methods in the permuted data are not discernable from the null expectation. Here λ_{gc} is the genomic control factor. (**C**) Number of identified enriched gene sets by iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are plotted against different empirical false discovery rates (FDR). iDEA is as the same powerful than other methods for GSE analysis. (**D**) Number of identified DE genes by iDEA (orange) and zingeR (blue) are plotted against different empirical FDR values. iDEA is more powerful than zingeR for DE analysis. (**E**) Heatmap shows the normalized expression level (log10-transformation with pseudo-count 0.1) for selected 30 DE genes (rows) identified by iDEA for cells in the two cell types (columns). Genes are sorted by Hierarchical clustering; cells are ordered by cell types (NP1: blue; Others: red). These DE genes clearly distinguish two compared cell types. (**F**) Bubble plot shows –log10 *p*-values for GSE analysis from iDEA (y-axis) for different gene sets. Gene sets are colored by six projects: FANTOM (red), HPCA (yellow), BLUEPRINT (blue), ENCODE (green), NOVERSHTERN (orange), IRIS (deep blue). The size of the dot represents the number of genes contained in the gene set. Names for ten of the gene sets that are closely related to CD4+ and CD8+ immune process are highlighted in the panel.

2.4 Discussion

We have presented a computational method, iDEA, for integrating DE analysis and GSE analysis in scRNA-seq studies. iDEA directly models summary statistics from existing scRNA-seq DE tools, produces well-calibrated *p*-values for enriched gene set detection, and provides increased power for both DE and GSE analyses. Modeling summary statistics in iDEA circumvents the need for explicit modeling of individual-level scRNA-seq data, allowing iDEA to be paired with existing DE tools for quick adaptation across a range of scRNA-seq data types. We have demonstrated the benefits of iDEA using both simulations and applications to three recently published scRNA-seq data sets.

We have primarily focused on scRNA-seq data, as we aimed to perform a comprehensive comparative study on GSE methods for scRNA-seq studies in addition to developing iDEA. However, the flexible modeling framework of iDEA can be equally applied to bulk RNA-seq studies. To illustrate this, we applied iDEA to an oral carcinoma bulk RNA-seq dataset (Tuch et al. 2010), where we show that iDEA can identify more DE genes and more enriched gene sets that are relevant to oral carcinogenesis (**Figure S2.23**; **APPENDIX A.3**).

We have primarily focused on modeling the marginal effect size estimates and standard errors from DE analysis, which is equivalent to modeling of marginal z-scores. Our modeling of z-scores follows that of (Efron 2001) (Efron and Tibshirani 2002) and effectively assumes that the prior distribution of true effect sizes is dependent on the standard errors, and subsequently the sample size. Such prior dependence on sample size appears to have relatively mild consequences in practical data analysis and has attractive theoretical properties (Narisetty and He 2014). Nevertheless, we have developed a variant of iDEA that does not require prior dependence on

sample size. The iDEA variant has similar performance as the original iDEA in the real data applications, properly controlling for type I error and displaying higher power than the other methods. For more details of the iDEA variant, please refer to the ref (Ma et al. 2020)

The DE analyses in our real data applications are performed by treating a pre-selected gene set as annotation based on prior biological knowledge. Certainly, selecting such gene set may not always be possible in every study. In the absence of a pre-selected gene set to serve as the annotation, we developed a Bayesian model averaging (BMA) approach (**APPENDIX A.4**) to aggregate DE evidence across all available gene sets. The BMA approach yields consistent results for majority of genes as compared to the pre-selection approach in the real data applications (**Figure S2.24**), demonstrating its utility for practical applications.

iDEA does not explicitly account for gene set overlap that may cause non-independence among gene sets. In practice, we found that the gene set overlap is generally small: the median number of overlapped genes among pairs of gene sets in the human data is only 1 (5 in the mouse data), as compared to the median gene set size of 143 (131 in the mouse data). A careful examination of the top identified enriched gene sets in the real data applications also suggest that gene set overlap does not appear to introduce excessive false signals (**Table S2.5 - Table S2.6; APPENDIX A.5**). In addition, the sparse data structure in scRNA-seq appears to further diminish the concern on gene-gene correlations. Indeed, GSE methods that do not explicitly account for gene-gene correlation (e.g., iDEA, PAGE, fGSEA and GSEA) appear to provide more calibrated *p*-values than methods that explicitly account for gene-gene correlation (e.g., CAMERA) in these real data applications. Nevertheless, we followed most existing GSE approaches and accounted for GSE test non-independence due to gene set overlap through permutation of gene labels. Such permutation retains the gene set overlap proportion under the empirical null: if one gene set

contains genes that are overlapped with genes in another gene set in the real data, then the overlapped number remains the same in the permuted data. Consequently, the test statistics on the two gene sets would be correlated in a somewhat comparable fashion between the permuted data and the real data. By estimating FDR based on such permuted null, we can account for test non-independence due to gene set overlaps.

Finally, we acknowledge that general caveat exists for DE analysis between cell types in scRNA-seq studies: because cell types are often inferred based on the whole gene expression matrix, DE analysis performed on the inferred cell types may lead to inflated DE test statistics with artificially smaller standard errors (Zhang, Kamath and Tse 2019). We have attempted to alleviate such issue by conducting our analysis on datasets where cell types are reasonably rigorously defined and validated through other experiments (**APPENDIX A.6**). Nevertheless, future methodological innovations are needed to account for the uncertainty associated with cell type inference for DE analysis between cell types in scRNA-seq studies.

2.5 Methods

2.5.1 iDEA overview

Here, we provide a brief overview of iDEA, with technical details provided in **APPENDIX A.1-A.2**. iDEA models all genes jointly and requires summary statistics from standard DE analysis for all genes. These summary statistics are in the form of marginal DE effect size estimate $\hat{\beta}_j$ and its standard error $\mathrm{se}(\hat{\beta}_j)$, $j = 1,2,\cdots p$, where p is the number of genes. We assume that the estimated DE effect size centers around the true effect size $\hat{\beta}_j \sim N(\beta_j, \mathrm{se}(\hat{\beta}_j)^2)$, and that the true effect size β_j follows a mixture of two distributions depending on whether j-th gene is a DE gene or not:

$$\beta_j \sim \pi_j \, N\left(0, \mathrm{se}\left(\hat{\beta}_j\right)^2 \sigma_\beta^2\right) + \left(1 - \pi_j\right)\delta_0, \tag{2.1}$$

where π_j is the prior probability of being a DE gene; σ_β^2 is a scaling factor that determines the DE effect size strength; and δ_0 is the Dirac function that represents a point mass at zero. Therefore, with proportion π_j, *j*-th gene is a DE gene and its DE effect size β_j follows a normal distribution with a large variance $\mathrm{se}\left(\hat{\beta}_j\right)^2 \sigma_\beta^2$. With proportion $1 - \pi_j$, *j*-th gene is a non-DE gene and its DE effect size is exactly zero. Note that our modeling above is also equivalent to modeling using marginal z-statistics,

$$z_j \sim \pi_j \, N\left(0, \sigma_\beta^2 + 1\right) + \left(1 - \pi_j\right) N(0,1), \tag{2.2}$$

where z_j is the marginal z-score on the DE evidence for the *j*-th gene.

In Equation (2.1), we have scaled the variance with respect to $\mathrm{se}\left(\hat{\beta}_j\right)^2$ using the scaling factor σ_β^2, so that our analysis results are scale invariant; that is, the results remain the same regardless of what the DE effect size is measured on. For the scaling factor, we followed existing statistical literature and chose the conjugate distribution for a variance parameter as the prior for σ_β^2. Specifically, we specify an inverse gamma prior on σ_β^2 : $\sigma_\beta^2 \sim \mathrm{InvG}(a_\beta, b_\beta)$ with $a_\beta = 3.0, b_\beta = 20.0$, which ensures a prior mean of 10 ($= b_\beta/(a_\beta - 1)$) and the existence of a prior variance (which requires $a_\beta > 2$). To integrate the gene set information into the above model, we model the gene-specific probability of being a DE gene as

$$\mathrm{logit}\left(\pi_j\right) = \log\left(\frac{\pi_j}{1-\pi_j}\right) = \tau_0 + a_j \tau_1. \tag{2.3}$$

where τ_0 is an intercept that determines the proportion of DE genes outside the gene set; a_j is a binary indicator on whether *j*-th gene belongs to the gene set ($a_j = 1$) or not ($a_j = 0$); and τ_1 is a

gene set enrichment parameter that determines the odds ratio of DE for genes inside the gene set versus genes outside the gene set. To facilitate computation, we introduce a vector of binary indicators $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^T$ to indicate whether each gene is a DE gene ($\gamma_j = 1$) or not ($\gamma_j = 0$). Therefore, the prior distribution of γ_j is effectively a Bernoulli distribution,

$$\gamma_j \sim \text{Bern}(\pi_j). \tag{2.4}$$

With proportion π_j, *j*-th gene is a DE gene and with proportion $1 - \pi_j$, *j*-th gene is a non-DE gene and its DE effect size is exactly zero. With the above model setup, we are primarily interested in inferring two parameters: the gene-specific indicator γ_j, which indicates whether *j*-th gene is a DE gene or not; and the enrichment parameter τ_1, which represents the enrichment of DE genes in the gene set. We aim to infer the posterior probability of $\gamma_j = 1$ as evidence for *j*-th gene being DE and test the null hypothesis H_0: $\tau_1 = 0$ that DE genes are not enriched in the gene set.

To achieve both goals, we develop an expectation maximization (EM)-Markov chain Monte Carlo (MCMC) algorithm for parameter estimation. Briefly, we treat the vector of both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as missing data and develop an iterative EM optimization algorithm that alternates between an expectation step and a maximization step. In the expectation step, the expectation of the log likelihood effectively requires computing the posterior probability of each gene being a DE gene, $P(\gamma_j = 1|\text{data})$, through MCMC. In the maximization step, we estimate the enrichment parameter through optimization, which is effectively equivalent to fitting a logistic regression model, where we treat the posterior probabilities for each gene being DE obtained from the expectation step as the outcome variable. While our EM-MCMC algorithm yields accurate parameter estimates, we found that the standard errors for the enrichment parameter obtained through the complete-data log-likelihood function is overly liberal and leads to *p*-value inflations ($\lambda_{\text{gc}} = 1.33$; **Figure**

S2.25A), a phenomenon that has been observed in many other settings (Spall 2005, Louis 1982). Therefore, we used the Louis method (Louis 1982) to obtain the corrected information matrix and produce calibrated *p*-values ($\lambda_{\mathrm{gc}} = 1.06$; **Figure S2.25B**).

2.5.2 Summary statistics and gene annotations

iDEA requires DE summary statistics in the form of fold change/effect size estimates and their standard errors as input. These summary statistics in principal can be obtained using any existing scRNA-seq DE methods, such as MAST (Finak et al. 2015) or zinger (Van den Berge et al. 2018) etc. Here, we primarily focus on presenting the results obtained based on input from zingeR, which directly outputs DE effect size estimates and their standard errors and which is the most recent DE method for scRNA-seq analysis. However, we also explored the benefits of pairing iDEA with different scRNA-seq DE methods in part of the **Results** section. Details of these DE methods are provided in the next subsection.

In addition to DE summary statistics, iDEA also requires pre-defined gene sets. For human data, we downloaded a total of 12,033 gene sets based on seven existing gene set/pathway databases annotated on the reference genome GRCh37 from MSigDB databases (http://software.broadinstitute.org/gsea/downloads.jsp). These databases include BioCarta (Nishimura 2001), KEGG (Kanehisa and Goto 2000), GO (Ashburner et al. 2000), PubChem Compound (Bolton et al. 2010), ImmuneSigDB (Godec et al. 2016), PID (Schaefer et al. 2009) and Reactome (Joshi-Tope et al. 2005). We divided the compiled gene sets into ten functional categories that include immunologic signatures (4,856 gene sets), chemical and genetic perturbations (2,379 gene sets), GO biological process (2,835 gene sets), GO molecular function (544 gene sets), GO cellular component (355 gene sets), oncogenic signatures (186 gene sets), Reactome (415 gene sets), KEGG (163 gene sets), PID (207 gene sets), and Biocarta (93 gene sets).

We merged the gene sets with summary statistics and filtered out gene sets that contain less than 20 genes and finally focused on a total of 11,474 gene sets in GSE analysis. For mouse data, we downloaded the gene ontology (GO) annotations of mouse genes in the GAF 2.0 format from the website (http://www.informatics.jax.org/downloads/reports/index.html#go). We merged the gene sets with summary statistics and filtered out gene sets that contain less than 50 genes and finally focused on a total of 2851 gene sets in GSE analysis. These GO terms were based on four categories: biological process (1,719 gene sets), cellular component (279 gene sets), molecular function (297 gene sets), and unannotated gene sets (556 gene sets). For 10x Genomics data, we collected a total of 489 gene sets consisting of cell type specific gene signatures from Xcell and these gene signatures were previously collected to annotate 64 distinct cell types and cell subsets (Aran et al. 2017). We merged the gene sets with summary statistics and filtered out gene sets that contain less than 10 genes to focus on a final set of 144 cell type specific gene sets. We note that we filtered out gene sets with a small number of genes (e.g., <10 or <20) as the pruning step due to computational reasons: for iDEA and other GSE methods, the gene set enrichment parameter estimation can become inaccurate and unstable for gene sets with small sizes.

2.5.3 Compared methods

For DE analysis in both simulations and real data applications, we compared iDEA with three existing approaches: (1) MAST (version 1.8.1), which outputs a coefficient (coef) as the effect size estimate and a corresponding *p*-value for each gene (Finak et al. 2015); (2) edgeR (version 3.8), which outputs a log fold change (logFC) as the effect size estimate and a corresponding *p*-value for each gene (Robinson et al. 2010). The edgeR function we used was the weighted version of edgeR: it first calculated the cell-level weights using ZINB-WaVE and then used these weights inside edgeR for final computation; (3) zingeR (version 1.0) (Van den Berge

et al. 2018), where we applied zingeR to obtain cell-level weights, which were further supplied to DESeq2 (version 1.18.1) for DE analysis. The output from this procedure consists of a log fold change (log2FoldChange) as the effect size estimate and a corresponding standard derivation (lfcSE) for every gene (Love et al. 2014). iDEA can be paired with each of these DE methods to use the corresponding summary statistics as input for analysis. In these DE methods, in order to extract summary statistics for iDEA, we treated either the logarithm fold change or fold change as the gene effect size $\hat{\beta}_j$, and back derived the standard error of $\hat{\beta}_j$ using the unsigned z-score by $\mathrm{se}(\hat{\beta}_j) = |\hat{\beta}_j/\mathrm{zscore}|$, where zscore was either directly available or was obtained by transforming the *p*-value via the R function qnorm (*p*-value*/2.0,* lower.tail=F). Afterwards, we used summary statistics obtained from these DE methods to fit iDEA.

For GSE analysis in both simulations and real data applications, we mainly compared iDEA with four existing approaches: (1) fGSEA (R version 1.8.0) (Sergushichev 2016); (2) CAMERA (inside limma, R version 3.8.3) (Wu and Smyth 2012); (3) PAGE (PGSEA, R version 1.56.0) (Kim and Volsky 2005); and (4) GSEA (Java version 2.2.4) (Subramanian et al. 2005). We used z-score statistics from zingeR DE analysis as input for all these methods. Here, the z-score statistics were calculated by the transformation of the unadjusted *p*-values, paired with the sign of log-fold change estimate: $\mathrm{zscore} = \Phi^{-1}(1 - \frac{\mathrm{pvalue}}{2})\mathrm{sign}(\mathrm{logFC})$, where $\Phi(\cdot)$ denotes the standard Gaussion cumulative distribution. We used the default settings for all GSE methods. We used the recommended interGeneCorrelation function in CAMERA to calculate the correlation between genes. In addition, we compared iDEA with the hypergeometric test in all real data applications. We counted the number of DE genes (defined as *p*-value < 0.05) and non-DE genes in the gene set as well as outside the gene set and performed hypergeometric test to obtain GSE *p*-value.

Note that, in the GSE analysis, we have primarily focused on comparing our method with traditional GSE methods that aim to identify gene sets whose genes are differentially expressed between cell types or treatment conditions. Different from these traditional GSE methods, several methods have been recently developed for scRNA-seq studies that are targeted for a completely different enrichment task: identifying gene sets whose genes show coordinated transcriptional heterogeneity. Exemplary such methods include the pathway and gene set overdispersion analysis (PAGODA) (Fan et al. 2016) and f-scLVM (Buettner et al. 2017). Because both PAGADA and f-scLVM are targeted for detecting coordinated expression heterogeneity (i.e., gene-gene correlation within a gene set) rather than the usual GSE analysis based on DE analysis, we did not compare our method and other GSE methods with them.

2.5.4 Simulations

We performed simulations to evaluate the performance of iDEA for both DE analysis and GSE analysis. In each simulation replicate, we simulated 10,000 genes. We randomly assigned a proportion of these genes to belong to a gene set of interest. We referred to the percentage of genes belonging to the gene set as the coverage rate (CR), which were set to be either 1%, 2%, 5%, or 10%; where 10% is close to the median CR of all analyzed pathways in the present study. We further introduced a binary indicator a_j to represent whether j-th gene belongs to the gene set ($a_j = 1$) or not ($a_j = 0$). Afterwards, we randomly assigned each gene to be a DE gene with probability π_j, which depends on a_j. In particular, the parameter π_j is in the form of

$$\pi_j = \frac{\exp(\tau_0 + a_j\tau_1)}{1+\exp(\tau_0 + a_j\tau_1)}, \tag{2.5}$$

where the intercept parameters τ_0 was set to be either -0.5, -1.0, -2.0, or -3.0 to present different proportions of DE genes in the data (e.g. $\tau_0 = -2$ represents that roughly 12% of genes are DE

genes; $\frac{\exp(-2)}{1+\exp(-2)} \approx 0.12$); while the gene set enrichment coefficient τ_1 was set to be either 0 (no enrichment of DE genes in the gene set), 0.25 (weak enrichment), 0.5, 1.0 (moderate enrichment), or 5.0 (strong enrichment). Note that the median gene set enrichment parameter estimate across all analyzed pathways in the real data applications is close to 0.5 while the highest enrichment parameter estimate is 17.

In order to compare the performance of iDEA of DE analysis with other count-based DE methods, we simulated scRNA-seq gene expression counts first. To make our simulations as realistic as possible, the simulations were performed based on parameters inferred from a published scRNA-seq data (Chu et al. 2016). Specifically, to simulate scRNA-seq gene expression counts, we selected two cell types that include endothelial cells (EC; 105 cells) and trophoblast-like cells (TB, 69 cells) from Chu et al (Chu et al. 2016). We fitted each gene using a zero-truncated negative binomial (ZTNB). Through the ZTNB model, we first inferred the gene-specific mean expression parameter λ_j and dispersion parameter ϕ_j in the negative binomial component of ZTNB through method of moments (Van den Berge et al. 2018). In particular, these parameter estimates are obtained iteratively through

$$\lambda_j^{(t+1)} = \frac{\sum_i Y_{ij}\left(1-\mathrm{f_{NB}}\left(\lambda_j^{(t)} N_i, \phi_j^{(t)}\right)\right)}{\sum_i N_i} \tag{2.6}$$

$$\phi_j^{(t+1)} = \frac{\sum_i \left(\lambda_j^{(t)} N_i\right)^2}{\sum_i Y_{ij}^2\left(1-\mathrm{f_{NB}}\left(\lambda_j^{(t)} N_i, \phi_j^{(t)}\right)\right) - \sum_i \left(\lambda_j^{(t)} N_i\right)^2 - \sum_i \left(\lambda_j^{(t)} N_i\right)}, \tag{2.7}$$

where Y_{ij} is the non-zero expression count for *i*-th cell and *j*-th gene in the real data (note that we ignored zero counts in this estimation step); N_i denotes the total read counts (i.e. read depth) for i-th cell; the superscripts (*t*) and (*t*+1) denote the *t*-th and (*t*+1)-th iteration estimates, respectively; $\mathrm{f_{NB}}(\cdot,\cdot)$ is the negative binomial density function.

In addition, we also followed (Van den Berge et al. 2018) to infer the zero proportion parameters p_{ij} in the ZTNB model by borrowing information across all genes. Specifically, we model the dropout probability for *i*-th cell and *j*-th gene, p_{ij}, using a semi-parametric additive logistic regression model:

$$z_{ij} \sim \text{Bern}(p_{ij}), \tag{2.8}$$

$$\log \frac{p_{ij}}{1-p_{ij}} = \text{s}(A_j) + \log(N_i) + \text{s}(A_j)\log(N_i), \tag{2.9}$$

where z_{ij} is an indicator on whether the observed count for *i*-th cell and *j*-th gene is zero or not; $\text{Bern}(p_{ij})$ denotes a Bernoulli distribution with the dropout parameter p_{ij}; $\text{s}(\cdot)$ is a non-parametric thin-plate spline; A_j is the average logarithm scale counts per million (CPM) calculated by aveLogCPM function in edgeR (Robinson et al. 2010). This way, the dropout probability becomes both cell-specific and gene-specific.

With the above estimated parameters, we simulated gene count through ZTNB model for both DE and non-DE genes. For DE genes, we simulated each DE effect size from a normal distribution with mean zero and standard deviation 3.5. For non-DE genes, we directly set the DE effect size to zero. We then calculated the true fold change of each gene as the exponential of effect size $\text{fc}_j = \exp(\beta_j)$, which is multiplied to the estimated mean gene expression levels $\hat{\lambda}_j$ in one population, resulting in a mean of $\hat{\mu}_{ij} = \hat{\lambda}_j \text{fc}_j$ for all cells in one population and a mean of $\hat{\mu}_{ij} = \hat{\lambda}_j$ for all cells in the other population. Afterwards, we simulated count data for *i*-th cell and *j*-th gene, C_{ij}, follows a negative binomial distribution $\text{NB}(\hat{\mu}_{ij}, \hat{\phi}_j)$. We set C_{ij} to be exactly zero with probability $\hat{p}_{ij}$. Note that the simulations do not exactly match the iDEA modeling assumptions, allows us to examine the robustness of iDEA.

We simulated the gene expression counts with τ_0 fixed to be -2, and with varying τ_1 (i.e., 0, 0.25, 0.5, 1.0, or 5.0) and varying CR (i.e., 1%, 2%, 5%, or 10%). With the simulated count data, we fitted different DE methods to obtain summary statistics, with which we fitted iDEA and other GSE methods. We evaluated the power of DE analysis and GSE analysis. To evaluate the type I error control of different GSE methods, we examined the null simulation settings ($\tau_1 = 0$). In each null setting, we permuted the gene labels 10 times to construct the permuted null sets, to which we applied different GSE methods. We then calculated the genomic inflation factor (λ_{gc}) for each GSE methods. Here, the genomic inflation factor (λ_{gc}) is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median. Specifically, we first convert the *p*-value for the gene sets to chi-squared test statistics then calculated λ_{gc} by dividing the resulting chi-squared test statistics by the expected median of a chi-squared distribution with one degree of freedom (0.4549364; qchisq (0.5,1) in R). To evaluate the power of different GSE methods, in each simulation setting, we obtained 1,000 simulation replicates with enriched pathways (i.e., $\tau_1 \neq 0$) and 9,000 simulation replicates without enriched pathways (i.e., $\tau_1 = 0$). We then evaluated the power of GSE analysis in detecting these 1,000 true signals given an FDR of 5%. To evaluate the power of different DE methods, we again computed power to detect true DE genes based on an FDR of 5%.

2.5.5 scRNA-seq datasets

We applied iDEA to analyze three published scRNA-seq data sets. The first scRNA-seq data is from Chu et al. (Chu et al. 2016) (GEO accession number GSE75748). It contains a total of 19,097 genes on 1,018 cells from seven cell types. The seven cell types include the human embryonic stem (ES) cell with subtypes H1 (212 cells) and H9 (162 cells); four ES derived linear-specific progenitor cell types that include neuronal progenitor cell (NPC, ectoderm derivatives,

173 cells), definitive endoderm derivatives cell (DEC, 138 cells), endothelial cell (EC, mesoderm derivatives, 105 cells), trophoblast-like cell (TB, extraembryonic derivatives, 69 cells); and human foreskin fibroblasts cell (HFF, 159 cells). We focused our analysis on five ES derived cell types (NPCs, DEs, ECs, TBs, and HFFs) and examined all pairs among them. For each pair, we filtered out lowly expressed genes that have more than 5 counts in at most two cells. The resulting number of analyzed genes ranges from 14,918 (EC vs TB) to 15,778 (DEC vs NPC). We considered batch information as a covariate when we fit zingeR to obtain summary statistics. For DE analysis of iDEA, we included the gene set vasculature development which is known to be important for vasculature progression and endothelial cell development (Tufro et al. 1999). To evaluate GSE analysis results, we examined the top 50 significant gene sets identified by each GSE method. To obtain the unbiased evaluation of different GSE methods, we used the R package RISmed (version 2.1.7) to query the related articles with the keywords: gene set name, cell type, and "embryonic development". We input one gene set at a time, and the number of gene set that do has the relevant literatures is counted to quantify the performance of different GSE methods.

The second scRNA-seq data is from Usoskin et al. (Usoskin et al. 2015) (GEO accession number GSE59739). This dataset contains a total of 19534 genes on 622 neuronal cells collected from the mouse lumbar dorsal root ganglion. These cells were classified into 11 neuronal cell types from four categories. The cell types include the neurofilament containing (NF) category: NF1 (31 cells), NF2 (48 cells), NF3 (12 cells) and NF4 (22 cells), NF5 (26 cells); nonpeptidergic nociceptors (NP) category: NP1 (125 cells), NP2 (32 cells), and NP3 (12 cells); peptidergic nociceptors (PEP) category: PEP1 (64 cells) and PEP2 (17 cells); and tyrosine hydroxylase containing (TH; 233 cells) category. NPs cell type versus the remaining cell types are shown in main results, while the pairs of NP1 cell type with each of the other ten cell types are

shown in **Figures S2.16 – 2.18**. For each pair, we filtered out lowly expressed genes that have more than 5 counts in at most two cells. The resulting number of analyzed genes ranges from 10,009 (comparing NP1 cell type vs NP3 cell type) to 10,948 (comparing NP1 cell type vs NF2 cell type). We included picking sessions as a covariate when we fit zingeR to obtain summary statistics (Van den Berge et al. 2018). For DE analysis of iDEA, we used the biological meaningful gene set neuron part (GO:0097458) (Guo et al. 2019) in the model.

The third scRNA-seq data is a peripheral blood mononuclear cells (PBMCs) data obtained from 10x Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k) (Zheng et al. 2017b). We downloaded the filtered gene/cell matrix that contains 2,700 cells and 32,738 genes. We processed the data using the R package Seurat (Butler et al. 2018) following the tutorial (https://satijalab.org/seurat/pbmc3k_tutorial.html) to obtain at a final set of 2,638 cells and 13,713 genes. We obtained clustering results from Seurat as shown in **Figure S2.16**. Here, we focus our analysis on 1,153 CD4+ T-cells and 305 CD8+ T-cells, to examine the performance of various methods in the challenging setting where the two examined cell types are similar to each other. We obtained summary statistics from zingeR and filtered out genes with *p*-values larger than 0.8 to focus on a final set of 1,696 genes. We did this due to the *p*-values obtained by the zingeR under the null are seriously left-skewed distributed. For DE analysis of iDEA, we used the gene signature CD8+ T-effector memory (Greenough et al. 2015) in the model.

In the real data applications, for both DE analysis and GSE analysis, we calculated power of different methods based on estimated FDR through permutations. Specifically, for DE analysis, we permuted the cell type label across cells ten times. We then applied different DE methods and obtained the empirical null distribution of test statistics (*p*-values or posterior estimates of $\boldsymbol{\gamma}$'s),

with which we calculated the empirical FDR for each threshold. For iDEA, in the permuted data, we also fixed the gene set enrichment parameters $\hat{\boldsymbol{\tau}}$ to be those estimated in the real data without re-estimating them. In our experience, re-estimating the enrichment parameters can lead to overly liberal FDR estimates and slows computation. For GSE analysis, we permuted the gene set label across all genes for each gene set ten times. We then applied different methods and obtained the enrichment *p*-values in the permuted null, with which we further calculated the empirical FDR for each *p*-value threshold.

2.5.6 Sensitivity analysis

In the main real data applications, we have fixed the hyperparameters for the inverse Gamma distribution ($a_\beta = 3.0, b_\beta = 20.0$, to ensure a prior mean $\frac{b_\beta}{a_\beta - 1}$ of 10) because there is insufficient information to estimate these parameters. Specifically, the inverse Gamma distribution serves as the prior for the variance parameter, which can be estimated by the effect sizes across many DE genes. Because there is only one variance parameter, it is impossible to estimate the hyperparameters in the inverse Gamma distribution for this variance parameter. Therefore, instead of estimating these hyperparameters, we performed sensitivity analysis to examine whether results would change with respect to the hyperparameters. To do so, we varied the hyperparameters and tested across a range of gene sets with different coverages in the three real datasets. Specifically, we varied the hyperparameters so that the prior mean of the inverse gamma distribution is 0.001, 0.1, 1, 10, 100. We also varied the coverage rate to be the 10^{th}, 30^{th},50^{th},70^{th}, 90^{th} percentile of the gene set size of the gene sets we used for the corresponding real data analysis. For example, for the human embryonic scRNA-seq dataset, we pick the gene set with coverage rate to be the 10^{th}, 30^{th},50^{th},70^{th}, 90^{th} percentile of the gene set size of the human gene sets we analyzed and set the

hyper parameter in the prior distribution of σ_β^2, (a_β, b_β) to be (3, 0.02), (3, 0.2), (3, 2), (3, 20), (3, 200) respectively. For the mouse sensory neuron scRNA-seq dataset and 10x Genomics PBMC scRNA-seq data, we followed the same procedure as the first scRNA-seq dataset.

2.5.7 Data and code availability

The datasets used in the present study are all publicly available. The human embryonic stem cell scRNA-seq dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75748. The Mouse sensory neuron scRNA-seq data is available at http://linnarssonlab.org/drg/. The 10x Genomics PBMC scRNA-seq data is available at 10x Genomics website https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k. For the gene sets we collected, the human gene sets are available from MSigDB databases http://software.broadinstitute.org/gsea/downloads.jsp and the mouse gene sets are available from the website http://www.informatics.jax.org/downloads/reports/index.html#go. The iDEA software package and source code have been deposited at www.xzlab.org/software.html. All scripts used to reproduce all the analysis is also available at the same website.

2.6 Supplementary Figures

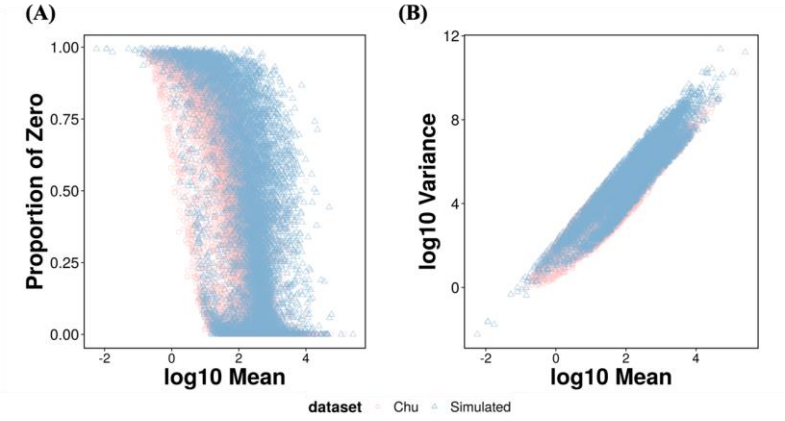


Figure S2.1 Characteristics of simulated data.

Simulated data has very similar characteristics as compared to the real scRNA-seq dataset. The data was simulated under the following parameters setting: $\tau_0 = -2, \tau_1 = 0.5$, and CR = 0.1. **(A)** Proportion of zero versus mean under log10 scale for both simulated data (blue) and real data (pink); **(B)** Mean-variance plot under log10 scale for both simulated data (blue) and real data (pink).

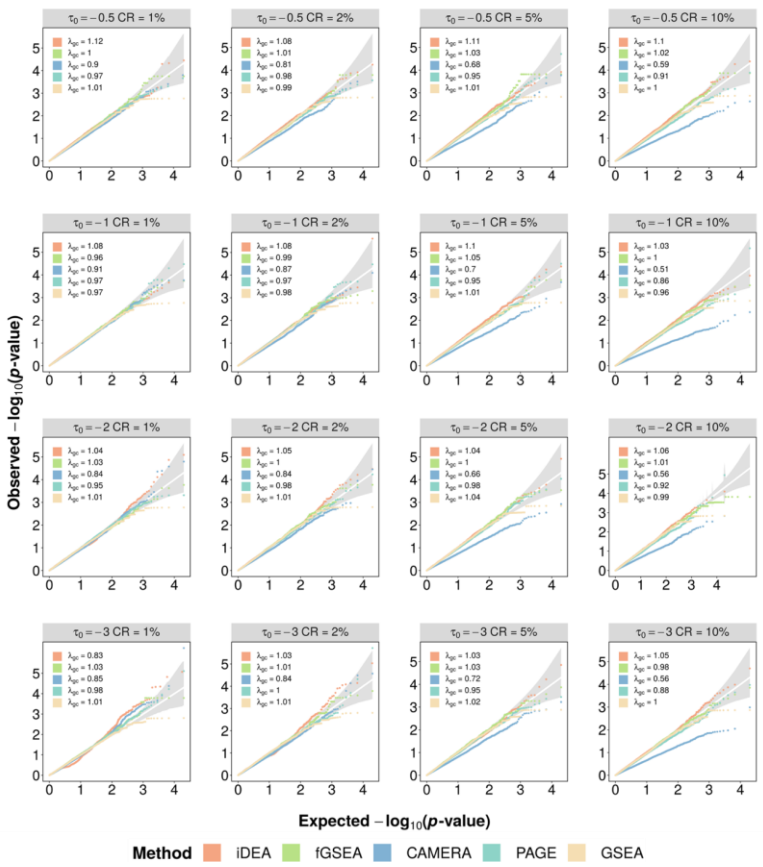


Figure S2.2 iDEA produces well-calibrated p-values for gene set enrichment analysis under null simulations.

Quantile-quantile plots of -log10(*p*-values) from iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown under different null scenarios with varying number of DE genes (denoted by the odd parameter τ_0; -0.5, -1.0, -2.0, or -3.0) and gene set coverage rates (CR; 1%, 2%, 5% or 10%). CR represents the percentage of genes inside the gene set. λ_{gc} is genomic control factor.

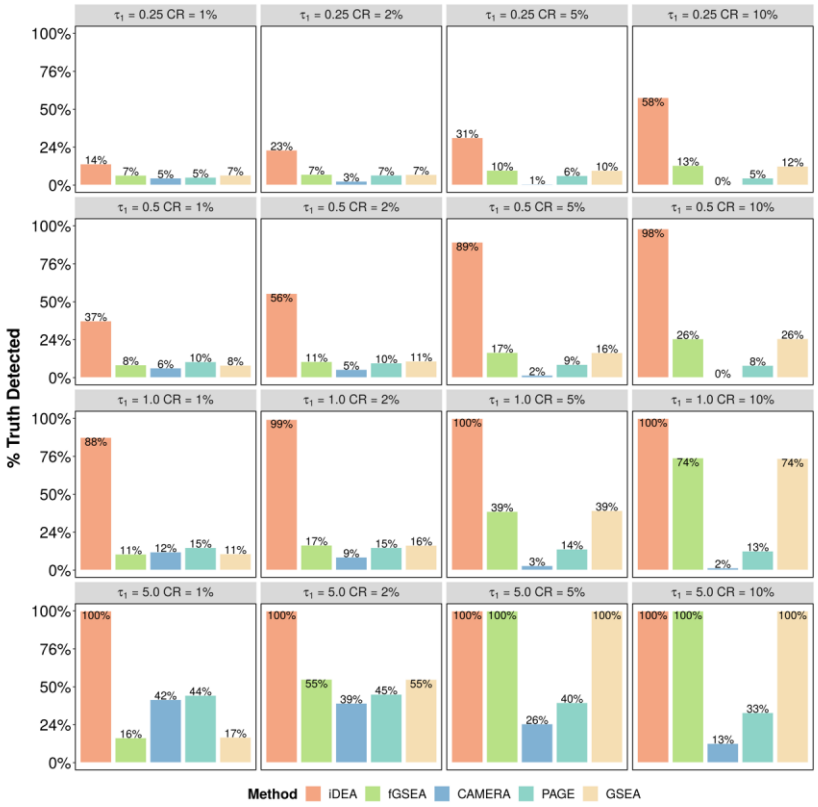


Figure S2.3 iDEA is more powerful than GSE methods for identifying enriched gene sets under alternative simulations.

The power plots from iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown under different scenarios with varying gene set enrichment coefficient (denoted by the odd parameter τ_1; 0.25, 0.5, 1.0 or 5.0) and gene set coverage rates (CR; 1%, 2%, 5% or 10%). CR represents the percentage of genes inside the gene set. Here, power was calculated based on an FDR of 5%.

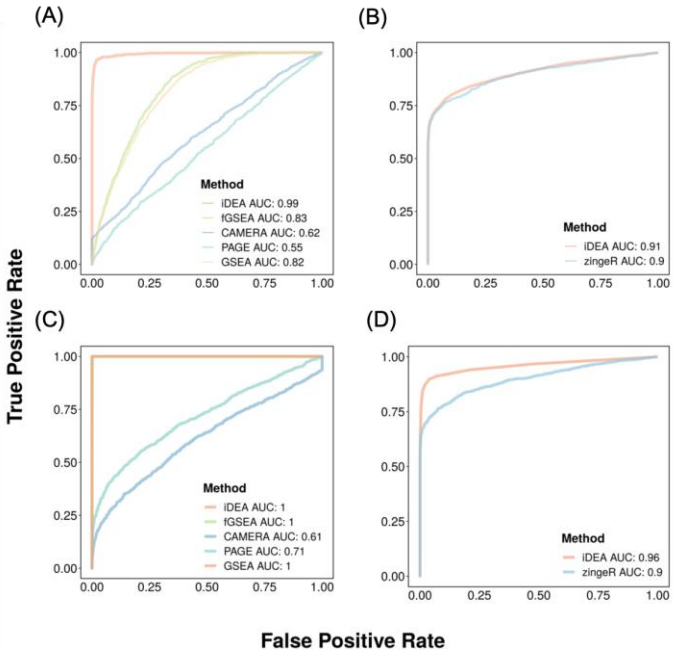


Figure S2.4 iDEA is more powerful for both GSE and DE analyses than existing approaches in power simulations.

The AUC of iDEA in identifying enriched pathways (**A** and **C**) and in identifying differentially expressed genes (**B** and **D**) are higher than that of the other methods. The compared GSE methods (**A** and **C**) include iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow). The compared DE methods (**B** and **D**) include iDEA (orange) and zingeR (skyblue). Simulations are performed under two parameter settings: $\tau_0 = -2$, $\tau_1 = 0.5$, and CR = 0.1 (**A** and **B**); $\tau_0 = -2$, $\tau_1 = 5$, and CR = 0.1 (**C** and **D**).

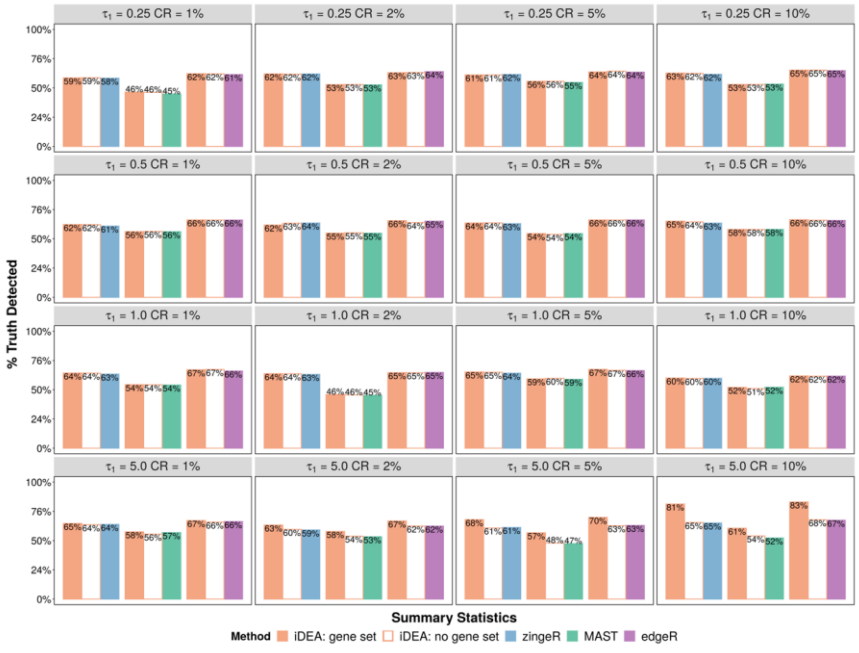


Figure S2.5 iDEA is more powerful than DE methods for identifying DE genes under alternative simulations when gene set enrichment parameter is larger.

Simulations were performed on one fixed scRNA-seq data set with $\tau_0 = -2$, varying τ_1 and CR. τ_1 is set to be 0.25, 0.5,1.0 or 5.0 and CR is set to be 1%, 2%, 5%, 10% respectively. In each simulation setting, power of DE results between common DE method (zinger (blue), MAST (green), edgeR (purple)) and iDEA (orange) with summary statistics obtained from that corresponding DE method when adding simulated gene set (filling color) or not (not filling color) is plotted. Here, power was calculated based on an FDR of 5%.

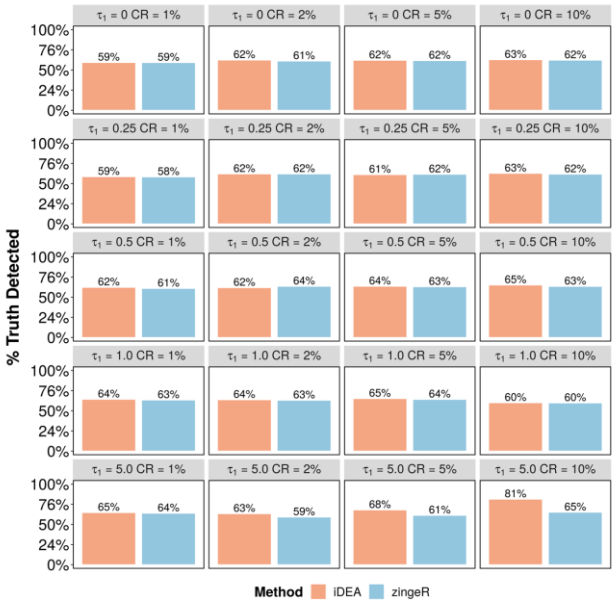


Figure S2.6 iDEA is more powerful in DE analysis than zingeR, when varying τ_1 and CR.

The data were simulated based on the parameter setting $\tau_0 = -2$, $\tau_1 = 0,\ 0.25, 0.5, 1.0$ or 5.0 and CR = 1%, 2%, 5% or 10%. iDEA identifies more significant gene sets on simulation studies when varying parameters. CR represents the percentage of genes inside the gene set. Here, power was calculated based on an FDR of 5%.

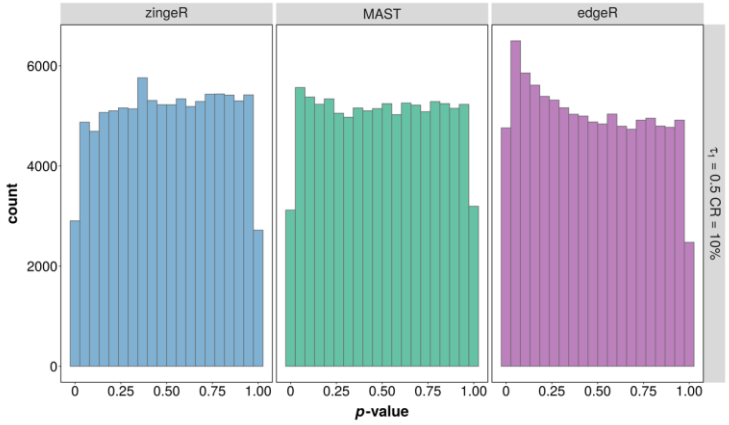


Figure S2.7 Distribution of marginal DE p-values from common DE methods.

The data were simulated based on the parameter setting $\tau_0 = -2$, $\tau_1 = 0.5$ and CR = 10%. *P*-values from zingeR (blue) and MAST (green) follow approximately a uniform distribution under the null while *P*-values from edgeR (purple) does not.

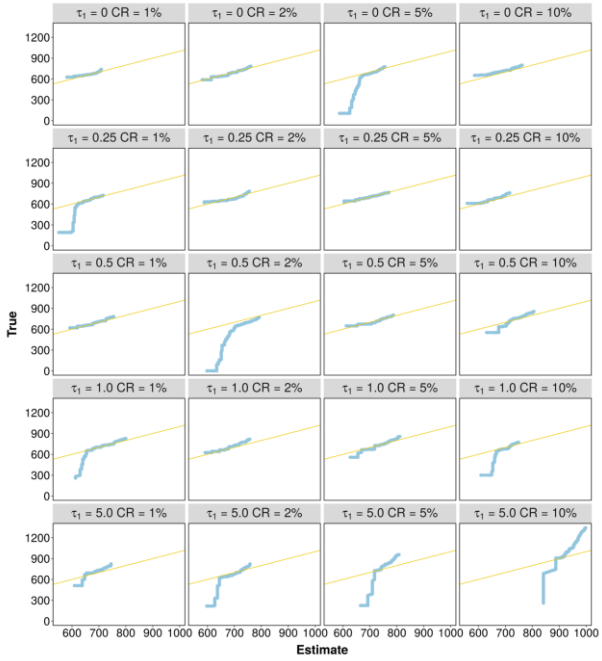


Figure S2.8 iDEA produces calibrated (or slightly conservative) FDR estimates.

Simulations were performed on one fixed scRNA-seq data set with $\tau_0 = -2$, varying τ_1 and CR. τ_1 is set to be 0, 0.25, 0.5,1.0 or 5.0 and CR is set to be 1%, 2%, 5%, 10% respectively. In each simulation setting, the scatterplot plot showed the number of detected signals based on true FDR (y-axis) versus the number of detected signals based on estimated FDR (x-axis). The yellow line is the reference line which represents y = x.

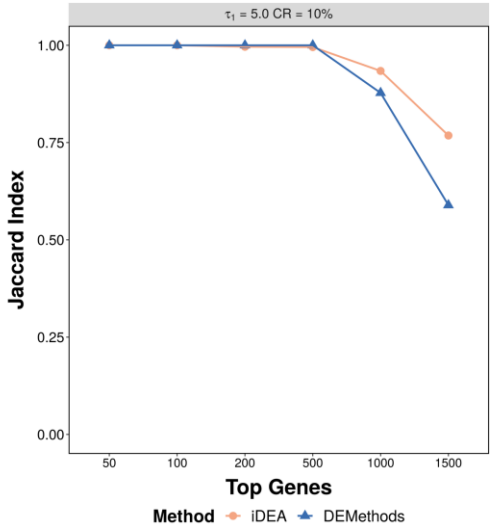


Figure S2.9 iDEA displays high consistency in detecting DE genes in simulations.

The data were simulated based on the parameter setting $\tau_0 = -2$, $\tau_1 = 5$, and CR = 0.1. The plot shows the Jaccard index for top DE genes between zingeR, edgeR, and MAST (blue) and the Jaccard index for top DE genes between iDEA when using summary statistics from zingeR, edgeR and MAST respectively(orange). CR represents the percentage of genes inside the gene set, τ_0 represents number of DE genes and τ_1 represents the gene set enrichment coefficient.

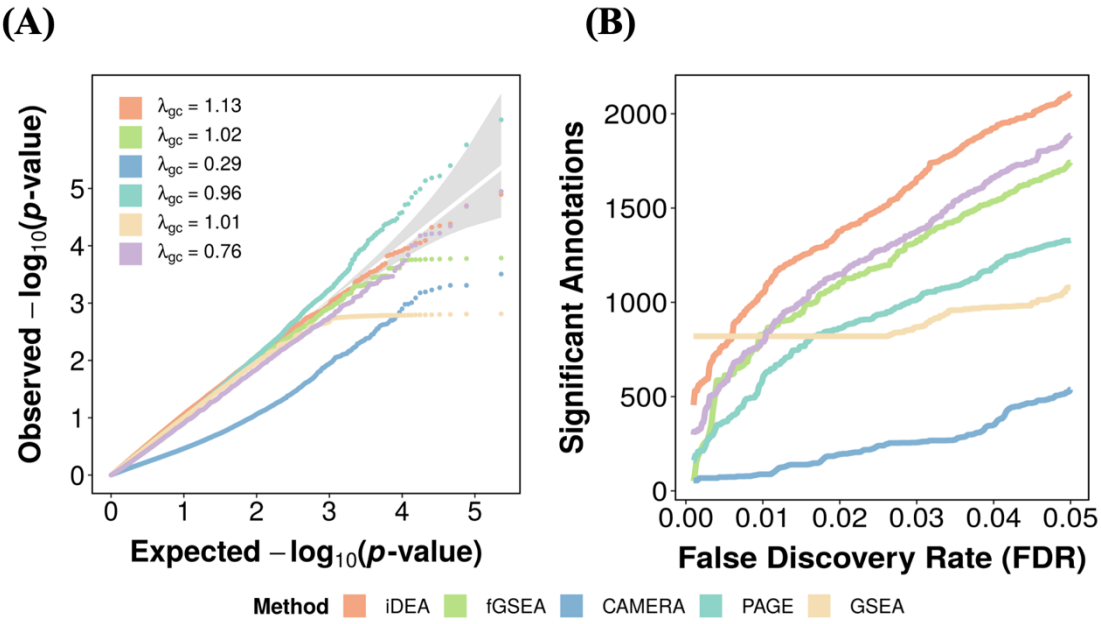


Figure S2.10 GSE Analysis including hypergeometric test results in human embryonic stem cell scRNA-seq dataset.

Results are shown for comparing definitive endoderm derivatives cell (DEC, 138 cells) and endothelial cell (EC, mesoderm derivatives, 105 cells). (**A**) Quantile-quantile plots of -log10(*p*-values) from GSE methods including iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue), GSEA (yellow) and Hypergeometric test (purple) are shown under permuted null; (**B**) Number of identified enriched gene sets by iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) GSEA (yellow) and Hypergeometric test (purple) are plotted against different empirical false discovery rates (FDR). Here λ_{gc} is the genomic control factor.

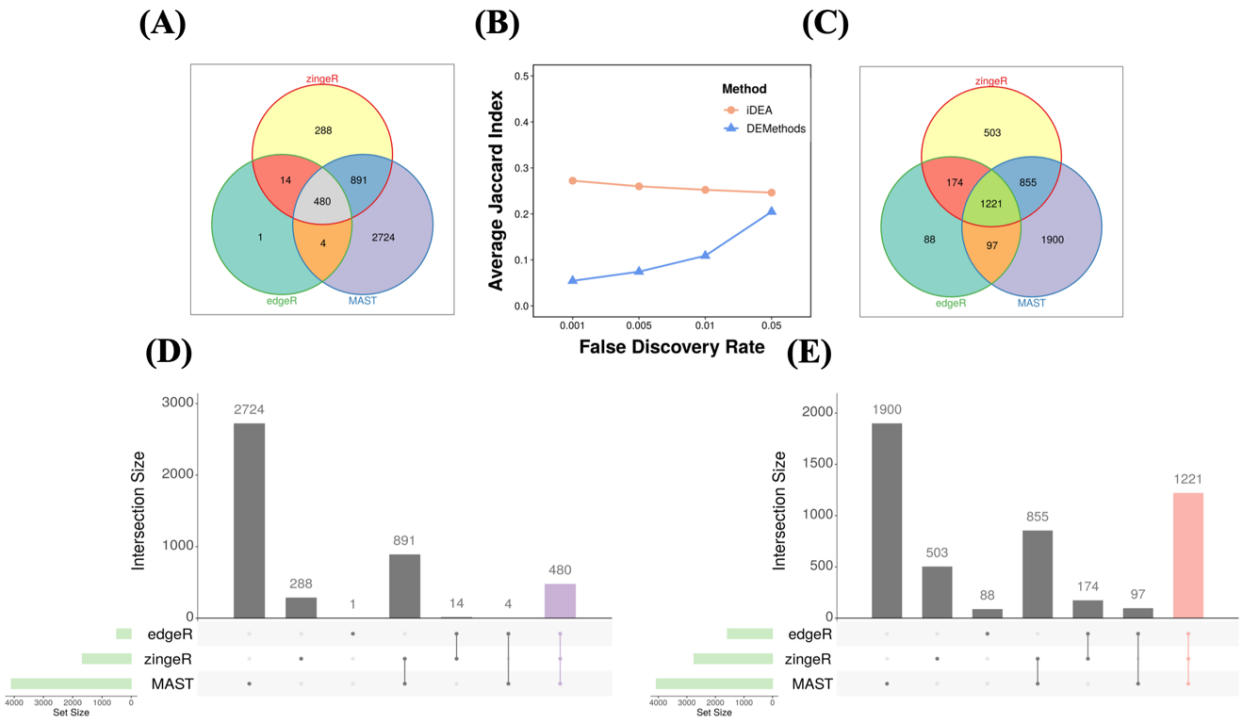


Figure S2.11 iDEA displays high consistency in detecting DE genes in human embryonic stem cell scRNA-seq data.

iDEA displays higher Jaccard index in the common DE genes. **(B)** Jaccard index for top DE genes at an FDR of 1% between zingeR, MAST, and edgeR, Jaccard index for top DE genes at an FDR of 1% between iDEA when using summary statistics from zingeR, MAST, and edgeR, respectively; **(A**) Overlap in top DE genes at an FDR of 1% between zingeR, MAST and edgeR; (**C**) Overlap in top DE genes at an FDR of 1% between iDEA when using summary statistics from zingeR, MAST and edgeR; **(D)** and **(E)** are just another visualization of the overlap corresponding to **(A)** and **(C)** by UpSetR (Conway, Lex and Gehlenborg 2017).

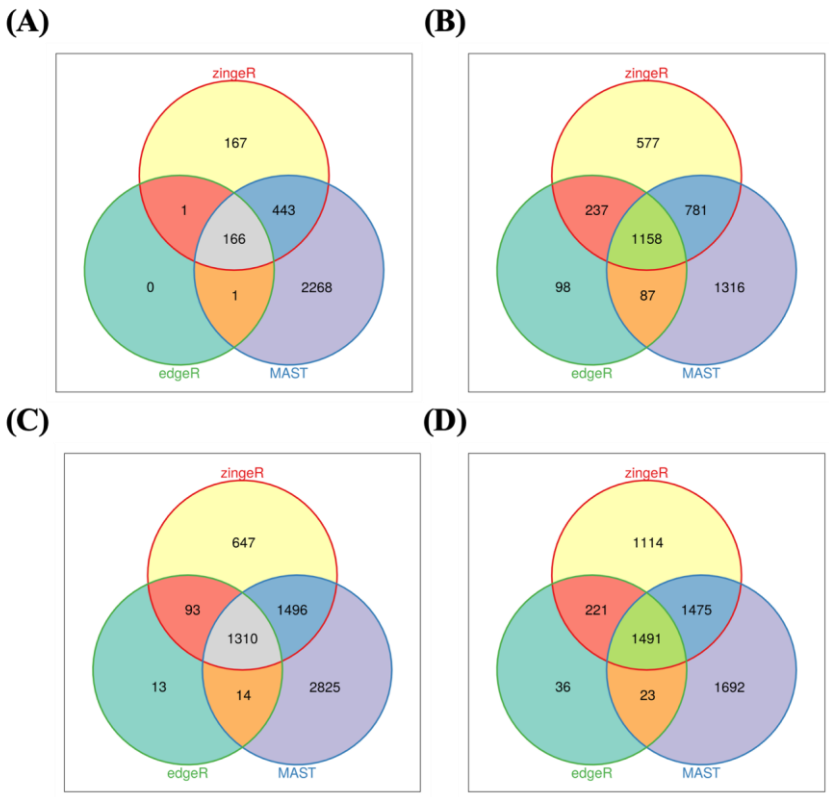


Figure S2.12 iDEA displays high consistency in detecting DE genes in human embryonic stem cell scRNA-seq data.

(**A**) Overlap in top DE genes at an FDR of 0.1% between zingeR, MAST and edgeR; (**B**) Overlap in top DE genes at an FDR of 0.1% between iDEA when using summary statistics from zingeR, MAST and edgeR respectively; (**C**) Overlap in top DE genes at an FDR of 5% between zingeR, MAST and edgeR; (**D**) Overlap in top DE genes at an FDR of 5% between iDEA when using summary statistics from zingeR, MAST and edgeR, respectively.

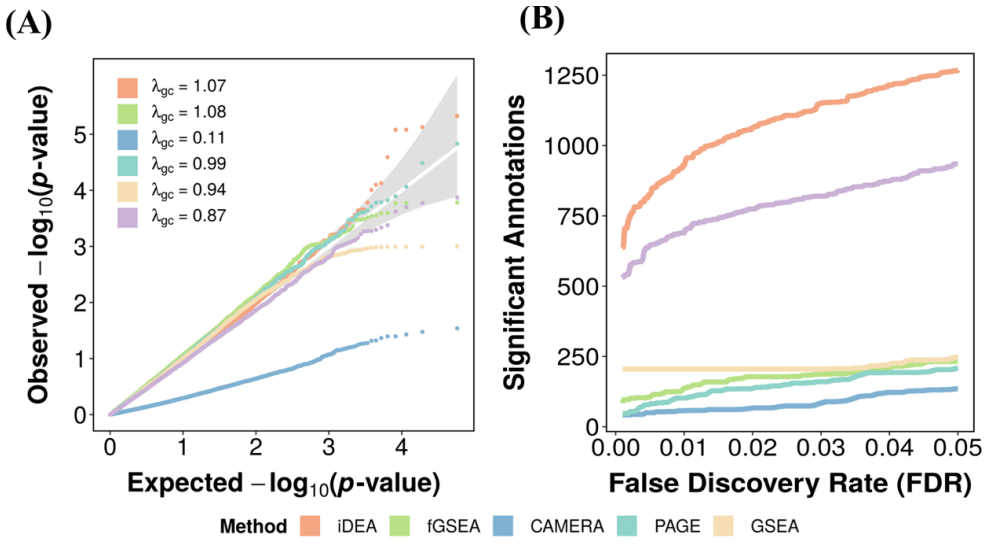


Figure S2.13 GSE Analysis including hypergeometric test results in mouse neuronal cell scRNA-seq dataset.

Results are shown for comparing nonpeptidergic nociceptors 1 (NP1) versus all the other cell types. (**A**) Quantile-quantile plots of -log10(*p*-values) from GSE methods including iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue), GSEA (yellow) and Hypergeometric test (purple) are shown under permuted null. (**B**) Number of identified enriched gene sets by iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) GSEA (yellow) and Hypergeometric test (purple) are plotted against different empirical false discovery rates (FDR). Here λ_{gc} is the genomic control factor.

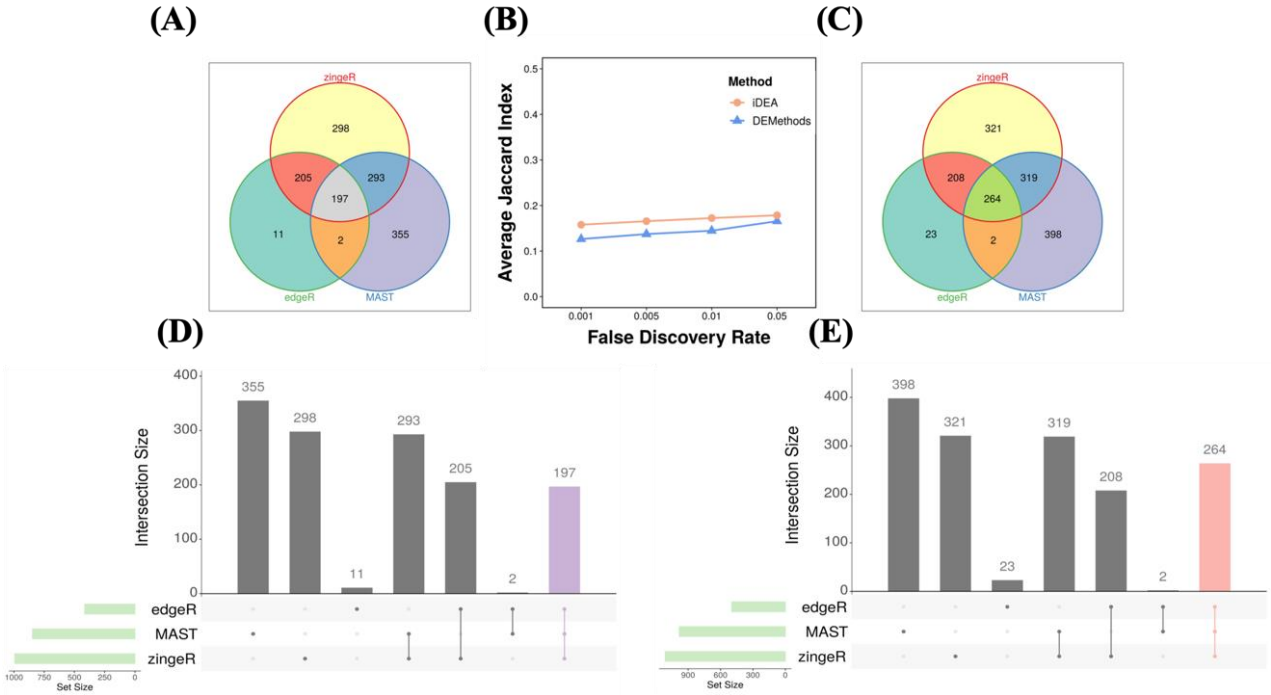


Figure S2.14 iDEA displays high consistency in detecting DE genes in mouse neuronal cell scRNA-seq data.

iDEA displays higher Jaccard index in the common DE genes. (**B**) Jaccard index for top DE genes at an FDR of 1% between zingeR, MAST and edgeR, Jaccard index for top DE genes at an FDR of 1% between iDEA when using summary statistics from zingeR, MAST and edgeR respectively(red); (**A**) Overlap in top DE genes at an FDR of 1% between zingeR, MAST and edgeR; (**C**) Overlap top DE genes at an FDR of 1% between iDEA when using summary statistics from zingeR, MAST and edgeR; **(D)** and **(E)** are just another visualization of the overlap corresponding to **(A)** and **(C)** by UpSetR (Conway et al. 2017).

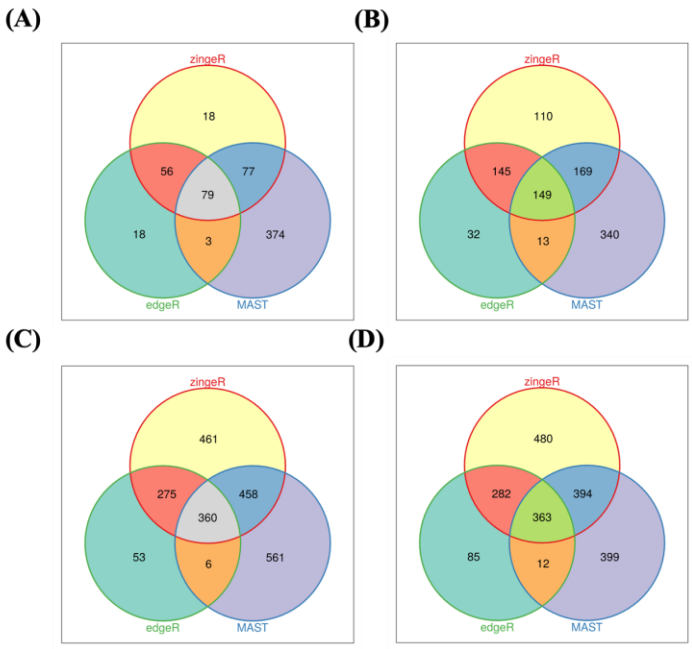


Figure S2.15 iDEA displays high consistency in detecting DE genes in mouse neuronal cell scRNA-seq data.

(**A**) Overlap in top DE genes at an FDR of 0.1% between zingeR, MAST and edgeR; (**B**) Overlap in top DE genes at an FDR of 0.1% between iDEA when using summary statistics from zingeR, MAST and edgeR respectively; (**C**) Overlap in top DE genes at an FDR of 5% between zingeR, MAST and edgeR; (**D**) Overlap in top DE genes at an FDR of 5% between iDEA when using summary statistics from zingeR, MAST and edgeR respectively.

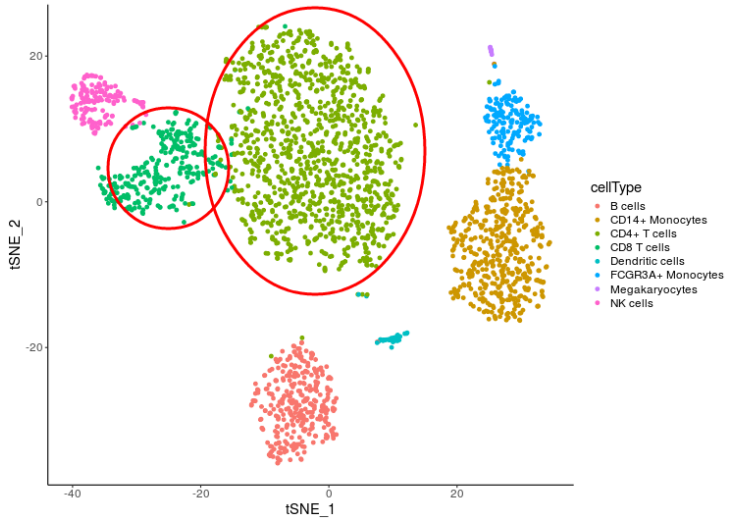


Figure S2.16 The scatterplot of first two t-SNE principal components for 10x Genomics data set.

There is a total of 8 cell types. CD4+ T cell type and CD8+ T cell type are highlighted in red circles. The cells are colored by the Seurat clustering method.

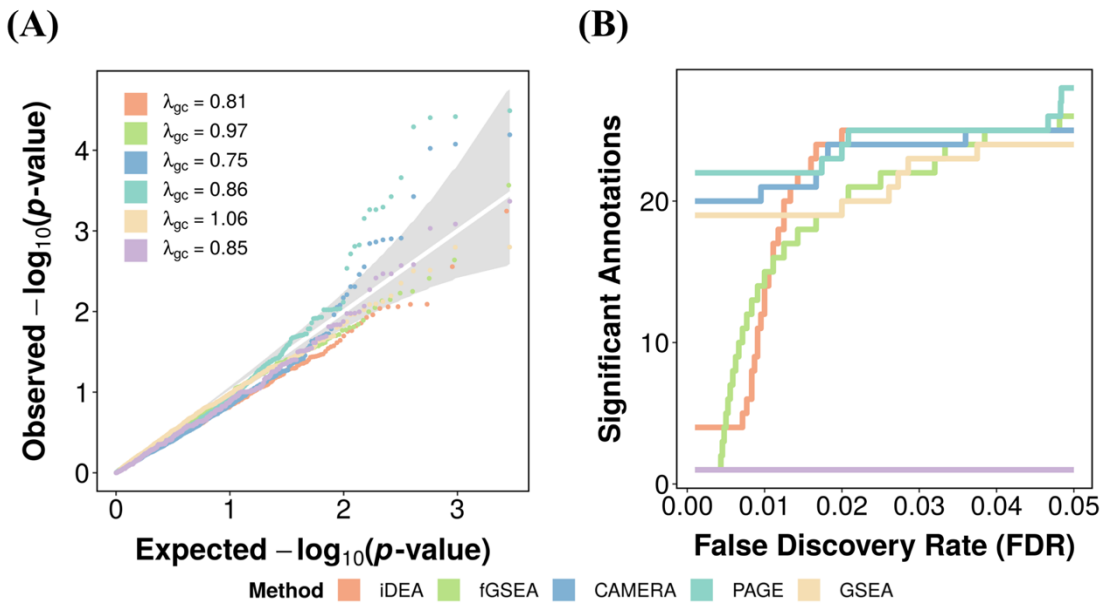


Figure S2.17 GSE Analysis including hypergeometric test results in 10x Genomics data set.

Results are shown for comparing CD4+ T cells versus CD8+ T cells. (**A**) Quantile-quantile plots of -log10(*p*-values) from GSE methods including iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue), GSEA (yellow) and Hypergeometric test (purple) are shown under permuted null. (**B**) Number of identified enriched gene sets by iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) GSEA (yellow) and Hypergeometric test (purple) are plotted against different empirical false discovery rates (FDR). Here λ_{gc} is the genomic control factor.

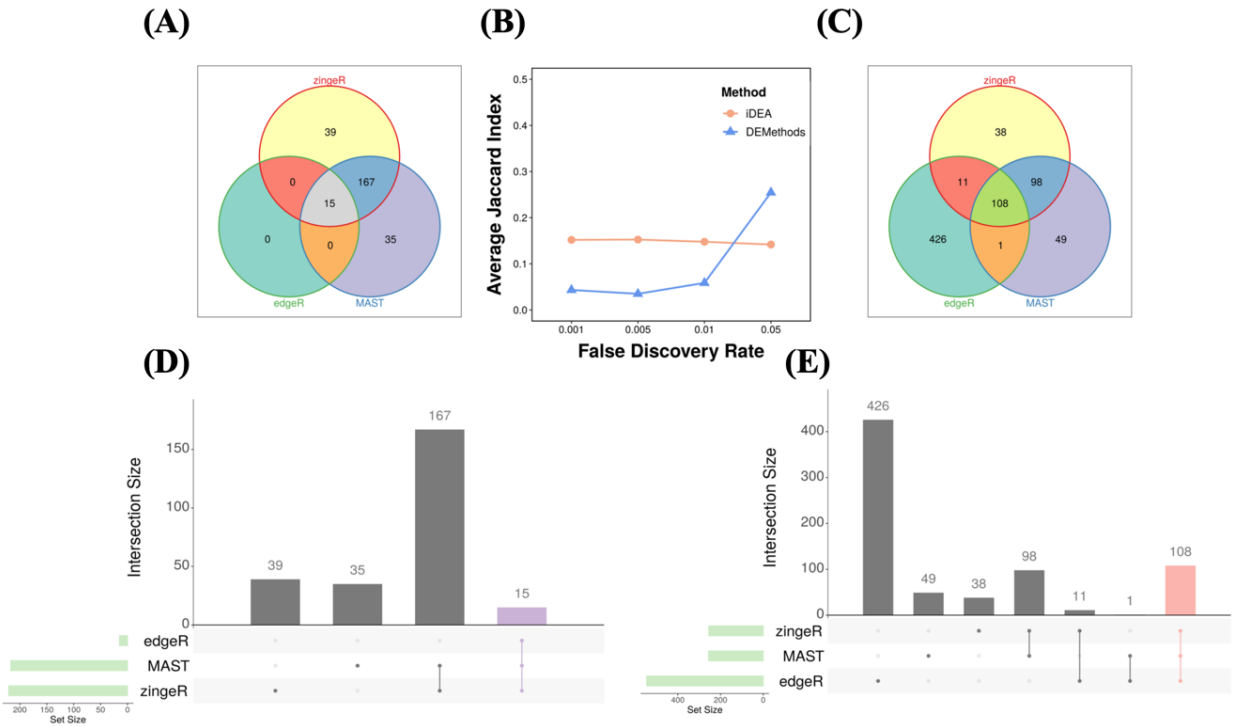


Figure S2.18 iDEA displays high consistency in detecting DE genes in 10x Genomics scRNA-seq data.

iDEA displays higher Jaccard index in the common DE genes. (**B**) Jaccard index for top DE genes at an FDR of 1% between zingeR, MAST and edgeR, Jaccard index for top DE genes at an FDR of 1% between iDEA when using summary statistics from zingeR, MAST and edgeR respectively(red); **(A**) Overlap in top DE genes at an FDR of 1% between zingeR, MAST and edgeR; (**C**) Overlap top DE genes at an FDR of 1% between iDEA when using summary statistics from zingeR, MAST and edgeR; **(D)** and **(E)** are just another visualization of the overlap corresponding to **(A)** and **(C)** by UpSetR (Conway et al. 2017).

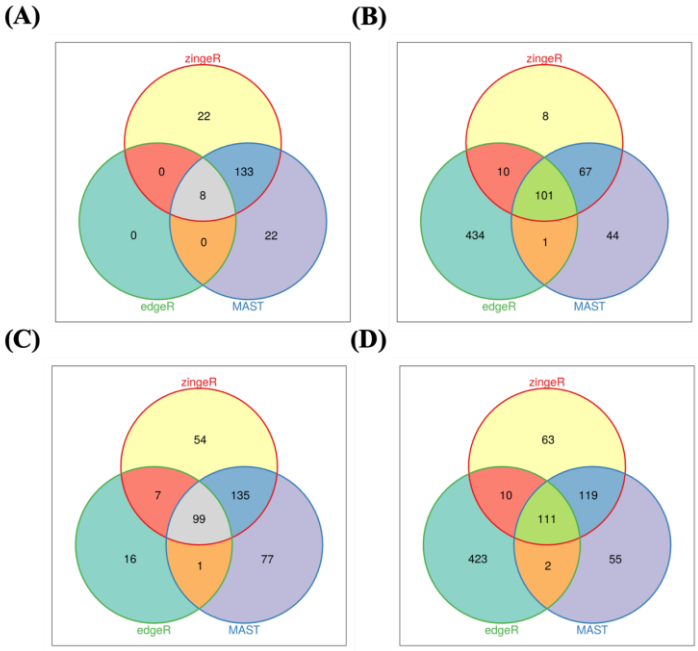


Figure S2.19 iDEA displays high consistency in detecting DE genes in 10x Genomics scRNA-seq data.

(**A**) Overlap in top DE genes at an FDR of 0.1% between zingeR, MAST and edgeR; (**B**) Overlap in top DE genes at an FDR of 0.1% between iDEA when using summary statistics from zingeR, MAST and edgeR respectively; (**C**) Overlap in top DE genes at an FDR of 5% between zingeR, MAST and edgeR; (**D**) Overlap in top DE genes at an FDR of 5% between iDEA when using summary statistics from zingeR, MAST and edgeR respectively.

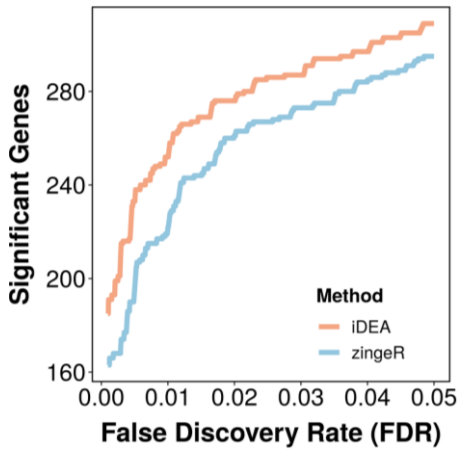


Figure S2.20 DE analysis results in the 10X Genomics scRNA-seq data.

Results are shown for comparing CD4+ T cells versus CD8+ T cells. The number of identified DE genes by iDEA (orange) and zingeR (blue) are plotted against different empirical FDR values. iDEA is more powerful than zingeR for DE analysis when there is no interesting gene set information provided.

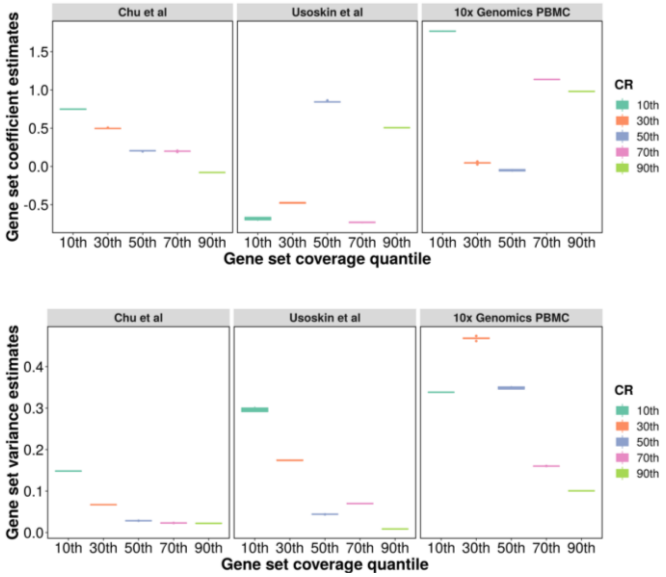


Figure S2.21 Sensitivity analysis of hyperparameters in prior distribution of σ_β^2.

Boxplot of the estimates of gene set coefficient and variance of gene set coefficient are displayed for three scRNA-seq datasets: human embryonic stem cell (Chu et al), mouse neuronal cell (Usoskin et al) and 10x Genomics PBMC scRNA-seq dataset. For each dataset, we tested the parameter estimates on gene sets with different coverage rate percentile among all gene sets we analyzed in that corresponding dataset. For each gene set with different coverage rate, estimates of gene set coefficient and variance were obtained under different prior distribution of σ_β^2. Parameter estimates are stable for gene sets across a wide range of prior distribution of σ_β^2. For each box plot, the bottom and the top of the box are the 25th and 75th quantiles, while the whiskers represent 1.5 * interquartile range from the lower and upper bounds of the box.

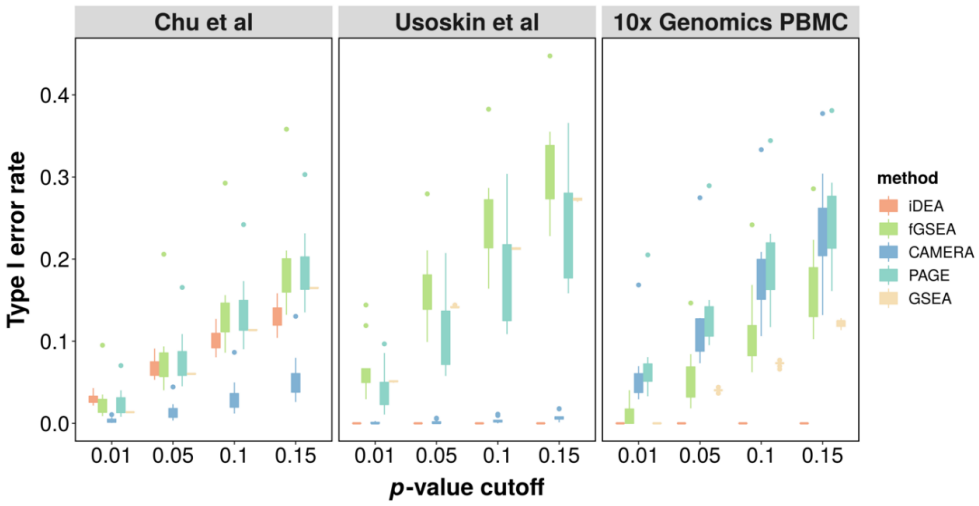


Figure S2.22 Type I error rate in real datasets.

We split the dataset within the same cell type (n = 10 replicates) to construct the true null distribution**.** Box plot of Type I error rate of iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown in human embryonic stem cell scRNA-seq dataset (Chu et al), mouse neuronal cell scRNA-seq dataset (Usoskin et al) and 10x Genomics PBMC scRNA-seq dataset. iDEA controlled type I error well in all three data sets. For each box plot, the bottom and the top of the box are the 25th and 75th quantiles, while the whiskers represent 1.5 * interquartile range from the lower and upper bounds of the box.

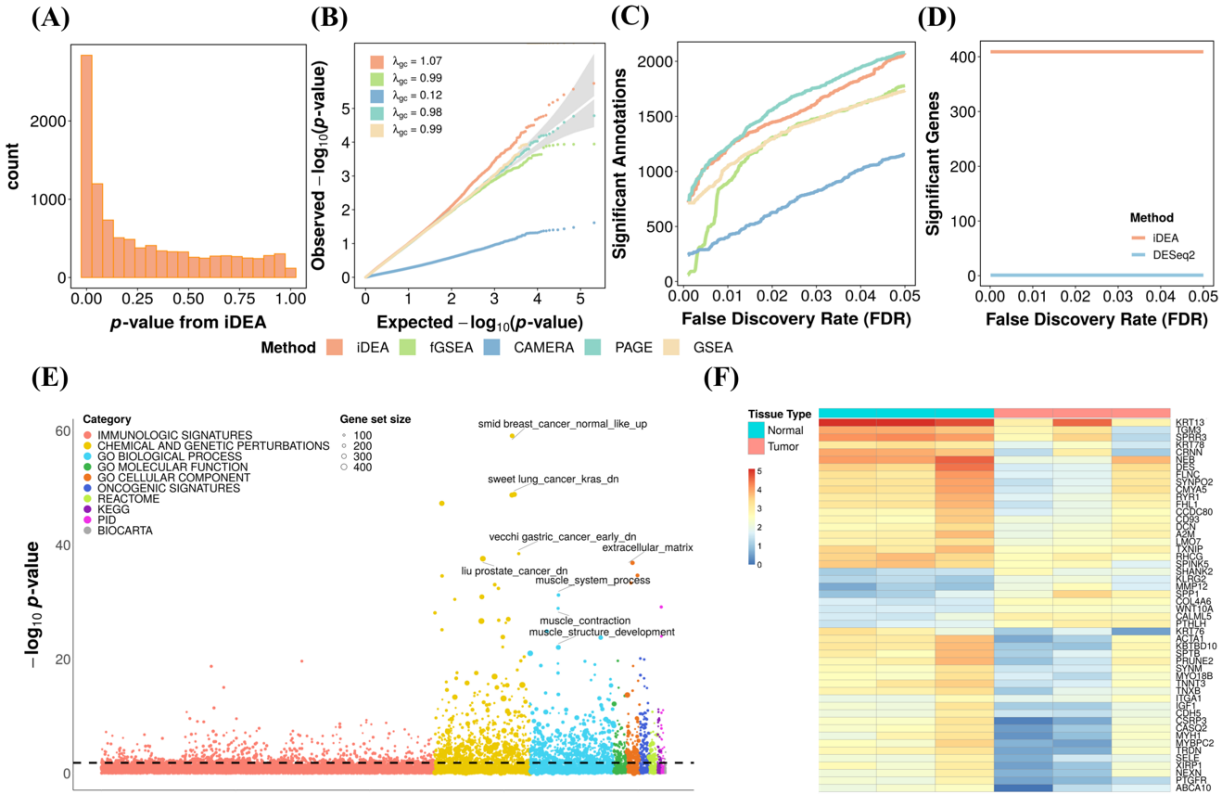


Figure S2.23 Analysis results in the bulk RNA-seq data.

Results are shown for comparing matched normal oral tissue versus oral squamous cell carcinoma. (**A**) *p*-values from iDEA for GSE analysis display expected enrichment of small *p*-values (for true signals) and a long flat tail towards large *p*-values. (**B**) Quantile-quantile plots of -log10(*p*-values) from GSE methods including iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are shown under permuted null. The *p*-values from iDEA, fGSEA, PAGE and GSEA are reasonably well calibrated. The *p*-values from CAMERA are overly conservative. Here λ_{gc} is the genomic control factor. (**C**) Number of identified enriched gene sets by iDEA (orange), fGSEA (green), CAMERA (navyblue), PAGE (skyblue) and GSEA (yellow) are plotted against different empirical false discovery rates (FDR). iDEA is as the same powerful as PAGE than other methods for GSE analysis. (**D**) Number of identified DE genes by iDEA (orange) and DESeq2 (blue) are plotted against different empirical FDR values. iDEA is more powerful than DESeq for DE analysis. (**E**) Heatmap shows the normalized expression level (log10-transformation with pseudo-count 0.1) for selected 50 DE genes (rows) identified by iDEA for cells in the two tissue types (columns). Genes are sorted by Hierarchical clustering; cells are ordered by tissue types (Normal: blue; Tumor: red). These DE genes clearly distinguish two compared tissues. (**F**) Bubble plot shows –log10 *p*-values for GSE analysis from iDEA (y-axis) for different gene sets. Gene sets are colored by ten categories: immunologic signatures (red), chemical and genetic perturbations (yellow), GO biological process (blue), GO molecular function (green), GO cellular component (orange), oncogenic signatures (deep blue), Reactome (grass-green), KEGG (purple), PID (rose), and Biocarta (grey). The size of the dot represents the number of genes contained in the gene set. Names for ten of the gene sets that are closely related to oral squamous cell carcinoma are highlighted in the panel.

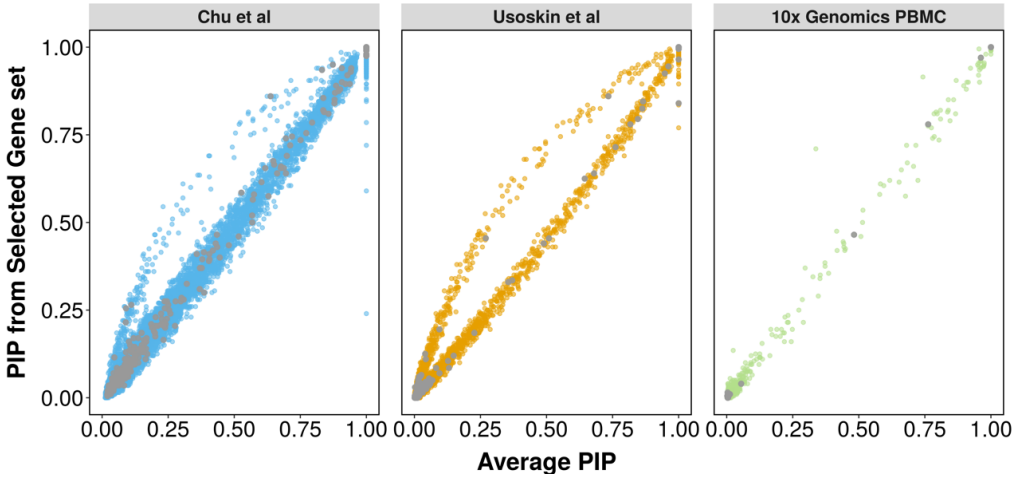


Figure S2.24 Posterior inclusion probabilities (PIPs) calculated by iDEA when adding specific gene set is highly correlated with averaging PIPs across all gene sets in all three scRNA-seq datasets.

Here, each dot represents each gene with x-axis represents the averaged pip a y-axis represents gene set specific pip. Genes in that selected gene set are highlighted by grey color. In Human embryonic stem cell scRNA-seq dataset (Chu et al), we added the gene set GO:0001944 (vasculature development). In Mouse Sensory neuron scRNA-seq dataset, we added the gene set GO:0097458 (neuron part). In 10x Genomics PBMC dataset, we added the gene set CD8+ T-effector memory Term.

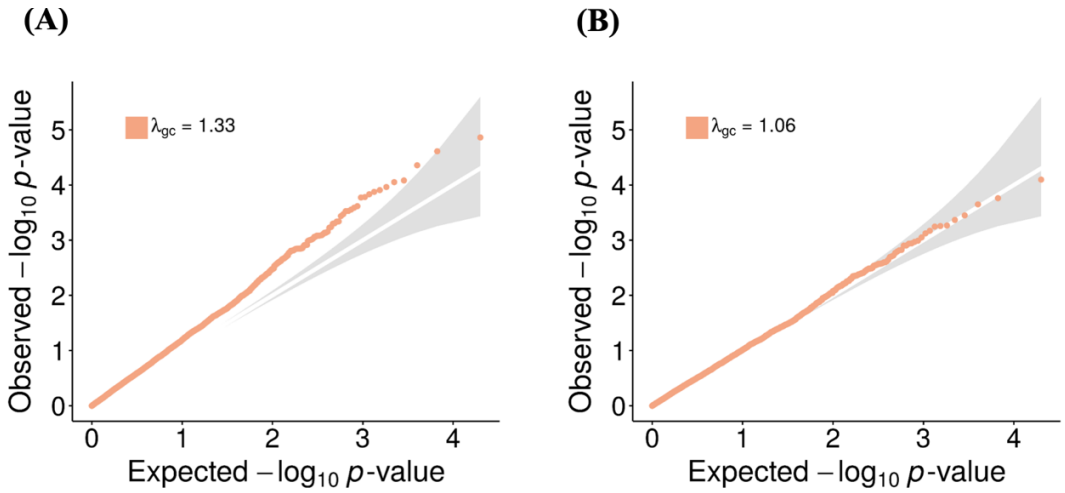


Figure S2.25 iDEA produces calibrated p-values in scRNA-seq based null simulations when using Louis Method to correct the observed information matrix.

Quantile-quantile plots of -log10(*p*-values) are shown for: iDEA without Louis Method (**A**); iDEA with Louis method **(B);** respectively under the null that simulated one fixed scRNA-seq data set and permute the gene set 10,000 times. Here, the other parameters are set to be $\tau_0 = -2, \tau_1 = 0$ and $\mathrm{CR} = 0.1$. CR represents the percentage of genes inside the gene set. λ_{gc} is genomic control factor.

2.7 Supplementary Tables

Gene Set	Coefficient	Variance	*P*-value
GO_VASCULATURE_DEVELOPMENT	0.000	0.016	7.340E-18
GO_BLOOD_VESSEL_MORPHOGENESIS	1.203	0.021	1.070E-16
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	1.267	0.024	3.650E-16
GO_ANGIOGENESIS	1.344	0.028	1.310E-15
LIU_PROSTATE_CANCER_DN	0.992	0.016	1.930E-15
SWEET_LUNG_CANCER_KRAS_DN	1.048	0.018	4.590E-15
ONDER_CDH1_TARGETS_2_DN	1.020	0.017	8.180E-15
SMID_BREAST_CANCER_NORMAL_LIKE_UP	1.110	0.021	9.870E-15
GO_SINGLE_ORGANISM_CELL_ADHESION	1.024	0.018	3.300E-14
GO_ANCHORING_JUNCTION	0.848	0.013	5.390E-14

Table S2.1 Top 10 enriched gene sets identified by iDEA on the human embryonic stem cell scRNA-seq dataset.

Note: only the top 10 gene sets identified by iDEA are listed here. For the full tables, please refer to https://www.nature.com/articles/s41467-020-15298-6.

Gene Ontology	Gene Ontology Term	Coefficient	Variance	*P*-value
GO:0044425	membrane part	1.040	0.004	2.260E-72
GO:0043005	neuron projection	1.240	0.007	8.680E-63
GO:0071944	cell periphery	0.996	0.005	4.630E-63
GO:0016020	membrane	0.995	0.005	3.970E-63
GO:0097458	neuron part	1.140	0.006	4.200E-62
GO:0005886	plasma membrane	0.994	0.005	2.510E-62
GO:0031224	intrinsic component of membrane	0.934	0.005	6.740E-56
GO:0044459	plasma membrane part	1.040	0.006	3.880E-51
GO:0045202	synapse	1.130	0.007	2.210E-50
GO:0006811	ion transport	1.180	0.008	3.850E-51

Table S2.2 Top 10 enriched gene sets identified by iDEA on the mouse neuronal cell scRNA-seq dataset.

Note: only the top 10 gene sets identified by iDEA are listed here. For the full tables, please refer to https://www.nature.com/articles/s41467-020-15298-6.

Signature	Project	Coefficient	Variance	*P*-value
CD8+ Tem	BLUEPRINT	1.780	0.121	2.960E-07
CD8+ Tem	NOVERSHTERN	2.240	0.359	1.860E-04
CD4+ memory T-cells	FANTOM	0.915	0.062	2.270E-04
CD8+ Tem	NOVERSHTERN	2.064	0.325	2.980E-04
CD4+ memory T-cells	FANTOM	0.798	0.054	5.770E-04
CD8+ Tcm	NOVERSHTERN	2.220	0.432	7.310E-04
CD8+ Tem	NOVERSHTERN	1.766	0.280	8.450E-04
CD8+ T-cells	HPCA	1.618	0.236	8.710E-04
CD8+ Tem	HPCA	1.039	0.107	1.468E-03
CD8+ Tem	BLUEPRINT	0.934	0.087	1.574E-03

Table S2.3 Top 10 enriched gene sets identified by iDEA on the 10x Genomics PBMC scRNA-seq dataset.

Note: only the top 10 gene sets identified by iDEA are listed here. For the full tables, please refer to https://www.nature.com/articles/s41467-020-15298-6.

Method	Signature	Project	*P*-value	Adjust *p*-value/FDR
fGSEA	CD4+ memory T-cells	FANTOM	2.490E-04	3.987E-03

	CD8+ Tem	NOVERSHTERN	2.740E-04	3.987E-03
	CD8+ Tem	NOVERSHTERN	2.790E-04	3.987E-03
	NK cells	HPCA	2.840E-04	3.987E-03
	CD8+ Tem	NOVERSHTERN	2.870E-04	3.987E-03
	NK cells	HPCA	2.870E-04	3.987E-03
	NK cells	HPCA	2.970E-04	3.987E-03
	CD8+ Tem	BLUEPRINT	3.400E-04	3.987E-03
	CD8+ Tem	HPCA	3.550E-04	3.987E-03
	Tgd cells	HPCA	3.580E-04	3.987E-03
CAMERA	CD8+ Tem	BLUEPRINT	5.650E-22	8.140E-20
	CD8+ Tem	HPCA	2.270E-13	1.640E-11
	Tgd cells	HPCA	6.660E-13	3.190E-11
	CD8+ Tem	HPCA	5.270E-12	1.900E-10
	NK cells	HPCA	9.870E-12	2.840E-10
	CD8+ Tem	BLUEPRINT	2.570E-11	6.150E-10
	NK cells	HPCA	2.990E-11	6.150E-10
	Tgd cells	HPCA	5.960E-10	1.070E-08
	CD8+ Tem	NOVERSHTERN	1.610E-07	2.500E-06
	NK cells	HPCA	1.740E-07	2.500E-06
PAGE	CD8+ Tem	BLUEPRINT	5.140E-44	0.000
	CD8+ Tem	HPCA	1.240E-26	0.000
	Tgd cells	HPCA	1.580E-25	0.000
	CD8+ Tem	BLUEPRINT	5.150E-23	0.000
	CD8+ Tem	HPCA	1.060E-21	0.000
	Tgd cells	HPCA	3.610E-21	0.000
	NK cells	HPCA	1.200E-20	0.000
	CD8+ Tem	NOVERSHTERN	2.190E-19	0.000
	CD8+ Tem	NOVERSHTERN	4.580E-19	0.000
GSEA	CD4+ MEMORY T-Cells	FANTOM	0.000E+00	0.036
	CD4+ MEMORY T-Cells	FANTOM	0.000E+00	0.026
	CD8+ T-Cells	HPCA	0.000E+00	0.023

	CD8+ TCM	HPCA	0.000E+00	0.006
	CD8+ TCM	NOVERSHTERN	0.000E+00	0.007
	CD8+ TEM	BLUEPRINT	0.000E+00	0.000
	CD8+ TEM	BLUEPRINT	0.000E+00	0.000
	CD8+ TEM	HPCA	0.000E+00	0.000
	CD8+ TEM	HPCA	0.000E+00	0.000
	CD8+ TEM	NOVERSHTERN	0.000E+00	0.000

Table S2.4 Top 10 gene sets identified by fGSEA, CAMERA, PAGE, GSEA respectively.

Note: only the top 10 gene sets identified by iDEA are listed here. For the full tables, please refer to https://www.nature.com/articles/s41467-020-15298-6.

***P*-value**	**Adjusted *p*-value**	**Set**	**Count**	**Set2**
5.456E-17	1.970E-13	intersection	332	GO_BLOOD_VESSEL_MORPHOGENESIS
1.132E-02	1.000E+00	set1	97	GO_BLOOD_VESSEL_MORPHOGENESIS
NA	NA	set2	0	GO_BLOOD_VESSEL_MORPHOGENESIS
3.397E-04	1.000E+00	intersection	49	SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP
2.937E-15	1.060E-11	set1	380	SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP
8.564E-20	3.092E-16	set2	261	SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP
1.327E-15	4.793E-12	intersection	268	GO_ANGIOGENESIS
1.031E-03	1.000E+00	set1	161	GO_ANGIOGENESIS
NA	NA	set2	0	GO_ANGIOGENESIS
2.819E-04	1.000E+00	intersection	41	LIU_PROSTATE_CANCER_DN
1.816E-14	6.557E-11	set1	388	LIU_PROSTATE_CANCER_DN
9.917E-17	3.581E-13	set2	383	LIU_PROSTATE_CANCER_DN
4.875E-06	1.760E-02	intersection	57	SWEET_LUNG_CANCER_KRAS_DN
1.097E-12	3.960E-09	set1	372	SWEET_LUNG_CANCER_KRAS_DN
4.693E-14	1.695E-10	set2	320	SWEET_LUNG_CANCER_KRAS_DN

Table S2.5 Results for the top gene set GO:0001944 with the combinations of the top 5 gene sets in the human embryonic stem cell scRNA-seq dataset.

Note: only combinations between the top 5 gene sets are listed here. For the full tables, please refer to https://www.nature.com/articles/s41467-020-15298-6. From the second gene set to the 50th gene set, we calculate the adjusted p-values for their intersection with the top first gene set as well as the disjoint parts. P-values were determined by two-sided Wald test and adjusted by Bonferroni correction.

***P*-value**	**Adjusted *p*-value**	**Set**	**Count**	**Set2**
5.530E-40	1.865E-36	intersection	674	GO:0043005
4.480E-22	1.511E-18	dis1	3014	GO:0043005
4.015E-10	1.354E-06	dis2	447	GO:0043005
1.497E-50	5.047E-47	intersection	2021	GO:0071944
1.346E-06	4.538E-03	dis1	1667	GO:0071944
2.991E-02	1.000E+00	dis2	819	GO:0071944
2.696E-55	9.092E-52	intersection	3688	GO:0016020
NA	NA	dis1	0	GO:0016020
9.927E-01	1.000E+00	dis2	1393	GO:0016020
1.179E-40	3.975E-37	intersection	878	GO:0097458
3.187E-19	1.075E-15	dis1	2810	GO:0097458
2.933E-09	9.890E-06	dis2	594	GO:0097458
1.920E-51	6.474E-48	intersection	2011	GO:0005886
2.378E-06	8.018E-03	dis1	1677	GO:0005886
6.089E-02	1.000E+00	dis2	751	GO:0005886

Table S2.6 Results for the top gene set GO:0044425 with the combinations of the top 5 gene sets in the mouse neuronal cell scRNA-seq dataset.

Note: only combinations between the top 5 gene sets are listed here. For the full tables, please refer to https://www.nature.com/articles/s41467-020-15298-6. From the second gene set to the 50th gene set, we calculate the adjusted p-values for their intersection with the top first gene set as well as the disjoint parts. P-values were determined by two-sided Wald test and adjusted by Bonferroni correction.

Chapter 3 Spatially Informed Cell Type Deconvolution for Spatial Transcriptomics

3.1 Abstract

Many spatially resolved transcriptomic technologies do not have single-cell resolution but measure the average gene expression for each spot from a mixture of cells of potentially heterogeneous cell types. Here, we introduce a deconvolution method, conditional autoregressive deconvolution (CARD), that combines cell type–specific expression information from single-cell RNA sequencing (scRNA-seq) with correlation in cell type composition across tissue locations. Modeling spatial correlation allows us to borrow the cell-type composition information across locations, improving accuracy of deconvolution even with a mismatched scRNA-seq reference. CARD can also impute cell type compositions and gene expression levels at unmeasured tissue locations, enable the construction of a refined spatial tissue map with a resolution arbitrarily higher than that measured in the original study, and perform deconvolution without a scRNA-seq reference. Applications to four datasets including a pancreatic cancer dataset identified multiple cell types and molecular markers with distinct spatial localization that define the progression, heterogeneity, and compartmentalization of pancreatic cancer.

3.2 Introduction

Spatially resolved transcriptomic technologies perform gene expression profiling on many tissue locations with spatial localization information (Burgess 2019), enabling the characterization of transcriptomic landscape on tissues (Soldatov et al. 2019, Prinz et al. 2011, Svensson, Teichmann and Stegle 2018, Dries et al. 2021, Pham et al. 2020, Biancalani et al. 2021, Fu et al.

2021, Fischl et al. 2014, Moses and Pachter 2022). Despite fast technological development, however, most technologies are of limited spatial resolution. Almost all sequencing-based technologies collect expression measurements on tissue locations that consist of a few to a few dozen single cells belonging to potentially distinct cell types (Asp et al. 2020, 10XGenomics , Rodriques et al. 2019, Ståhl et al. 2016). Because each measured location contains a mixture of cells, these sequencing-based technologies effectively quantify the average expression level across many cells on the location. Consequently, performing cell type deconvolution on tissue locations becomes an essential analytic task for disentangling the spatial localization of cell types and characterizing the complex tissue architecture (Liao et al. 2020, Rao et al. 2021).

Deconvolution of spatial transcriptomics data requires cell type specific gene expression information and tailored spatial methods. Cell type specific gene expression information are nowadays readily available from single-cell RNA sequencing (scRNA-seq) studies (Hwang et al. 2018), which have been previously used for deconvoluting bulk RNA-seq data (Cobos et al. 2020) by recently developed deconvolution methods including MuSiC (Wang et al. 2019), SCDC (Dong et al. 2020), and Bisque (Jew et al. 2020). These methods can in principle be directly applied to spatial transcriptomics and are being adapted so by several recently developed methods (Elosua-Bayes et al. 2021, Song and Su 2021, Lopez et al. 2022, Biancalani et al. 2021, Danaher et al. 2022, Gayoso et al. 2022, Andersson et al. 2020, Kleshchevnikov et al. 2022, Dong and Yuan 2021, Cable et al. 2021) such as RCTD (Cable et al. 2021), stereoscope (Andersson et al. 2020), SPOTlight (Elosua-Bayes et al. 2021), cell2location (Kleshchevnikov et al. 2022), and spatialDWLS (Dong and Yuan 2021) (details in **APPENDIX B.1**). All these methods, however, do not make use of the rich spatial localization information available in spatial transcriptomics.

Spatial localization information in spatial transcriptomics measures the relative distance between tissue locations and contains potentially invaluable information for deconvolution. Specifically, a tissue is composed of multiple cell types that are segregated in a spatially correlated fashion into tissue domains (Stoltzfus et al. 2020, Dudas et al. 2008, Bove et al. 2017, van Vliet et al. 2018), which are characterized by a domain-specific composition of cell types, with similar cell types colocalized spatially (Phillips et al. 2021, Schürch et al. 2020). Histological characterization of various tissues (Hawrylycz et al. 2014b, 10XGenomics), including the Hematoxylin and Eosin (H&E) staining images accompanying spatial transcriptomics datasets (Ståhl et al. 2016), highlight the spatial segregation of cell types and neighboring cell type composition similarity. In single cell resolution spatial transcriptomics (Xia et al. 2019, Eng et al. 2019), we also observed that similar cell types tend to colocalize, with colocalization pattern decaying with distance (Ma and Zhou 2022). Consequently, neighboring locations on the tissue likely contain more similar cell type compositions as compared to locations that are far away. Therefore, modeling the neighborhood similarity in cell type compositions and accommodating their spatial correlation would allow us to borrow composition information across locations on the entire tissue section to enable accurate deconvolution of spatial transcriptomics on each individual location.

Here, we develop a method, named Conditional AutoRegressive based Deconvolution (CARD), to perform such spatially informed deconvolution of cell types for spatial transcriptomic. CARD builds upon a non-negative matrix factorization model to use the cell type specific gene expression information from scRNA-seq data for deconvoluting spatial transcriptomics. A unique feature of CARD is its ability to accommodate the spatial correlation structure in cell type composition across tissue locations by a conditional autoregressive modeling assumption (Banerjee, Carlin and Gelfand 2014, Lee 2011). As a result, CARD can take advantage of the

spatial correlation structure to enable accurate and robust deconvolution of spatial transcriptomics across technologies with different spatial resolutions and in the presence of mismatched scRNA-seq references. In addition, modeling spatial correlation allows CARD to impute cell type compositions as well as gene expression levels on new locations of the tissue, facilitating the construction of a refined spatial map with an arbitrarily high resolution for any spatial transcriptomics technologies -- both these features are in direct contrast to a recent method BayesSpace (Zhao et al. 2021) that can only enhance Spatial Transcriptomics (ST) or 10x Visium data with a fixed resolution of either six or nine times higher than that of the original. Importantly, an extension of CARD is also capable of performing reference-free deconvolution without a scRNA-seq reference. We develop a computationally efficient algorithm for constrained maximum likelihood inference, making CARD scalable to data with tens of thousands of spatial locations and tens of thousands of genes. We illustrate the benefits of CARD through extensive simulations and applications to four published spatial transcriptomics studies with distinct technologies, spatial resolutions, tissue structures, and scRNA-seq references.

3.3 Results

3.3.1 Simulations

CARD is described in **Methods**, with its method schematic shown in **Figure 3.1**. We performed simulations to evaluate the performance of CARD and compared it with six existing deconvolution methods: MuSiC, SPOTlight, RCTD, cell2location, spatialDWLS, and stereoscope (details in **Methods**). Briefly, we used a scRNA-seq data (Zeisel et al. 2018) to construct spatial transcriptomics and we varied a noise level parameter p_n to modify cell type compositions and spatial correlation patterns across locations (**Figure S3.1**). The simulated data are realistic, preserving data features observed in the published spatial transcriptomics data (**Figure S3.2**). We

examined four simulation settings, each of which consists of five simulation replicates. In each replicate, we applied various deconvolution methods to deconvolute the spatial transcriptomics data, using either the same set of scRNA-seq data or its modified version or another set as reference. We then followed (Wang et al. 2019) and quantified the deconvolution performance by computing the root mean square error (RMSE) between the estimated cell type composition and the underlying truth on each location. We primarily displayed RMSE difference plots where we contrasted the RMSE of other methods with respect to CARD following (Yang and Zhou 2020, Zhou et al. 2013). We kept the original RMSE and rank plots in the supplements, which show consistent results.

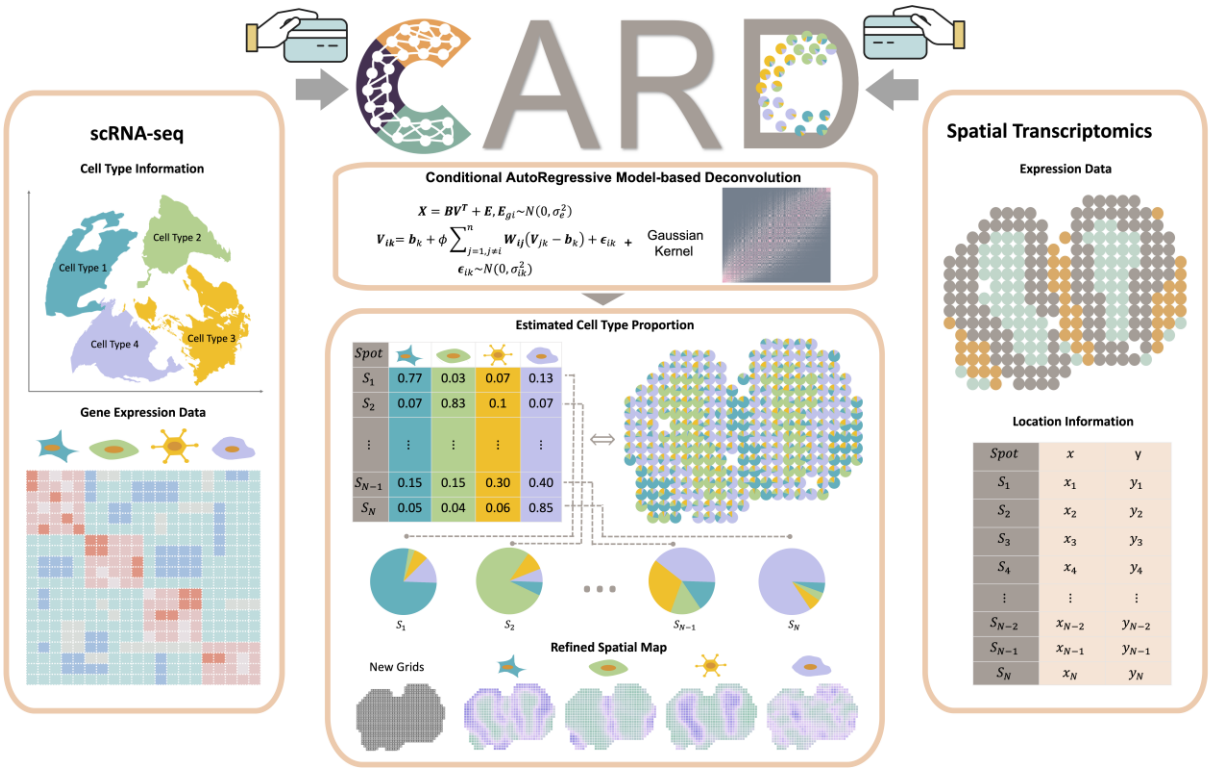


Figure 3.1 Schematic overview of CARD

CARD is designed to deconvolute spatial transcriptomics data and infer cell type composition on each spatial location based on the reference scRNA-seq data. CARD requires a scRNA-seq data with cell type specific gene expression information (left box) along with the spatial transcriptomics data with localization information (right box). With these two inputs, CARD performs deconvolution through a non-negative matrix factorization framework and outputs the estimated

cell type composition across spatial locations (bottom box). A unique feature of CARD is its ability to account for the spatial correlation of cell type compositions across spatial locations through a conditional autoregressive (CAR) model (top box). By accounting for the spatial correlation of cell type compositions across spatial locations, CARD is also capable of imputing cell type compositions and gene expression levels on locations not measured in the original study, facilitating the construction of a refined high-resolution spatial map on the tissue (bottom box).

We first explored a baseline analysis scenario (scenario I), where we used the same scRNA-seq data used in the simulations for deconvolution. Here, CARD outperforms all other deconvolution methods across all simulation settings (median RMSE = 0.079), with 9%, 8%, 33%, 7%, 23%, and 18% improvement in terms of RMSE as compared to MuSiC (0.087), RCTD (0.086), SPOTlight (0.118), cell2location (0.085), spatialDWLS (0.103), and stereoscope (0.096), respectively (**Figure 3.2 scenario I, Figure S3.3 - Figure S3.4**). In addition, CARD identifies the dominant cell type on each spatial location accurately as measured by AUC and ARI (**Figure S3.5**).

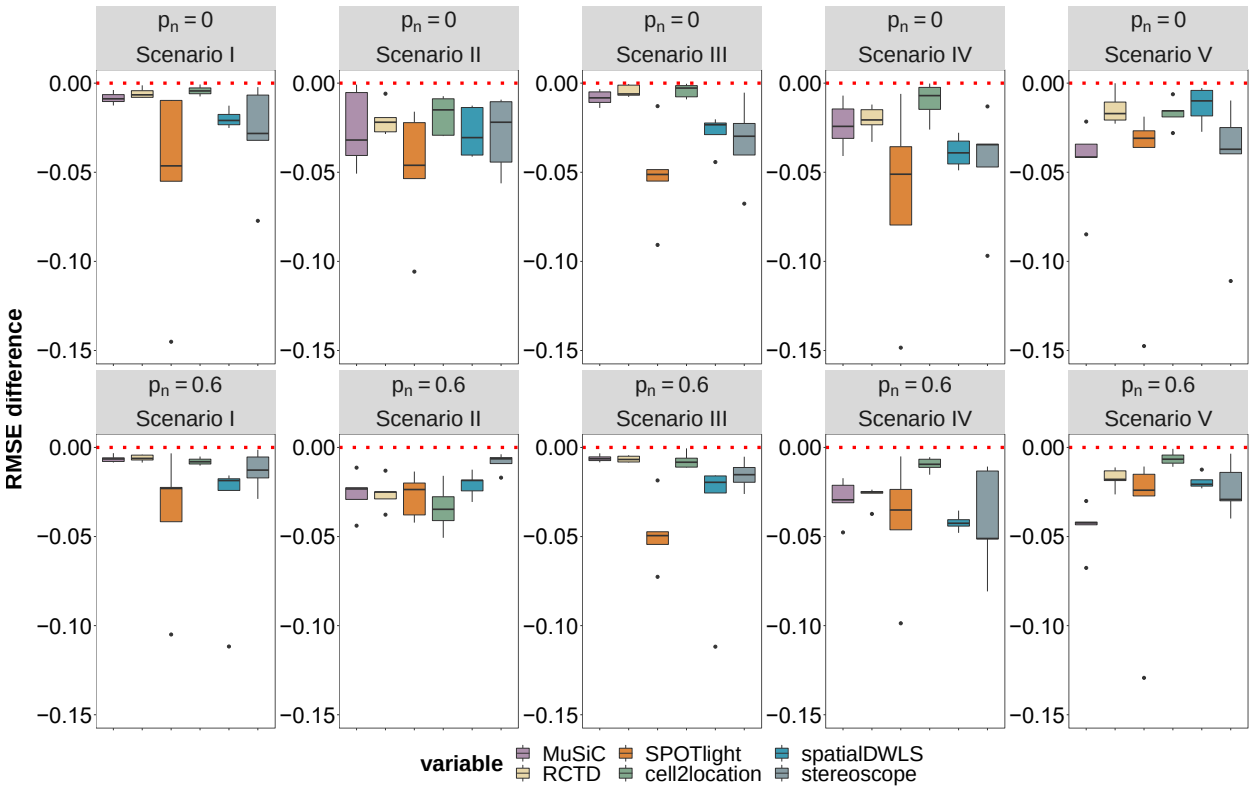


Figure 3.2 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenarios I-V.

In the analysis scenario I, the same scRNA-seq dataset used in simulations is used as the reference for deconvolution. In the analysis scenario II, the same scRNA-seq data but with one missing cell type (e.g., Neuron cells) is used as the reference for deconvolution. In the analysis scenario III, the same scRNA-seq data but with one additional cell type (e.g., Blood cells) is used as the reference for deconvolution. In the analysis scenario IV, the same scRNA-seq reference data but with miss-classified cell type in the reference for deconvolution. In the analysis scenario V, the different scRNA-seq reference sequenced from a different platform but with similar cell types is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (yellow), SPOTlight (orange), cell2location (green), spatialDWLS (blue), and stereoscope (blue gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. We further contrasted RMSE of the other methods with respect to that of CARD by computing an RMSE difference to remove the unnecessarily difficulty level variation across replicates. An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. Differences of RMSE across five simulation replicates (n = 5) were displayed in the form of box plots. Each boxplot ranges from the third and first quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

To examine the robustness of different deconvolution methods, we explored four additional scenarios (**APPENDIX** B.2) where we either removed one cell type in the scRNA-seq reference (scenario II); added one cell type (scenario III); used miss-classified cell types (scenario IV); or used another scRNA-seq data sequenced on a different platform for deconvolution (scenario V). Compared to scenario I, the performance of all methods remains similar in scenarios III (except SPOTlight) and generally reduces in other scenarios, though their relative rank remains largely consistent across scenarios. In addition, CARD outperforms the other methods in all settings, with its performance gain more apparent than scenario I (**Figure 3.2**). Specifically, in scenario II, CARD loses a median of 3% accuracy across settings as compared to using the original scRNA-seq data (**APPENDIX B.3**). However, CARD is more accurate than the other methods across settings with 13% ~ 32% accuracy improvement (**Figure 3.2, Figure S3.6**). In scenario III, CARD only loses a median of 0.4% accuracy across settings as compared to using the original scRNA-

seq data. It remains the most accurate method across settings with 7% ~ 40% accuracy improvement over the other methods (**Figure 3.2, Figure S3.7**). In scenario IV, CARD loses a median of 4% accuracy as compared to using the original scRNA-seq data (**Figure 3.2**). However, CARD is again more accurate than the other methods across settings (**Figure 3.2, Figure S3.8**), with 6% ~ 32% accuracy improvement across misclassified cell types (**Figure S3.9**). In scenario V, CARD loses a median of 10% accuracy across settings as compared to using the original scRNA-seq data. But it remains the most accurate method across settings with 5% ~ 35% accuracy improvement over the other methods (**Figure 3.2, Figure S3.10**).

We examined the deconvolution accuracy of different methods at distinct cell type resolution levels (**APPENDIX B.2**) and found that the deconvolution accuracy of most methods improved initially with increasing number of sub-cell types (**Figure S3.11**) and reached a saturation point with sufficiently large number of sub-cell types, where many sub-cell types are no longer distinguishable from each other. Regardless of the cell type resolution, the relative performance of most deconvolution methods remains consistent (Ma and Zhou 2022). We also carried out additional model-based simulations where we can more effectively control for spatial correlation and found as expected that the advantage of CARD over the other methods shows a clear dependency on spatial correlation (details in (Ma and Zhou 2022)).

3.3.2 Mouse olfactory bulb data

We applied CARD and the other methods to analyze four published spatial transcriptomics data that include two obtained from Spatial Transcriptomics (ST), one from Slide-seq, and one from 10x Visium (details in **APPENDIX B.4**). In each data, the majority of marker genes (92% by Moran's I test and 54% by Geary's C test) display statistically significant spatial autocorrelation

(adjusted p-value < 0.05; **Table S3.1**), with the semivariance generally increasing with distance and the expression correlation between locations decreasing with distance, supporting cell type composition similarity between neighboring locations (Ma and Zhou 2022). We used scRNA-seq data from sequencing platforms different from the spatial transcriptomics for deconvolution.

We first examined the mouse olfactory bulb (MOB) data (Ståhl et al. 2016), where we used a scRNA-seq data (Tepe et al. 2018) from 10x Chromium on the same tissue for deconvolution (**Table S3.2**). The MOB data consists of four main anatomic layers organized in an inside out fashion annotated based on H&E staining: the granule cell layer (GCL), the mitral cell layer (MCL), the glomerular layer (GL), and the nerve layer (ONL) (**Figure 3.3A**). The cell type compositions inferred by CARD accurately depict such expected layered structure (Nagayama, Homma and Imamura 2014), as is evident by visualizing either the first principal component (PC1) of the estimated cell type composition matrix (**Figure S3.12**) or the inferred dominant cell types (**Figure 3.3B**). In contrast, MuSiC, SPOTlight, spatialDWLS, and stereoscope were unable to distinguish the three outer layers from each other, while RCTD was unable to clearly distinguish the nerve layer from the glomerular layer. RCTD, cell2location, and spatialDWLS showed a blurry boundary between GCL and MCL/GL on top of the tissue section, while cell2location could not clearly identify the boundaries between MCL and GL.

Careful examination of the cell type composition and corresponding cell type marker genes in different layers further confirm the accuracy of CARD deconvolution (**Figure 3.3C - Figure 3.3D**). For example, CARD distinguished correctly the adjacent MCL and GL, with distinct enrichment of mitral/tufted cells and periglomerular cells in the two layers, respectively, despite the similarity between these two cell types; while others cannot (**Figure S3.13**, (Ma and Zhou 2022)). We also observed that multiple cell types inferred by CARD show spatially co-localization

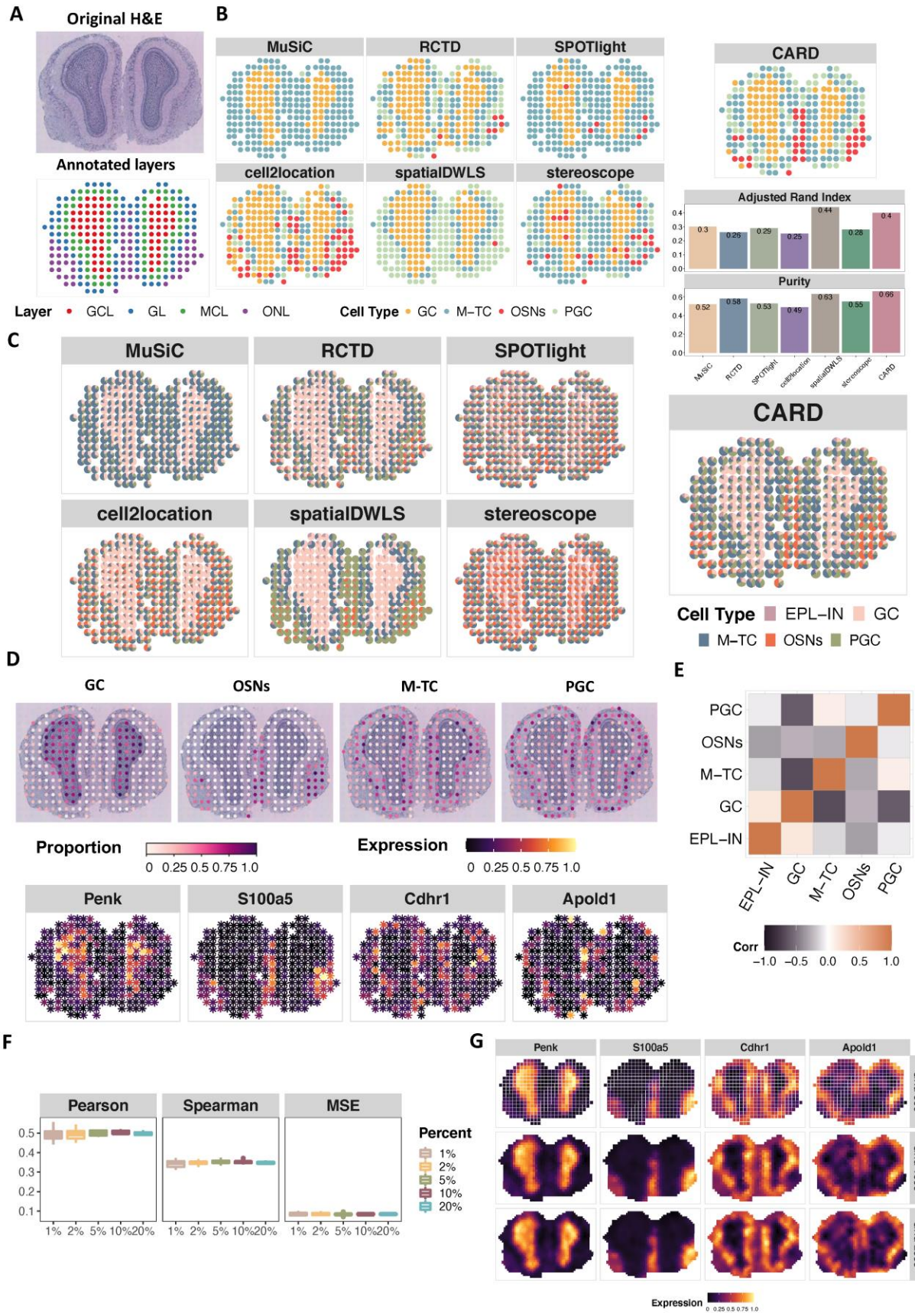

A
Original H&E
Annotated layers
Layer GCL GL MCL ONL
B
MuSiC
RCTD
SPOTlight
cell2location
spatialDWLS
stereoscope
CARD
Adjusted Rand Index
0.3 0.25 0.29 0.25 0.44 0.28 0.4
Purity
0.52 0.58 0.53 0.49 0.63 0.55 0.66
MuSiC RCTD SPOTlight cell2location spatialDWLS stereoscope CARD
Cell Type GC M–TC OSNs PGC
C
MuSiC
RCTD
SPOTlight
cell2location
spatialDWLS
stereoscope
CARD
Cell Type EPL–IN GC M–TC OSNs PGC
D
GC
OSNs
M-TC
PGC
Proportion 0 0.25 0.5 0.75 1.0
Expression 0 0.25 0.5 0.75 1.0
Penk
S100a5
Cdhr1
Apold1
E
PGC
OSNs
M–TC
GC
EPL–IN
EPL–IN GC M–TC OSNs PGC
Corr −1.0 −0.5 0.0 0.5 1.0
F
Pearson
Spearman
MSE
0.5
0.4
0.3
0.2
0.1
1% 2% 5% 10% 20%
Percent
1%
2%
5%
10%
20%
G
Penk
S100a5
Cdhr1
Apold1
Grid 500
Grid 1000
Grid 2000
Expression 0 0.25 0.5 0.75 1.0

Figure 3.3 Analyzing the mouse olfactory bulb data.

(A) Hematoxylin and eosin (H&E) staining of the olfactory bulb (top panel) displays four anatomic layers that are organized in an inside out fashion (bottom panel): the granule cell layer (GCL), the mitral cell layer (MCL), the glomerular layer (GL), and the nerve layer (ONL). **(B)** Left panel shows on each spatial location the dominant cell type inferred from four different deconvolution methods. The examined cell types include granule cells (GC), olfactory sensory neurons (OSNs), periglomerular cells (PGC) and mitral/tufted cells (M-TC). Compared deconvolution methods include MuSiC, RCTD, SPOTlight, cell2location, spatialDWLS, stereoscope and CARD. Right bottom panel displays the adjusted rand index (ARI; y-axis) and the Purity (y-axis), which quantify the similarity between the inferred dominant cell types from different methods (x-axis) and the anatomic layers annotated based on the H&E image. **(C)** Spatial scatter pie plot displays inferred cell type composition on each spatial location from different deconvolution methods. **(D)** Top panels display on each spatial location the proportion of each of the four cell types inferred by CARD. Bottom panels display the expression levels of four corresponding cell type specific marker genes. **(E)** Correlations in cell type proportion across spatial locations between pairs of cell types inferred by CARD. Color is scaled by the correlation value. **(F)** Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10). A fixed percentage of locations are masked as missing (x-axis) and CARD is used to impute the gene expression on the masked locations. Three different metrics (y-axis) are used to evaluate imputation accuracy in terms of the similarity between the imputed expression and true expression on masked locations: Pearson's correlation, Spearman's correlation and mean square error (MSE). Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box. **(G)** CARD imputes gene expression for four marker genes on a fine grid set of spatial locations (number of grid points = 500, 1,000, or 2,000), resulting in a refined spatial map of gene expression.

patterns (**Figure 3.3E**).

A key benefit of CARD is its ability to model the spatial correlation structure across spatial locations, which facilitates the imputation of cell type composition and gene expression on locations not measured in the original study. We performed location masking analysis for CARD and validated that the imputed expression levels are highly consistent with the truth regardless of the percentage of masked locations (Pearson's correlation=0.44-0.56; **Figure 3.3F**, **Figure S3.14**). Imputation on new locations allows us to construct a refined spatial map of cell type composition or gene expression with arbitrarily high spatial resolution (details in **Methods**), which captures fine grained details of the layered structure in the olfactory bulb (**Figure 3.3G**, **Figure S3.15** - **Figure S3.16**) and facilitates the identification of marker genes with spatial expression patterns

(**Figure S3.17**, (Ma and Zhou 2022)). In contrast, the fixed resolution enhancement by BayesSpace failed to capture the expected spatial expression pattern for a few marker genes at high resolution (Ma and Zhou 2022). We quantitatively compared the performance of CARD and BayesSpace for resolution enhancement by performing clustering analysis on the imputed expression data. We found that the clustering results based on CARD displayed a clear inside-out layered structure that resembles the anatomic organization of the olfactory bulb, more so than that obtained with the original scale data or by BayesSpace (**Figure S3.18**). CARD is also computationally efficient: CARD takes only 0.4 seconds to construct the refined expression map for all genes, is 5,816 times faster than BayesSpace, and represents a scalable solution for fine map reconstruction in much larger datasets.

3.3.3 Human pancreatic ductal adenocarcinomas data

The second data we examined is a human pancreatic ductal adenocarcinomas (PDAC) data from spatial transcriptomics (Moncada et al. 2020). For deconvolution, we first used a matched scRNA-seq data for the same patient obtained through inDrop (Moncada et al. 2020) (denoted as PDAC-A). The PDAC data contains multiple tissue regions (cancer, pancreatic, ductal, and stroma regions) annotated by histologists based on H&E staining (Moncada et al. 2020) (**Figure 3.4A**). Through deconvolution, CARD located various pancreatic and tumoral cell types into different tissue regions (**Figure 3.4B**). The PC1 of the estimated cell type composition matrix from CARD can clearly capture a gross regional segregation between cancer and non-cancer regions, between the ductal and stroma regions, and between the pancreatic and ductal regions. In contrast, none of the other methods were as effective in differentiating these regions (**Figure S3.19** - **Figure S3.20**). The dominant cell types on each location from CARD also capture the segregation between cancer and non-cancer regions (**Figure S3.21**), with the neoplastic cells such as cancer clone A and clone

B cells highly enriched in the former (Wilcoxon test p-value = 1.9e-48, 1.1e-43 respectively, **Figure 3.4D**). CARD also reveals distinct distribution of two macrophage subpopulations between the cancer and non-cancer regions (**Figure 3.4D**), representing a key functional signature of the regional compartmentalization of the cancer tissue that were missed by the other methods (**Figure S3.22**)

CARD further divides the cancer region into two sub-regions, a pattern missed by the other methods: an upper subregion dominated by cancer clone A cells with an enrichment of marker gene *Tm4sf1*, and a bottom subregion dominated by cancer clone B cells with an enrichment of marker gene *S100a4* (**Figure 3.4B - Figure 3.4C**, **Figure S3.23**). *S100A4* is a prognostic marker for early-stage pancreatic cancer and its spatial enrichment suggests that the bottom cancer subregion is likely an early cancer region. In contrast, *Tm4sf1* is essential for PDAC migration and invasion (Zheng et al. 2015, Fu et al. 2020, Xu et al. 2020) and its spatial enrichment suggests that the upper cancer subregion is likely a late-stage cancer region with metastasis capability. Indeed, the upper cancer subregion is also detected by CARD to be enriched with fibroblast cells, along with fibroblast cell marker gene Cd248 (**Figure 3.4C**), a cell type known to be associated with advanced TNM stage (Zhang et al. 2017).

CARD also localizes many other cell types into specific tissue regions, consistent with the expression pattern of the corresponding marker genes (**Figure 3.4C**). In contrast, none of the other methods capture the expected spatial localization of both ductal centroacinar and terminal ductal cells. In addition, acinar cells inferred by CARD are enriched in the normal pancreatic tissue region; but they are inferred by the other methods to be either absent in the pancreatic region or diffused outward from the pancreatic region to the stroma region and cancer region. Several cell types inferred by CARD are also co-localized spatially in PDAC (**Figure 3.4F**), such as those

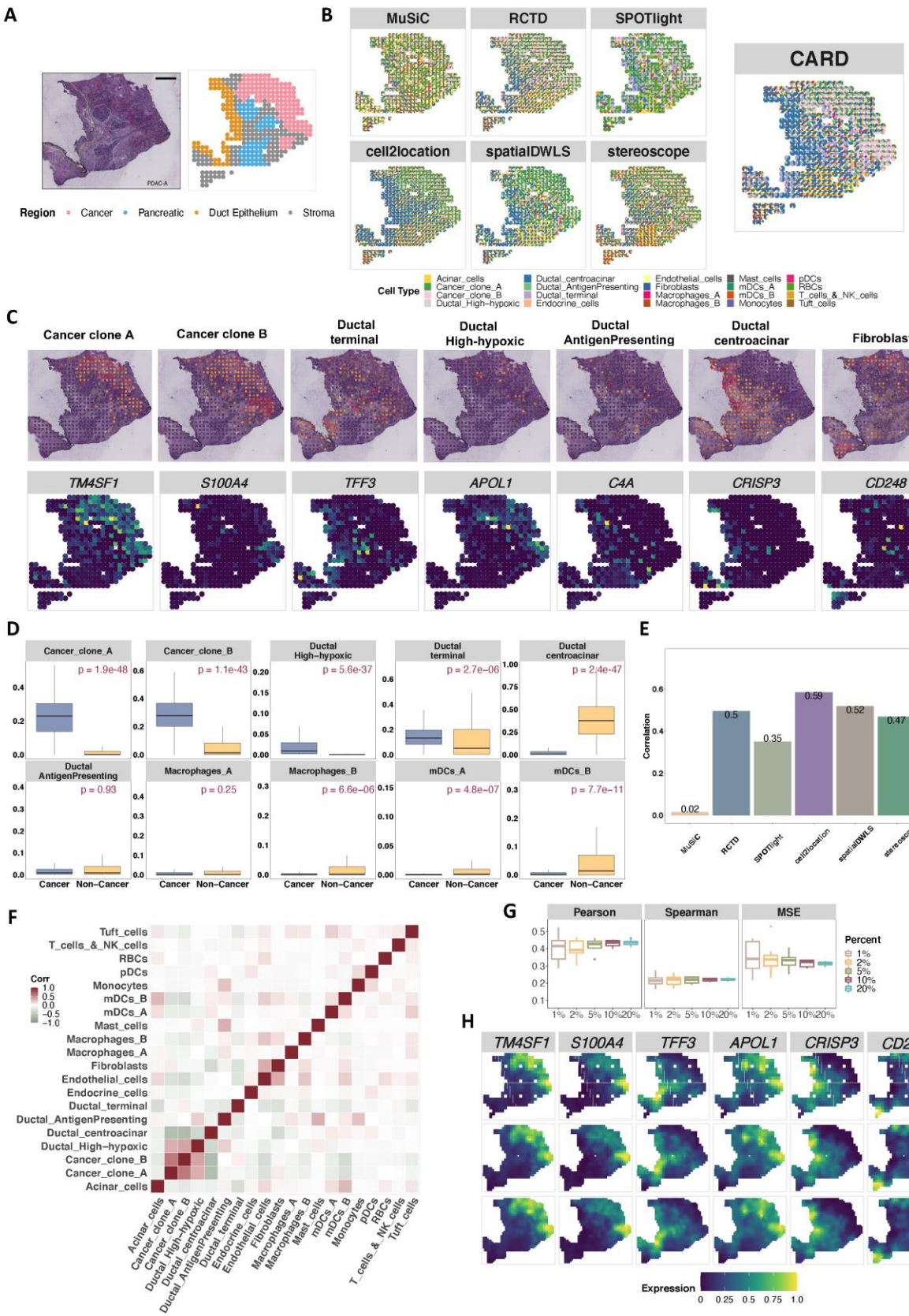

A
PDAC-A
Region Cancer Pancreatic Duct Epithelium Stroma
B
MuSiC
RCTD
SPOTlight
CARD
cell2location
spatialDWLS
stereoscope
Cell Type
Acinar_cells
Cancer_clone_A
Cancer_clone_B
Ductal_High−hypoxic
Ductal_centroacinar
Ductal_AntigenPresenting
Ductal_terminal
Endocrine_cells
Endothelial_cells
Fibroblasts
Macrophages_A
Macrophages_B
Mast_cells
mDCs_A
mDCs_B
Monocytes
pDCs
RBCs
T_cells_&_NK_cells
Tuft_cells
C
Cancer clone A
Cancer clone B
Ductal terminal
Ductal High-hypoxic
Ductal AntigenPresenting
Ductal centroacinar
Fibroblasts
TM4SF1
S100A4
TFF3
APOL1
C4A
CRISP3
CD248
D
Cancer_clone_A
p = 1.9e-48
Cancer_clone_B
p = 1.1e-43
Ductal High−hypoxic
p = 5.6e-37
Ductal terminal
p = 2.7e−06
Ductal centroacinar
p = 2.4e-47
Ductal AntigenPresenting
p = 0.93
Macrophages_A
p = 0.25
Macrophages_B
p = 6.6e−06
mDCs_A
p = 4.8e−07
mDCs_B
p = 7.7e−11
Cancer
Non−Cancer
E
Correlation
0.02
0.5
0.35
0.59
0.52
0.47
0.75
MuSiC
RCTD
SPOTlight
cell2location
spatialDWLS
stereoscope
CARD
F
Corr
Tuft_cells
T_cells_&_NK_cells
RBCs
pDCs
Monocytes
mDCs_B
mDCs_A
Mast_cells
Macrophages_B
Macrophages_A
Fibroblasts
Endothelial_cells
Endocrine_cells
Ductal_terminal
Ductal_AntigenPresenting
Ductal_centroacinar
Ductal_High−hypoxic
Cancer_clone_B
Cancer_clone_A
Acinar_cells
G
Pearson
Spearman
MSE
Percent
1%
2%
5%
10%
20%
H
TM4SF1
S100A4
TFF3
APOL1
CRISP3
CD248
Grid 500
Grid 1000
Grid 2000
Expression
0 0.25 0.5 0.75 1.0

Figure 3.4 Analyzing the pancreatic ductal adenocarcinoma (PDAC) data.

(A) Hematoxylin and eosin (H&E) staining of the PDAC (left panel) displays four regions (right panel) annotated from the original publication (Moncada et al. 2020): Cancer, Pancreatic, Duct and stroma regions. **(B)** Spatial scatter pie plot displays inferred cell type composition on each spatial location from different deconvolution methods. Compared deconvolution methods include MuSiC, RCTD, SPOTlight, cell2location, spatialDWLS, stereoscope and CARD. **(C)** Top panels display on each spatial location the proportion of each of the cell types inferred by CARD. Bottom panels display the expression levels of corresponding cell type specific marker genes. **(D)** Comparisons of cell type proportions inferred by CARD in cancer region (n = 137) vs non-cancer region (n = 289) with p-value tested by two-sided Wilcoxon Rank Sum test. **(E)** Correlation between mean cell type proportions inferred by CARD and that in the matched scRNA-seq reference data. **(F)** Correlations in cell type proportion across spatial locations between pairs of cell types inferred by CARD. Color is scaled by the correlation value. **(G)** Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10). A fixed percentage of locations are masked as missing (x-axis), and CARD is used to impute the gene expression on the masked locations. Three different metrics (y-axis) are used to evaluate imputation accuracy in terms of the similarity between the imputed expression and true expression on masked locations: Pearson's correlation, Spearman's correlation and mean square error (MSE). **(H)** CARD imputes gene expression for four marker genes on a fine grid set of spatial locations (number of grid points = 500, 1,000, or 2,000), resulting in a refined spatial map of gene expression. Each boxplot in (D) and (G) ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

between ductal high hypoxic cells and cancer cells and those between endothelial cells and fibroblast cells, supporting the role of the former in forming the hypoxic and nutrient-poor tumor microenvironment (TME) and the role of the later in pancreatic-cancer stroma interaction of the tumor microenvironment (Nielsen, Mortensen and Detlefsen 2016, Morvaridi et al. 2015). The mean cell type proportions inferred by CARD in the ST data are also highly correlated with that measured in the scRNA-seq dataset obtained on the same patient, more so than that obtained by the other methods (**Figure 3.4E**).

Next, we examined the robustness of deconvolution by using unmatched scRNA-seq datasets (**Table S3.2**). Despite the platform and sample differences in the scRNA-seq references, we found that the estimated cell type compositions for the major cell types are consistent across different scRNA-seq references, with the highest consistency achieved by CARD (**Figure S3.24**).

Regardless of which unmatched scRNA-seq data was used, CARD shows superior performance than the other methods in capturing the gross segregation of cancer and non-cancer regions, identifying two distinct cancer subregions, accurately localizing cell types, and revealing a possible TME supporting tumor progression (Zheng et al. 2017a, Comito et al. 2020, Lambrechts et al. 2018, Junya et al. 2019) (Details see the ref (Ma and Zhou 2022)).

Finally, we found that the imputed gene expression by CARD are highly consistent with the truth across a range of masking percentages (Pearson's correlation=0.29-0.52; **Figure 3.4G**, **Figure S3.25**). Such consistency is higher when the matched scRNA-seq data from the same patient is used as the reference, as compared to using an unmatched scRNA-seq data (**Figure S3.26**). The high-resolution spatial map of cell type composition or gene expression obtained by CARD also reveals refined boundaries between different tissue subregions (**Figure S3.27**) and the spatial expression pattern of marker genes (**Figure 3.4H**, **Figure S3.28**). Besides marker genes, CARD also discovered multiple genes that display clear spatial expression pattern in the refined spatial map but not in the original map (**Figure S3.29**). In contrast, the high-resolution map of BayesSpace does not show a clear pattern of multiple known marker genes and additional genes (Ma and Zhou 2022). Clustering analysis on CARD imputed high resolution data also revealed clear segregation of the two cancer sub-regions, the normal pancreatic region, and the ductal region, more so than the original data or the refined data by BayesSpace (**Figure S3.29**).

3.3.4 Mouse hippocampus data from multiple sources

We analyzed two mouse hippocampus datasets: one directly on hippocampus measured using Slide-seq V2 (Stickels et al. 2021) and the other on a coronal brain section containing hippocampus measured using 10x Visium. We used the hippocampus scRNA-seq dataset by Drop-

seq (Cable et al. 2021, Saunders et al. 2018) for deconvoluting both datasets (**Table S3.2**). We only applied cell2location to the 10x Visium data but not the Slide-seq V2 data due to its heavy computational burden.

The hippocampus primarily consists of three regions -- the CA1/CA2 region, the CA3 region, and the dentate gyrus -- all visualizable by total UMI counts per location displayed on the tissue (**Figure 3.5A**). The cell type compositions inferred by CARD accurately depict the three anatomic structures of hippocampus, with the compositional PC1 capturing the curved shape of hippocampus accurately, more so than the other three methods (**Figure 3.5A**, **Figure S3.31**). The dominant cell type on each location inferred by CARD also matches the expectation (**Figure 3.5B**): CA1 cells are highly enriched in CA1; CA3 cells mainly localize in CA3; dentate cells reside in a C-shaped ring region of dentate gyrus; ependymal cells form an irregular and columnar shape and line the ventricles of the brain (Del Bigio 2009); while choroid cells reside right below the ependymal cells and locate in the choroid plexus (Ramachandran 2002) along with Cajal-Retzius cells (Meyer 2010) (**Figure S3.32**). In contrast, MuSiC is unable to localize the main cell types such as CA1 and CA3 cells correctly and thus unable to reveal the main structures of the hippocampus. SPOTlight detects an incorrectly diffused pattern of ependymal cells and incorrectly locates many CA3 cells to the CA1 region or outside hippocampus. RCTD, spatialDWLS, and stereoscope perform similarly, all locating CA3 cells incorrectly in CA1 (**Figure 3.5B**; (Ma and Zhou 2022)), with the CA1 cell marker gene enriched in locations dominated by CA3 cells inferred by these methods (**Figure S3.33**). Additionally, they all allocate different cell types to hippocampus structures that appear to be much wider than expected (Rodriques et al. 2019, Stickels et al. 2021, Hawrylycz et al. 2014b) (**Figure 3.5A - Figure 3.5B**). Careful examination of marker genes further confirms the accuracy of CARD deconvolution (**Figure 3.5C**). We quantified

the deconvolution performance of different methods by examining the expression levels of the marker genes on each of the three hippocampal structures inferred based on the estimated cell type composition by different methods. Quantifications again support more accurate deconvolution by CARD than the other methods (**Figure 3.5D, Figure S3.34**).

We observed that multiple cell types inferred by CARD are co-localized together (**Figure S3.35**). The highest co-localization occurs between Slc17a6/Vglut2 neurons and entorhinal cells, highlighting the cell compositional architecture underlying the hippocampus-entorhinal cortex network (Wozny et al. 2018). The imputed gene expression by CARD are consistent with the truth across a range of masking percentages (**Figure S3.36**). Although the resolution of this dataset is already high, the refined spatial map of cell type composition by CARD again reveals refined boundaries between different subregions of hippocampus (**Figure S3.37**), with the refined gene expression recovers strong spatial pattern for various marker genes (**Figure 3.5E**, **Figure S3.38**) and additional genes (**Figure S3.39**). We examined the reliability of the refined spatial map by creating a low-resolution version of the Slide-seq V2 data and then applied CARD to construct a refined spatial expression map at the original Slide-seq V2 resolution. We found that the refined spatial map recovers a consistent and sometime stronger spatial pattern than the original Slide-seq V2 data (Ma and Zhou 2022), supporting the accuracy and effectiveness of refined spatial map construction. Here, we were unable to apply BayesSpace due to both its heavy computational burden and its required input of pixel coordinates that are not available from Slide-seq technologies.

Finally, we examined the hippocampus region from the 10x Visium data. Again, CARD captures the key structures of the hippocampus (**Figure 3.5F - Figure 3.5G**). The estimated cell type compositions on CA1, CA3 and dentate gyrus from both CARD and MuSiC matched the

corresponding structures on the H&E image, while those from the other methods appear to also occupy regions outside the expected structure boundaries (**Figure 3.5F**, **Figure S3.40**), a pattern confirmed with quantifications (**Figure S3.41**).

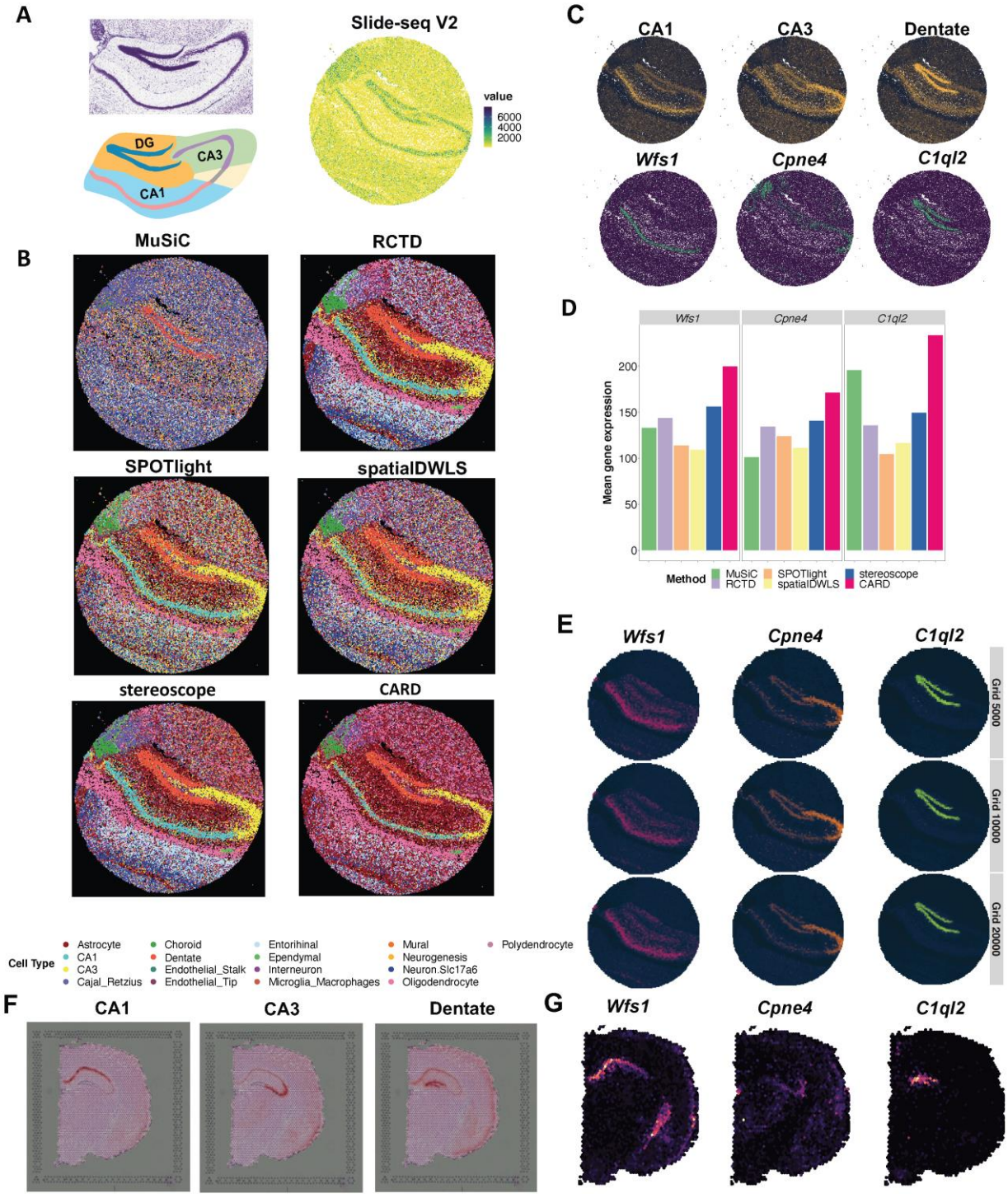


Figure 3.5 Analyzing the hippocampus region in the Slide-seq V2 and 10x Visium Mouse Brain (Coronal) data.

(A) The UMI counts of Slide-seq V2 data (right panel) displays the structure and the shape of hippocampus tissue, highly consistent with the image from Allen Reference Atlas (left panel) **(B)** The dominant cell type on each location inferred from four different deconvolution methods. Compared deconvolution methods include MuSiC, RCTD, SPOTlight, spatialDWLS, stereoscope and CARD. **(C)** Top panels display on each spatial location the proportion of each of the cell types inferred by CARD. Bottom panels display the expression levels of corresponding cell type specific marker genes. The examined cell types are CA1 cells, CA3 cells and dentate cells. **(D)** Bar plots display the comparisons of the mean gene expression level of marker genes in the major regions inferred by different deconvolution methods; **(E)** CARD imputes gene expression for four marker genes on a fine grid set of spatial locations, resulting in a refined spatial map of gene expression. **(F)** The proportion of each of the cell types on each location inferred by CARD in the 10x Visium dataset. **(G)** The expression levels of corresponding cell type specific marker genes in the 10x Visium dataset.

3.3.5 Extension of CARD for reference-free deconvolution

We further developed CARDfree, an extension of CARD for reference-free cell type deconvolution that does not require scRNA-seq reference data. The technical details of CARDfree is provided in the ref (Ma and Zhou 2022). CARDfree only requires users to input a list of gene names for previously known cell type markers, which determines the dimensionality of the input gene expression matrix. Compared to CARD, CARDfree yields generally similar cell type composition estimates in the real data, but likely with lower accuracy. For example, CARDfree captures the general tissue domain segregation pattern as CARD in both MOB and PDAC data, though it was unable to differentiate the two cancer sub-regions as CARD did (**Figure S3.42**). CARDfree does not perform as well as CARD in the high-resolution Slide-seq V2 data and did not identify the CA3 structure based on its estimated cell type proportions, as the Slide-seq V2 data is highly sparse and thus could be benefited from reference-based deconvolution. However, in the hippocampus region of the Slide-seq V2 data, we did notice that CARDfree identified a region with a unique cell type composition ((Ma and Zhou 2022), CT15 colored in blue) that was not found by other deconvolution methods. This region appears to part of the entorhinal cortex,

which consists of endothelial tip cells that are highly related to angiogenesis in mouse brain (Wälchli et al. 2014). The results suggest that reference-free deconvolution may sometimes have added benefits.

3.4 Discussion

We have presented CARD for accurate and spatially informed deconvolution of spatial transcriptomics. CARD is computationally efficient: it is 0.8-7,761.8 times faster while using 0.2%-109% of the physical memory as compared to the other deconvolution methods (**Figure S3.43**); it is 5,875-7,028 times faster and uses only 14%-17% of the physical memory as compared to BayesSpace in creating refined spatial maps (**Figure S3.44 - Figure S3.45**). We have demonstrated the benefits of CARD in both simulations and applications to four spatial transcriptomics datasets.

We have primarily focused on examining the sequencing-based technologies that measure the average gene expression from a mixture of cells on each tissue location. Non-sequencing-based technologies, such as seqFISH (Lubeck et al. 2014) and MERFISH (Chen et al. 2015), mostly rely on single molecular fluorescent *in situ* hybridization (smFISH) and are directly of single cell resolution. However, it remains computationally challenging to detect the accurate boundaries between cells on the smFISH image data, especially when the cell density is high (Moffitt et al. 2018, He et al. 2021, Moen et al. 2019). Consequently, the expression data measured on each "single cell" in smFISH may consist of transcripts from a mixture of neighboring cells. Therefore, CARD can also be applied to analyze these datasets. In a mouse cortex data from seqFISH+ (Eng et al. 2019), we found that the cell type compositions inferred by CARD clearly displayed a layered structure that resembled the laminar organization of the cortex, with each layer harboring a distinct composition of neuronal populations (Ma and Zhou 2022).

We have presented an extension of CARD, CARDfree, for reference-free deconvolution. CARDfree requires a post-processing step to correctly label the inferred cell types. Such post-processing often requires cell type specific gene expression profiles and can be challenging to carry out accurately. For example, in PDAC, CARDfree infers cell type composition on each location for 20 inferred cell types. But it is not trivial to find the name for each inferred cell type: for instance, it is not easy to tell whether the inferred cell type #14 (CT14) corresponds to the ductal centroacinar cell or the endothelial cell, as markers for both cell types are enriched in locations with a high proportion of CT14 cells (Ma and Zhou 2022). Therefore, new computational algorithms are likely needed for labeling cell types inferred from reference-free deconvolution methods. We also present another extension of CARD (Ma and Zhou 2022) to facilitate the construction of single-cell resolution spatial transcriptomics from non-single-cell resolution spatial transcriptomics. Such extension requires knowing the spatial localization information for all single cells on the tissue, which remains challenging to obtain from non-single-cell resolution spatial transcriptomics. Because the spatial transcriptomics data itself does not contain information for inferring the single cell positions, H&E image segmentation is often required to identify single cells on the tissue and extract their locations. However, common software is not always accurate in inferring the location for single cells (Ma and Zhou 2022). In addition, aligning H&E image with spatial transcriptomics can be computationally challenging (Bergenstråhle, Larsson and Lundeberg 2020). Future efforts are needed to address these challenges.

Additional extensions of CARD are possible. First, CARD models normalized spatial transcriptomes data and could be benefited from extensions for direct modeling of raw count data using an over-dispersed Poisson model (Sun et al. 2017, Sun et al. 2018). Second, we only explored the use of the Gaussian kernel (Sun, Zhu and Zhou 2020a) for modeling spatial correlation.

Exploring the use of other kernels such as the periodic kernels or incorporating histological image information such as image intensity level as additional coordinates (Hu et al. 2021, Fu et al. 2021), which can be readily done in CARD, may capture diverse and rich spatial correlation patterns in the future. Third, the spatial imputation feature of CARD facilitates not only the construction of a refined spatial map but also the selection of scRNA-seq references when multiple scRNA-seq resources are available. Specifically, we can evaluate through data masking the imputation accuracy resulted from pairing with different scRNA-seq references and select the scRNA-seq data with the best imputation accuracy for deconvolution. In PDAC, the matched scRNA-seq indeed produced the best imputation performance and would be selected as the optimal reference data for deconvolution.

3.5 Methods

3.5.1 CARD method overview

We present an overview of CARD here, with its technical details provided in **APPENDIX B.5**. CARD is a deconvolution method for spatial transcriptomics studies with regional resolution. These studies perform transcriptomic profiling on multiple tissue locations, each of which contains multiple single cells. CARD aims to estimate the cell type composition on each tissue location while properly accounting for the spatial correlation among them. CARD requires both spatial transcriptomics data and a single cell RNA-seq (scRNA-seq) data as input. The scRNA-seq data serves as a reference and consists of K cell types with a set of G cell type informative genes. Cell types and informative genes in scRNA-seq can be obtained through standard analysis pipelines for clustering and informative gene identification (Soneson and Robinson 2018, Duò, Robinson and Soneson 2018). In scRNA-seq, we denote $\boldsymbol{B}$ as the G by K cell type specific expression matrix for the informative genes, where each element represents the mean expression level of an informative

gene in a specific cell type. The expression matrix $\boldsymbol{B}$ is commonly referred to as the reference basis matrix. In the spatial transcriptomics data, we denote $\boldsymbol{X}$ as the G by N gene expression matrix for the same set of informative genes measured on N spatial locations. We denote $\boldsymbol{V}$ as the N by K cell type composition matrix, where each row of $\boldsymbol{V}$ represents the proportions of the K cell types on each spatial location. Our objective is to estimate $\boldsymbol{V}$ given both $\boldsymbol{X}$ from the spatial transcriptomics data and $\boldsymbol{B}$ constructed from the scRNA-seq data. To do so, we consider a non-negative matrix factorization model to link the three matrices:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{V}^{\boldsymbol{T}} + \boldsymbol{E}, \tag{B.1}$$

where each element in $\boldsymbol{V}$ is constrained to be non-negative; and $\boldsymbol{E}$ is an G by N residual error matrix with each element independently and identically following a normal distribution $\boldsymbol{E}_{gi} \sim N(0, \sigma_e^2)$. A detailed biological interpretation of equation (B.1) in the context of deconvolution is provided in **APPENDIX B.5**.

The non-negative matrix factorization model in equation (B.1) has been applied for cell type deconvolution in bulk RNA-seq studies. However, this model is not directly applicable for deconvoluting spatial transcriptomics as it does not account for the spatial correlation structure in the cell type compositions across locations. Intuitively, cell type compositions on two neighboring locations of a tissue are likely to be similar to each other, more so than those on locations that are far away. Consequently, the cell type compositions on neighboring locations contain valuable information for inferring the cell type composition on the location of interest. The similarity in cell type compositions on neighboring locations effectively induces spatial correlation among rows of $\boldsymbol{V}$ in the above factorization model. Thus, modeling spatial correlation in $\boldsymbol{V}$ is relevant for spatial transcriptomics as it would allow us to borrow cell type composition information across spatial locations to enable accurate estimation of $\boldsymbol{V}$. To accommodate the spatial correlation in $\boldsymbol{V}$, we

specify a conditional autoregressive (CAR) (De Oliveira 2010, Besag 1974, Banerjee et al. 2014) modeling assumption on each column of $\boldsymbol{V}$. Specifically, for the column/cell type *k*, we assume:

$$\boldsymbol{V}_{ik} = \boldsymbol{b}_k + \phi \sum_{j=1, j\neq i}^{n} \boldsymbol{W}_{ij}(\boldsymbol{V}_{jk} - \boldsymbol{b}_k) + \boldsymbol{\epsilon}_{ik} \quad \text{(B.2)}$$

where $\boldsymbol{V}_{ik}$ represents the proportion of cell type k on the i-th location; $\boldsymbol{b}_k$ is the k-th cell type specific intercept that represents the average cell type composition across locations; W is a N-by-N non-negative weight matrix with each element $\boldsymbol{W}_{ij}$ specifying the weight used for inferring the cell type composition on the i-th location based on the cell type composition information on the j-th location; ϕ is a spatial autocorrelation parameter that determines the strength of the spatial correlation in cell type composition; and $\boldsymbol{\epsilon}_{ik}$ is the residual error that follows a normal distribution $\boldsymbol{\epsilon}_{ik} \sim N(0, \sigma_{ik}^2)$. The CAR modeling assumption on V effectively expresses the composition of the k-th cell type on the i-th location, $\boldsymbol{V}_{ik}$, as a weighted summation of the k-th cell type compositions on all other locations, $\boldsymbol{V}_{jk}$ $(j \neq i)$. Consequently, the CAR modeling assumption on V allows us to borrow information across locations to infer the cell type composition on the location of interest.

We follow (Sun et al. 2020a) to express the weight matrix $\boldsymbol{W}$ in the form of a Gaussian kernel function constructed based on the Euclidean distance between pairs of spatial locations (details in **APPENDIX B.5**). The Gaussian kernel function has been widely used to model a range of correlation patterns that decay over distance across tissue locations in many other analytic tasks in spatial transcriptomics (Vanhatalo, Pietiläinen and Vehtari 2010, Rousset and Ferdy 2014). While we primarily focus on using a Gaussian kernel for $\boldsymbol{W}$, our method and software can easily incorporate other types of kernels to capture diverse spatial correlation patterns encountered in different data sets. With the Gaussian kernel matrix $\boldsymbol{W}$, we further obtain a row-standardized weight matrix $\widetilde{\boldsymbol{W}}$ through transformation $\widetilde{\boldsymbol{W}}_{ij} = \boldsymbol{W}_{ij}/\boldsymbol{W}_{i+}$, with $\boldsymbol{W}_{i+} = \sum_{j=1}^{n} \boldsymbol{W}_{ij}$. Because the

weight matrix and the residual error variance need to satisfy the symmetric condition (Cressie 1992, Rue and Held 2005), we set $\sigma_{ik}^2 = \boldsymbol{\lambda}_k \,/\, \boldsymbol{W}_{i+}$ to ensure $\widetilde{\boldsymbol{W}}_{ij}\sigma_{jk}^2 = \widetilde{\boldsymbol{W}}_{ji}\sigma_{ik}^2$, where $\boldsymbol{\lambda}_k$ is a scalar. With the above parameterization, we can follow the Brook's Lemma (Brook 1964, Besag 1974) to obtain the joint distribution for the *N*-size column vector $\boldsymbol{V}_k$ as

$$\boldsymbol{V}_k \sim MVN(\boldsymbol{b}_k \mathbf{1}_N, \boldsymbol{\Sigma_k}), \tag{B.3}$$

where $\mathbf{1}_N$ is a *N*-vector of 1's; $\boldsymbol{\Sigma_k} = \left(\boldsymbol{I}_N - \phi\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{M_k}$ is a positive definite covariance matrix with $\boldsymbol{M_k} = diag\left(\sigma_{1\mathrm{k}}^2, \ldots \sigma_{Nk}^2\right)$; and *MVN* denotes a multivariate normal distribution (details in **APPENDIX B.5**).

Equations (B.1) and (B.3) together define a factor model with a CAR modeling assumption on the latent factors to induce spatial correlation across rows of $\boldsymbol{V}$. By modeling the spatial correlation in $\boldsymbol{V}$, our model allows us to borrow cell type composition information across spatial locations for spatially informed cell type deconvolution. We developed a constrained optimization algorithm in the maximum likelihood framework to estimate the cell type composition matrix $\boldsymbol{V}$, with non-negativity constraints on each of its elements (details in **APPENDIX B.5**). Our algorithm treats the hyper-parameters ($\boldsymbol{b}_k, \boldsymbol{\lambda}_{\mathrm{k}}$, ϕ and σ_e^2) as unknown and infers these parameters based on the data at hand to ensure optimal deconvolution performance. Our algorithm has several computational advantages that make it highly computationally efficient. First, the modeling framework of CARD is in essence a linear factor model, expressing the mean gene expression profile in the spatial transcriptomics as a linear function of that from scRNA-seq. The linear factor modeling framework streamlines the inference procedure and facilitates scalable computation. Second, CARD makes use of the fast multiplicative updating rules (Lee and Seung 2000, Janecek and Tan 2011) for updating the nonnegative cell type composition matrix in each optimization

iteration. The multiplicative updating rules allow for algorithmic optimization without explicit inverse of the spatial covariance matrix, which is otherwise required for spatial deconvolution, and which incurs heavy computation burden (**APPENDIX B.5**). Third, CARD takes advantage of the modern computing architecture and explicitly expresses the most computationally intensive part of the algorithm in the form of large matrix operations instead of multiple scalar operations. For example, it updates the entire cell type composition matrix jointly at each optimization iteration instead of updating each element in the cell type composition matrix on each spatial location separately. Finally, while CARD is implemented in R, its core deconvolution algorithm is implemented with an efficient C++ code that is linked back to the main functions of CARD through Rcpp, ensuring scalable computation.

3.5.2 Imputation and construction of high-resolution spatial maps for cell type composition and gene expression

A key feature of CARD is its ability to model the spatial correlation structure in V. By modeling the spatial correlation in V, CARD can predict and impute the cell type compositions on new, unmeasured, spatial locations on the tissue. Imputing cell type compositions on new locations would allow us to obtain a refined cell type composition map of the tissue with a spatial resolution much higher than that measured in the original study. To enable imputation and construction of a refined cell type composition map, we first outlined the shape of the tissue by applying a two-dimensional concave hull algorithm (Park and Oh 2012) on the existing locations. We then created an equally spaced grid within the tissue outline and set the number of grid points to exceed the number of spatial locations measured in the original study. We denote the cell type composition matrix on the original N spatial locations as V and denote the corresponding matrix on the $\boldsymbol{N}^*$ new locations as $\boldsymbol{V}^*$. Based on equation (B.3), the $(\boldsymbol{N}+\boldsymbol{N}^*)$-sized cell type composition vector for the

k-th cell type, $(\boldsymbol{V_k}, \boldsymbol{V_k^*})^{\boldsymbol{T}}$, follows a multivariate normal distribution $\boldsymbol{MVN}(\boldsymbol{b_k}\mathbf{1}_{\boldsymbol{N+N^*}}, \boldsymbol{\Sigma})$. We partition the covariance matrix $\boldsymbol{\Sigma}$ into $\begin{bmatrix} \boldsymbol{\Sigma_{oo}} & \boldsymbol{\Sigma_{on}} \\ \boldsymbol{\Sigma_{no}} & \boldsymbol{\Sigma_{nn}} \end{bmatrix}$, where o are the indices that correspond to the original locations while n are the indices that correspond to the new locations. We can then estimate $\boldsymbol{V_k^*}$ via its conditional mean

$$\widehat{\mathrm{V}}_{\mathrm{k}}^{*} = \mathrm{b_k}1_{\mathrm{N}^*} + \Sigma_{\mathrm{no}}\Sigma_{\mathrm{oo}}^{-1}(\mathrm{V_k} - \mathrm{b_k}1_{\mathrm{N}}), \tag{B.4}$$

where the parameters on the right-hand side of the equation are replaced by the corresponding estimates. The estimates $\widehat{\boldsymbol{V}}_{\boldsymbol{k}}^{*}$ on the new locations are almost always non-negative as they are effectively represented as a weighted summation of the non-negative cell type proportions on the original locations. To ensure scalable imputation, we used a sparse approximation of the covariance matrix $\boldsymbol{\Sigma}$ by using only the nearest 10 neighbors for each location. With the imputed cell type compositions, we can further impute the gene expression levels on the new locations by multiplying the above conditional mean in equation (B.4) with the basis matrix to obtain $\boldsymbol{B}\,\widehat{\boldsymbol{V}}_{\boldsymbol{k}}^{*}$.

3.5.3 Basis matrix construction

We constructed the reference basis matrix $\boldsymbol{B}$ following the main ideas of MuSiC (Wang et al. 2019) using three detailed steps (details in **APPENDIX B.5**). (1) We selected genes that are expressed in both the scRNA-seq reference data and the spatial transcriptomic data. (2) We selected among them the candidate cell type informative genes with a mean expression level in a given cell type at least 1.25 log fold higher than its mean expression level across all remaining cell types. (3) We removed among them the outlier genes that show high expression heterogeneity within a cell type by calculating gene-specific expression dispersion (Ma and Zhou 2022). In particular, we calculated the expression dispersion as the variance to mean ratio for each gene in each cell type. We then obtained the gene-specific dispersion by averaging the estimated

expression dispersion across cell types. We finally removed the top 1% genes with the largest gene-specific dispersion values.

3.5.4 Simulations and deconvolution analysis evaluation

All simulations are described in the **APPENDIX B.2**. In each simulation replicate, we calculated the true cell type proportions on each spatial location as the number of cells in each cell type divided by the total number of cells on the location. We denote the true cell type composition matrix as $\boldsymbol{V}$. After we obtained the estimated cell type composition matrix $\widehat{\boldsymbol{V}}$, we evaluated deconvolution performance by computing the root mean square error (RMSE) between $\widehat{\boldsymbol{V}}$ and $\boldsymbol{V}$ through

$$RMSE = \sqrt{\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\boldsymbol{V}_{ik} - \widehat{\boldsymbol{V}}_{ik}\right)^2},$$

where N=260 is the total number of spatial locations and K is the total number of cell types. Note that the above formula for RMSE calculation is based on all cell types (**APPENDIX B.3**).

3.5.5 Compared methods

We compared CARD with four deconvolution methods: (1) MuSiC (Wang et al. 2019) (version 0.1.1), (2) SPOTlight (Elosua-Bayes et al. 2021) (version 0.1.0), and (3) RCTD (Cable et al. 2021) (version 1.1.0), (4) cell2location (Kleshchevnikov et al. 2022) (version 0.07a), (5) spatialDWLS (Dong and Yuan 2021)(implemented in the R package Giotto, version 1.0.4), (6) stereoscope (Andersson et al. 2020)(version 0.2.0). For all methods, we followed the tutorial on the corresponding GitHub pages and used the recommended default parameter settings for deconvolution analysis. cell2location requires users to input additional parameters. For these

parameters, we set them to be close to what we used in the simulations and to be close to what we best know of in the real data applications. Specifically, in the simulations, we set "cells_per_spot" to be a random number from a uniform distribution U (8, 12) with an expected value of 10. We set "factors_per_spot" and "combs_per_spot" to be exactly the number of cell types available in the corresponding scRNA-seq reference. In the real data applications, we set "cells_per_spot" to be 30 for the mouse olfactory spatial transcriptomics data and human pancreatic ductal adenocarcinoma data and set it to be 10 for the 10x Visium data. We set both the "factors_per_spot" and "combs_per_spot" to be 7 following the software tutorial.

We also compared the high-resolution spatial map constructed by CARD with a recently developed method BayesSpace (version 1.1.4). Because BayesSpace only implements that neighborhood structure suitable for Spatial Transcriptomics (ST) and 10x Visium, we only evaluated its performance on the mouse olfactory spatial transcriptomics data and human pancreatic ductal adenocarcinoma (PDAC) data. We followed the tutorial on GitHub and used the recommended default parameter settings for resolution enhancement. Specifically, we set the required number of clusters qs based on their recommended pseudo-log-likelihood as the following: qs = 5 for mouse olfactory spatial transcriptomics data and qs = 8 for PDAC data. Note that BayesSpace is restricted in creating a neighborhood structure that has a fixed number of sub-spots at each location in the original data (5 for Visium technology and 9 for ST technology). In order to compare the high-resolution spatial gene expression constructed by CARD and BayesSpace on the same set of sub-spots, we applied CARD to directly impute the gene expression on the sub-spots generated by BayesSpace. Afterwards, we performed PCA dimension reduction on the high-resolution data and applied the K-means algorithm analysis on the top 20 PCs to cluster

spatial locations into six clusters for the mouse olfactory data and eighteen clusters for the PDAC following the original studies.

3.5.6 Real data analyses

All real datasets used in the present study are described in the **APPENDIX B.4**. We first examined cell type composition similarity in these real datasets. Because we do not know the true cell type composition in these data, we used cell type marker genes as surrogates to examine the spatial distribution of cell types on the tissue (Ralston and Shaw 2008). We reasoned that, if the cell type composition is similar among neighboring locations, then we would also expect the cell type maker genes to show spatial correlation in their expression pattern on the tissue. Therefore, for each of the three spatial transcriptomics datasets examined in the present study, we looked at one marker at a time (from the same set of markers in real data applications) and examined its spatial autocorrelation pattern by carrying out spatial autocorrelation tests using Moran I (Li, Calder and Cressie 2007) and Geary's C (Radeloff et al. 2000). Note that we were unable to carry out Moran's I test (Bivand et al. 2011) and Geary's C test (Bivand et al. 2011)on the large SlideseqV2 dataset due to heavy computational cost. Besides examining cell type marker genes, we also calculated correlation in the expression profile of the marker genes between neighboring locations (Ma and Zhou 2022). Intuitively, if the cell type composition is similar between neighboring locations, then the expression profile of marker genes will also be correlated between neighboring locations, more so than that between locations that are far away.

Next, we applied different methods to deconvolute the above datasets. In each analysis, we supplied the same spatial transcriptomics data and the same scRNA-seq data as input for all methods (preprocessing details in **APPENDIX B.4**). After deconvolution, we followed

(Teschendorff et al. 2020) to assign the dominant cell type on each spatial location and examined the distribution of each cell type on the tissue. For the two datasets that contain a matched H&E image (MOB and PDAC), we compared the distribution of the dominant cell types inferred from spatial transcriptomics with the tissue structures annotated based on the H&E image. Specifically, we obtained tissue structure annotations based on the H&E image, overlayed spatial transcriptomics locations on top the H&E image, and manually annotated each measured location in spatial transcriptomics with the tissue structure annotations extracted from the H&E image. For the MOB dataset, we annotated four main structural layers in the olfactory bulb: the granule cell layer (GCL, which contains n = 67 spatial locations), the mitral cell layer (MCL, n = 75), the glomerular layer (GL, n =80), and the nerve layer (ONL, n = 55). For the PDAC dataset, we annotated four main structural regions on the cancer tissue: cancer region (n = 137), ductal region (n = 72), pancreatic region (n = 70), and stroma region (n = 147). In the MOB dataset, because each olfactory layer is dominated by one cell type, we directly compared the dominant cell type inferred from CARD with the layer annotations based on H&E image via adjusted rand index (ARI) and Purity, using the *compare* function in the *igraph* R package (v1.0.0) and *purity* function in the *funtimes* R packages (v8.1), respectively. In the PDAC dataset, because each tissue region is substantially more heterogenous than that in the MOB data and contains potentially multiple cell types, using ARI would penalize methods that detected fine tissue regions that were not detected in the original study. Therefore, we carefully examined the distribution of inferred cell types on each annotated tissue region based on the transcriptomic profile and existing biological literature.

Because CARD directly models spatial correlation, CARD can be used to impute gene expression on unmeasured locations. To evaluate the accuracy of such imputation, we performed location masking analysis. Specifically, in each real data application, we randomly masked a fixed

percentage of the spatial locations to be missing, used the unmasked spatial locations to perform CARD deconvolution, relied on the cell type composition estimates obtained on the unmasked locations to predict and impute the cell type composition on the masked locations, and further imputed the gene expression levels on the masked locations. We then compared the imputed gene expression level with the measured expression level on the masked locations using RMSE. RMSE serves as an indicator on how accurate CARD imputation works, which also reflects its deconvolution performance. The magnitude of RMSE can vary substantially across datasets depending on factors such as the sequencing read depth per location. In the analysis, we set the mask percentage to be either 1%, 2%, 5%, 10% or 20% for all datasets.

3.5.7 Data and code availability

This study made use of publicly available datasets. These include the mouse olfactory bulb dataset (https://www.spatialresearch.org/resources-published-datasets/doi-10-1126science-aaf2403/), human pancreatic ductal adenocarcinoma (PDAC) dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111672), mouse hippocampus Slide-seqV2 dataset (https://singlecell.broadinstitute.org/single_cell/study/SCP948/robust-decomposition-of-cell-type-mixtures-in-spatial-transcriptomics), and mouse brain (coronal section) 10x Visium (https://www.10xgenomics.com/resources/datasets/). For the scRNAseq references used in this study, they are all publicly available with details provided in supplementary tables 2-3. The CARD software package and source code have been deposited at www.xzlab.org/software.html. All scripts used to reproduce all the analysis are also available at the same website.

3.6 Supplementary Figures

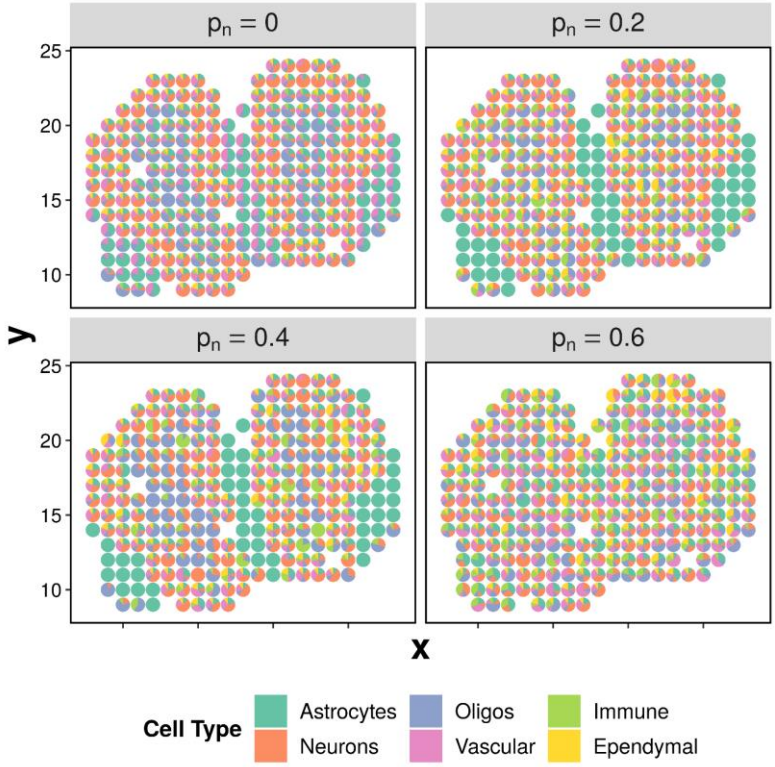


Figure S3.1 Simulated cell type proportions.

Simulated cell type proportions capture a wide variety of spatial patterns and display layer-structural patterns across spatial locations. Here, as an example of one simulation replicate, data were simulated under different spatial correlation strength as represented by the proportion of noisy locations (p_n, columns) with $n_c = 10$. High p_n corresponds to low spatial correlation. Each panel of the figure was plotted as a pie plot of cell type compositions with 6 cell types we used in the simulation.

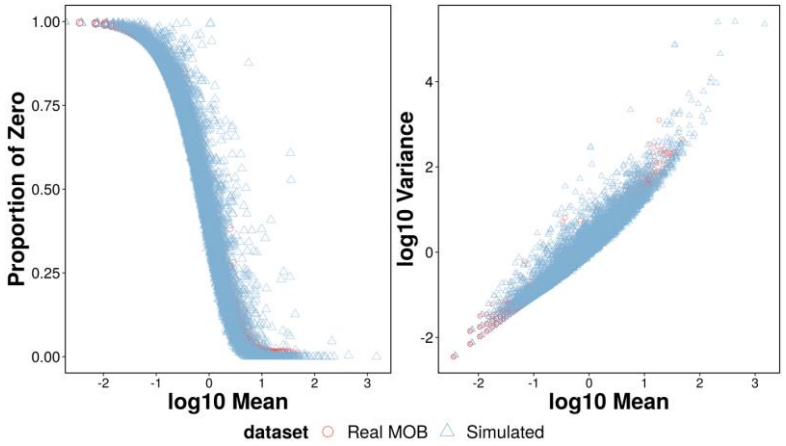


Figure S3.2 Simulated data are realistic, preserving data features observed in the published spatial transcriptomics data.

Specifically, as an example of one simulation replicate, the data was simulated under the following parameters setting: the number of cells on each spatial location (n_c = 10) and the spatial correlation strength as represented by the proportion of noisy locations (p_n = 0). (**A**) Proportion of zero versus mean under log10 scale for both simulated data (blue) and published spatial transcriptomics data (pink); (**B**) Mean-variance plot under log10 scale for both simulated data (blue) and published spatial transcriptomics data (red).

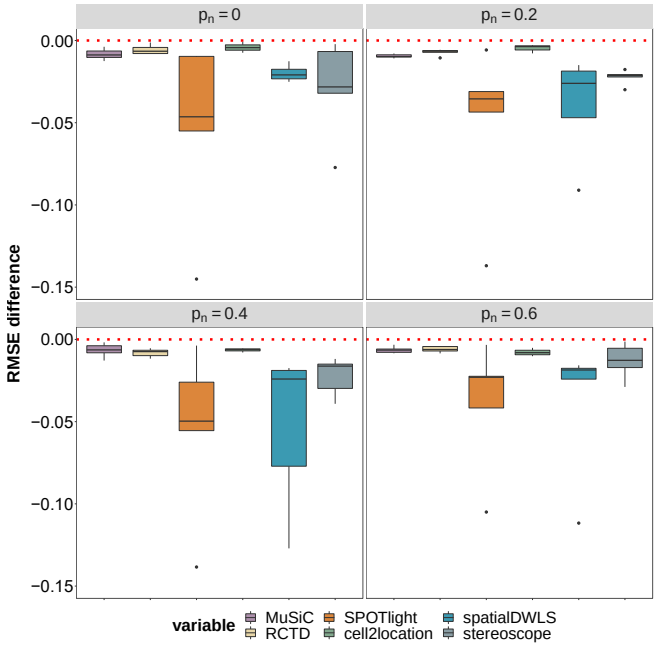


Figure S3.3 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario I.

In the analysis scenario I, the same scRNAseq dataset used in simulations is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (yellow), SPOTlight (orange), cell2location (green), spatialDWLS (blue), and stereoscope (blue-gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. We further contrasted RMSE of the other methods with respect to that of CARD by computing an RMSE difference. An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. **Differences of RMSE** across five simulation replicates (n = 5) were displayed in the form of box plots. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

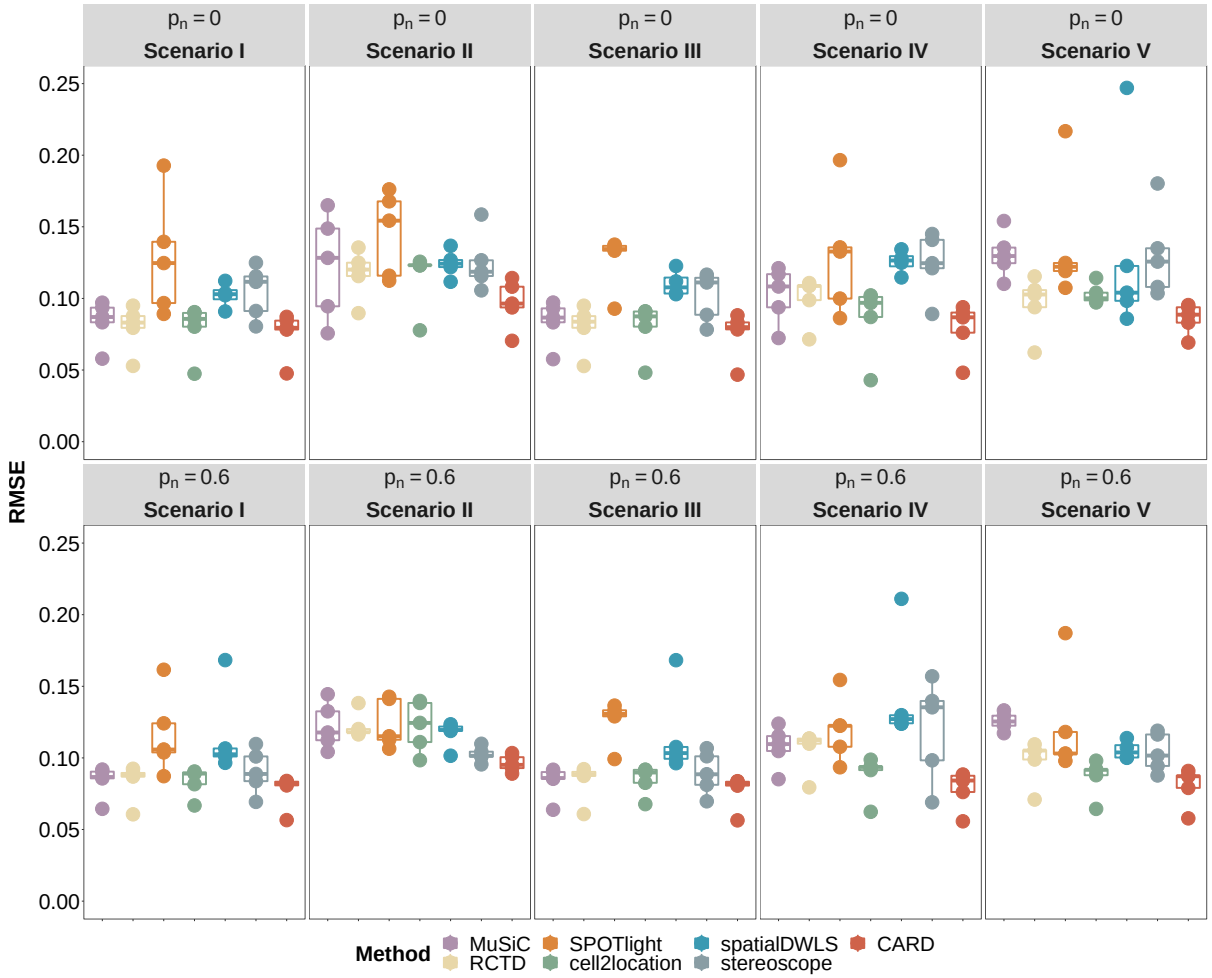


Figure S3.4 Comparison of deconvolution accuracy of different methods in simulations under all simulation scenarios.

In the analysis scenario I, the same scRNA-seq dataset used in simulations is used as the reference for deconvolution. In the analysis scenario II, the same scRNA-seq data but with one missing cell type (e.g., Neuron cells) is used as the reference for deconvolution. In the analysis scenario III, the same scRNA-seq data but with one additional cell type (e.g., Blood cells) is used as the reference for deconvolution. In the analysis scenario IV, the same scRNA-seq reference data but with miss-classified cell type in the reference for deconvolution. In the analysis scenario V, the different scRNA-seq reference sequenced from a different platform but with similar cell types is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (yellow), SPOTlight (orange), cell2location (green), spatialDWLS (blue), stereoscope (blue-gray), and CARD (red). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. **RMSE** across five simulation replicates (n = 5) were displayed in the form of box plots. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

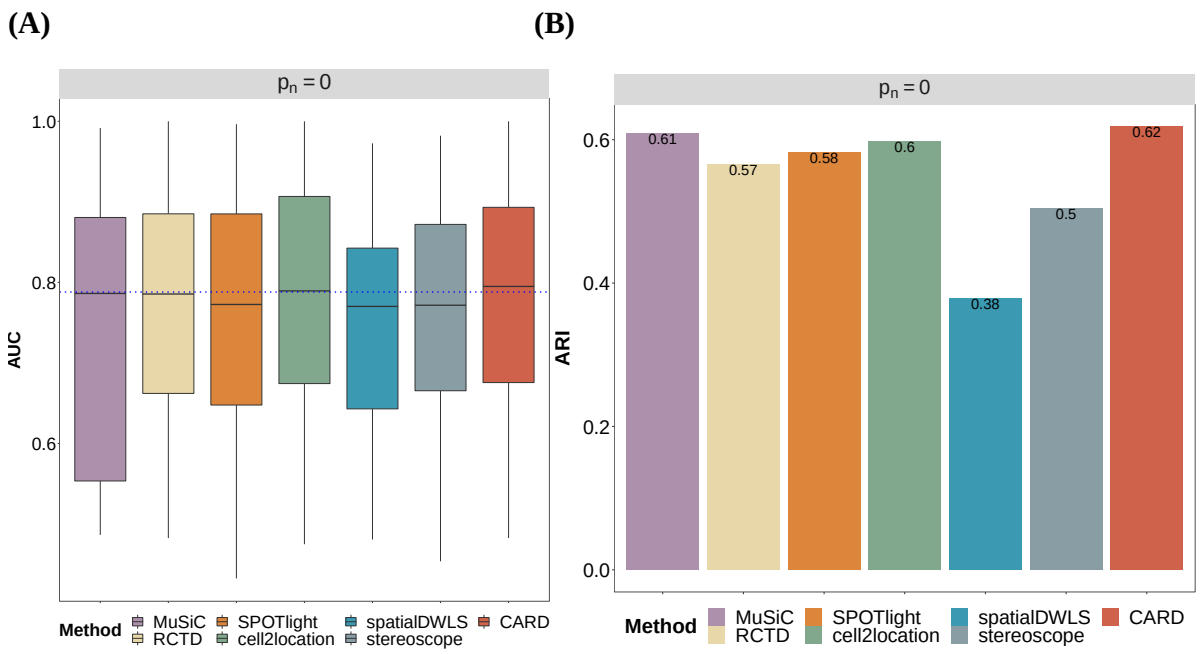


Figure S3.5 Deconvolution accuracy on detecting the dominant cell type at each spatial location for each method at Simulation Scenario I.

Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (wheat), SPOTlight (orange), cell2location (green), spatialDWLS (skyblue), stereoscope (blue gray), and CARD (red). Here the percentage of noisy locations equals to 0 ($P_n = 0$). **(A)** Boxplots display for each cell type (n = 6) the AUC between the binary labeled dominant cell type inferred by each method and the true binary dominant cell type for each spatial location. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box. **(B)** Bar plots display for the ARI between inferred dominant cell type and the true dominant cell type for each spatial location. A higher AUC and ARI indicate a higher accuracy at detecting the dominant cell types.

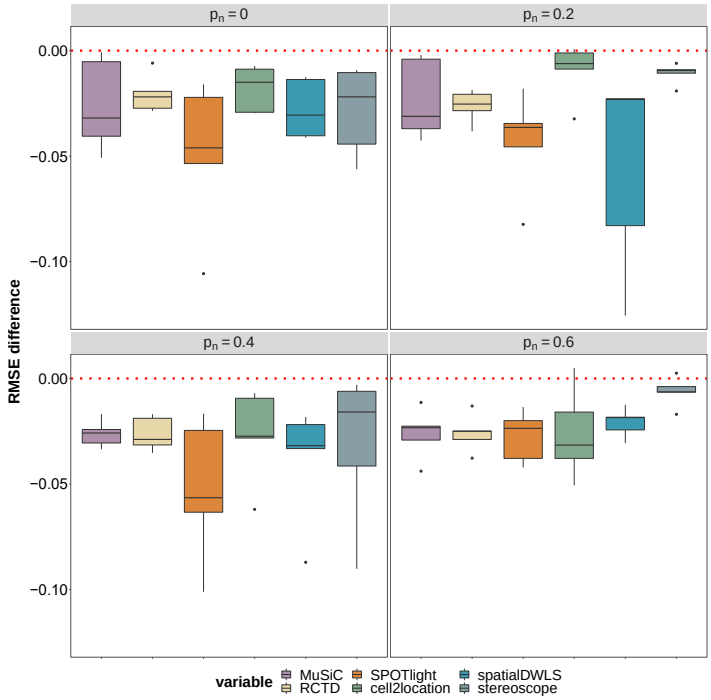


Figure S3.6 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario II.

In the analysis scenario II, the same scRNAseq dataset is used in simulations but missing one cell type (e.g., Neuron cells) is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (yellow), SPOTlight (orange), cell2location (green), spatialDWLS (blue), and stereoscope (blue-gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. We further contrasted RMSE of the other methods with respect to that of CARD by computing an RMSE difference. An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. Differences of RMSE across five simulation replicates (n = 5) were displayed in the form of box plots. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box. For results of other missing cell types, please see details in ref (Ma and Zhou 2022)

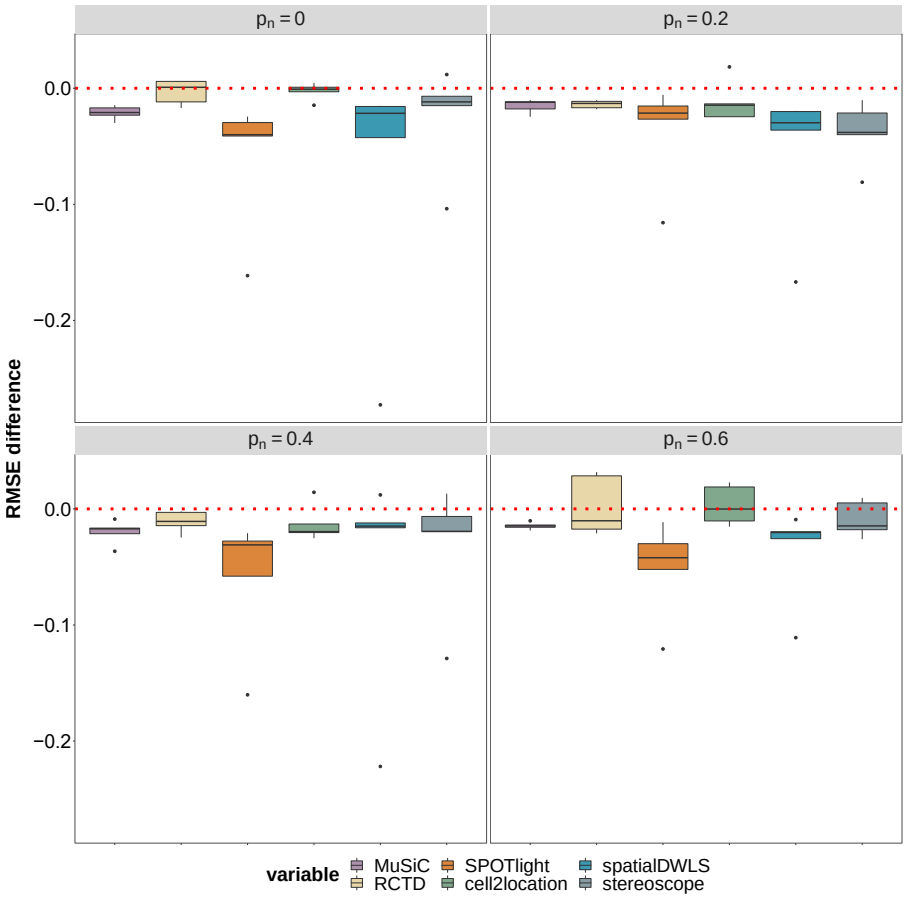


Figure S3.7 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario II.

In the analysis scenario II, the same scRNAseq dataset used in simulations but missing one cell type (e.g., Astrocytes) is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (yellow), SPOTlight (orange), cell2location (green), spatialDWLS (blue), and stereoscope (blue-gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. We further contrasted RMSE of the other methods with respect to that of CARD by computing an RMSE difference. An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. Differences of RMSE across five simulation replicates (n = 5) were displayed in the form of box plots. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

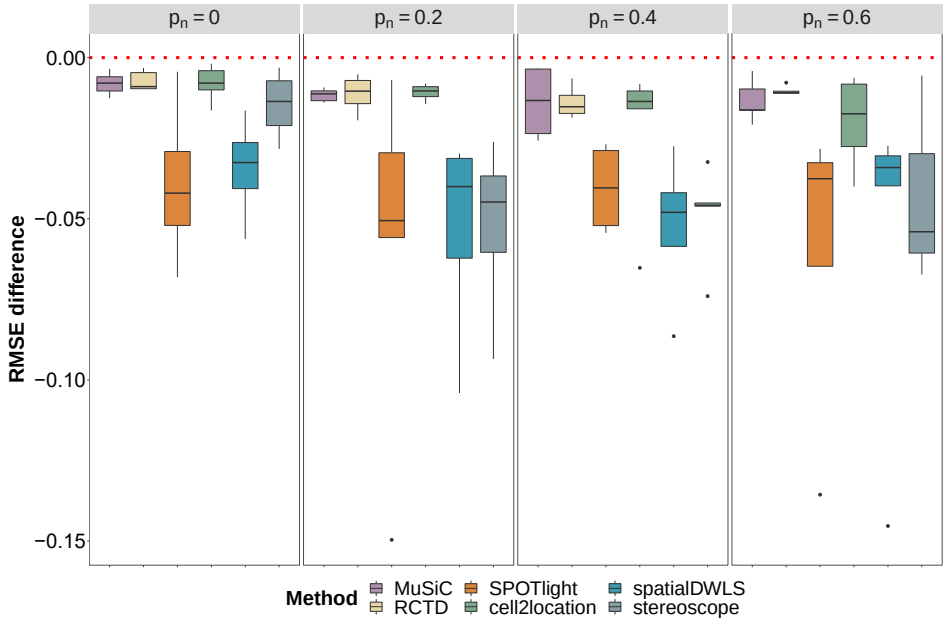


Figure S3.8 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario IV

In the analysis scenario IV, the same scRNA-seq data but with manually merged cell type included is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (wheat), SPOTlight (orange), cell2location (green), spatialDWLS (skyblue), and stereoscope (blue gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. RMSE difference with CARD were displayed across all five replicates (n = 5) in the format of boxplot when the two cell types underlying the merged one are (**A)** similar to each other (e.g., astrocytes and ependymal cells, with mean gene expression correlation between the two equaling 0.8); **(B)** or very different from each other (e.g., neurons and immune cells, with mean gene expression correlation between the two equaling 0.3). An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. Differences of RMSE across five simulation replicates (n = 5) were displayed in the form of box plots. In (A) and (B), each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

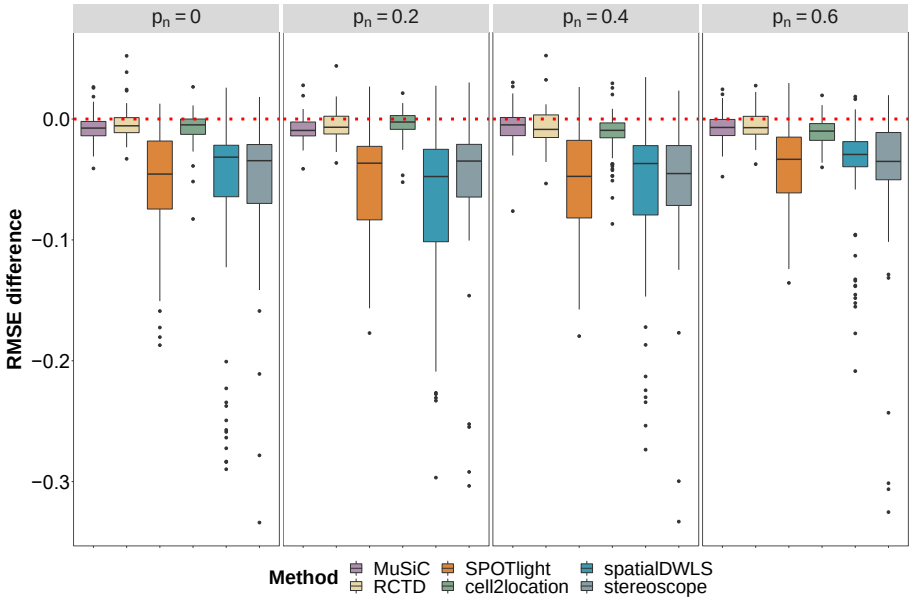


Figure S3.9 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario IV across all possible combinations of the merged cell types

In the analysis scenario IV, the same scRNA-seq data but with manually merged cell type included is used as the reference for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (wheat), SPOTlight (orange), cell2location (green), spatialDWLS (skyblue), and stereoscope (blue gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. Boxplot displaying the contrasted RMSE of the other methods with respect to that of CARD by computing an RMSE difference. An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. Differences of RMSE across five simulation replicates (n = 5) were displayed in the form of box plots. Here, the RMSE difference is displayed across all cell type combinations underlying the merged one. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

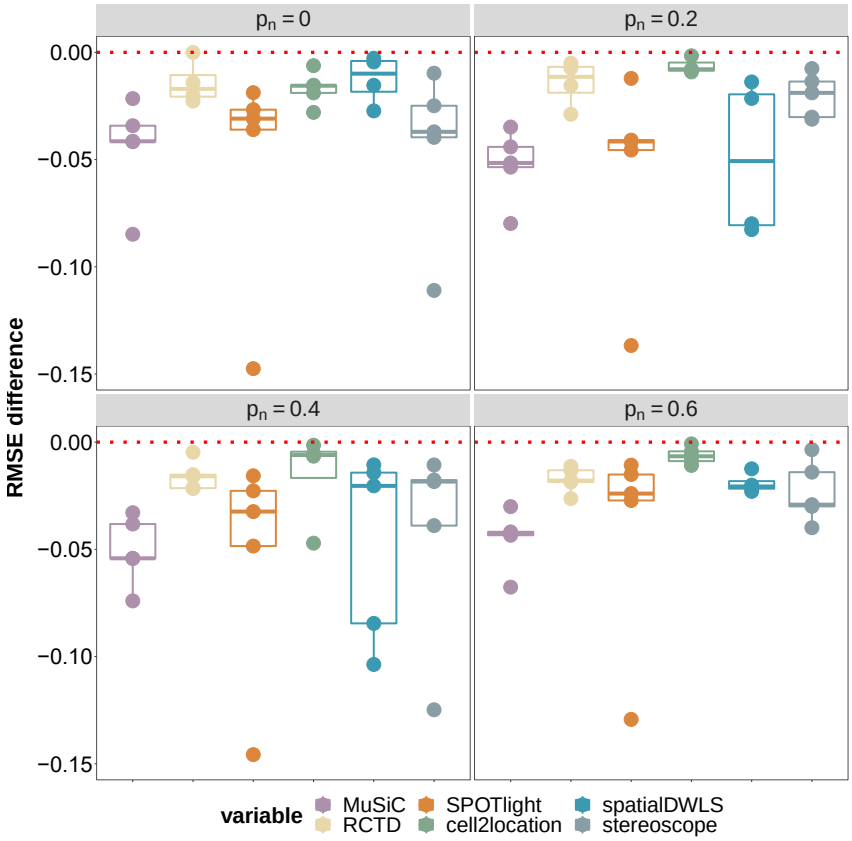


Figure S3.10 Comparison of deconvolution accuracy of different methods in simulations under the analysis scenario V.

In the analysis scenario V, a similar scRNA-seq data but sequenced from different platform was used for deconvolution. Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (wheat), SPOTlight (orange), cell2location (green), spatialDWLS (skyblue), and stereoscope (blue gray). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. Boxplot displaying the contrasted RMSE of the other methods with respect to that of CARD by computing an RMSE difference. An RMSE difference (y-axis) below zero suggests that CARD performs better than other methods. Differences of RMSE across five simulation replicates (n = 5) were displayed in the form of boxplots. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

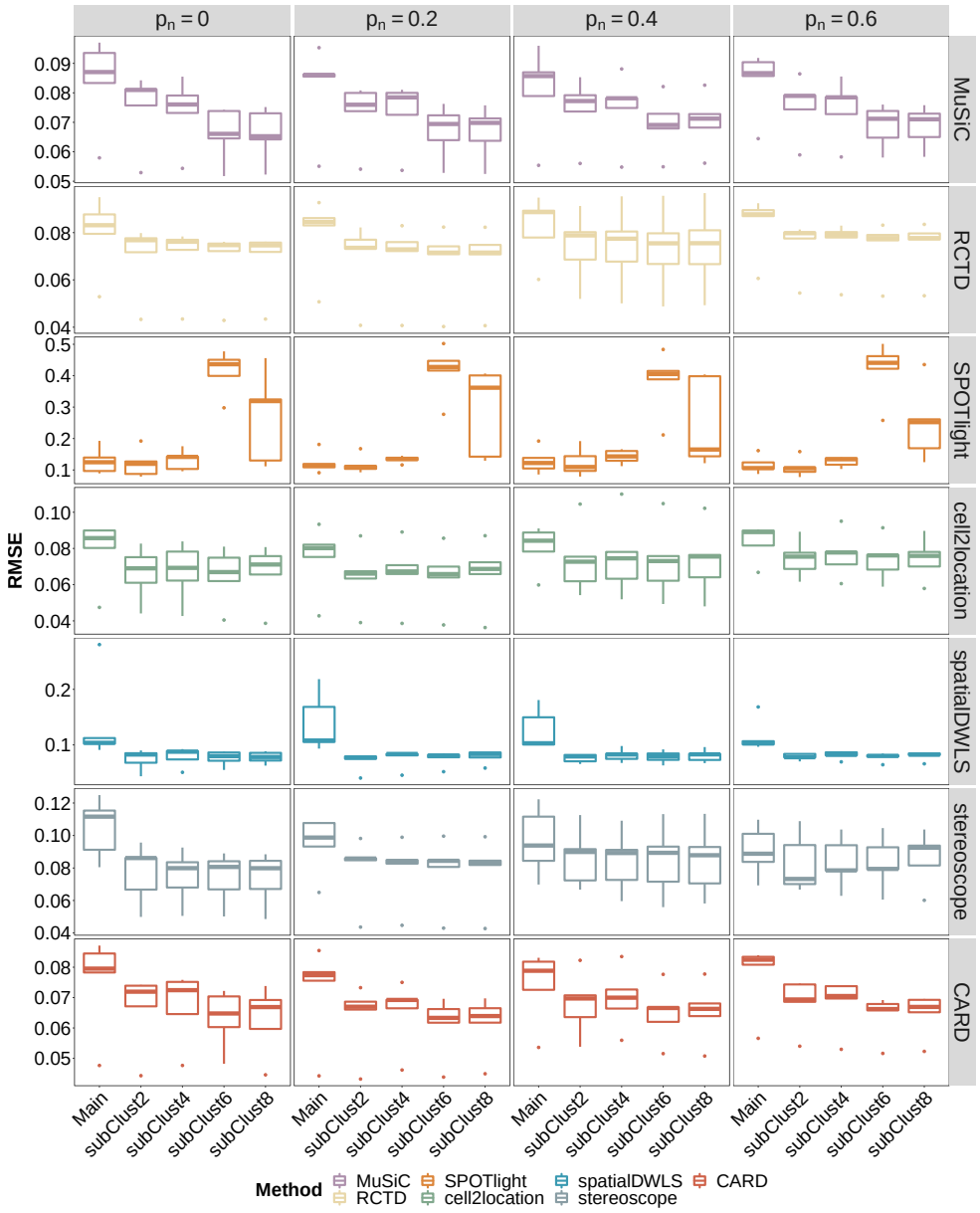


Figure S3.11 Comparison of deconvolution accuracy of different methods at the major cell type level (Scenario I) and at the sub-cell type level.

Compared deconvolution methods (x-axis) include MuSiC (purple), RCTD (wheat), SPOTlight (orange), cell2location (green), spatialDWLS (skyblue), stereoscope (blue gray), and CARD (red). Simulations were performed under different spatial correlation strength as represented by the proportion of noisy locations (p_n). High p_n corresponds to low spatial correlation. We calculated the root mean square errors (RMSE) between the estimated cell type compositions and the true cell type compositions for each method to measure its deconvolution performance. Boxplot displaying the RMSE of each method versus the number of sub-cell types ranging from 0 (Main), 2, 4, 6, to 8 across 5 replicates (n = 5). Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

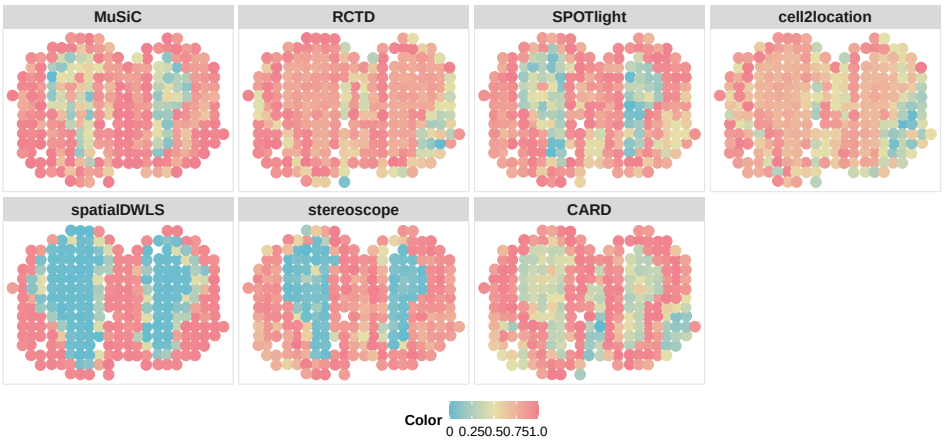


Figure S3.12 Scatterplot of the first principal component of the estimated cell type compositions matrix of mouse olfactory bulb ST data.

Specifically, the first principal component of the estimated cell type compositions by CARD accurately depicts the expected layered structure of mouse olfactory bulb Spatial Transcriptomics data. Here, each dot represents one location and is colored by the first principal component correspondingly.

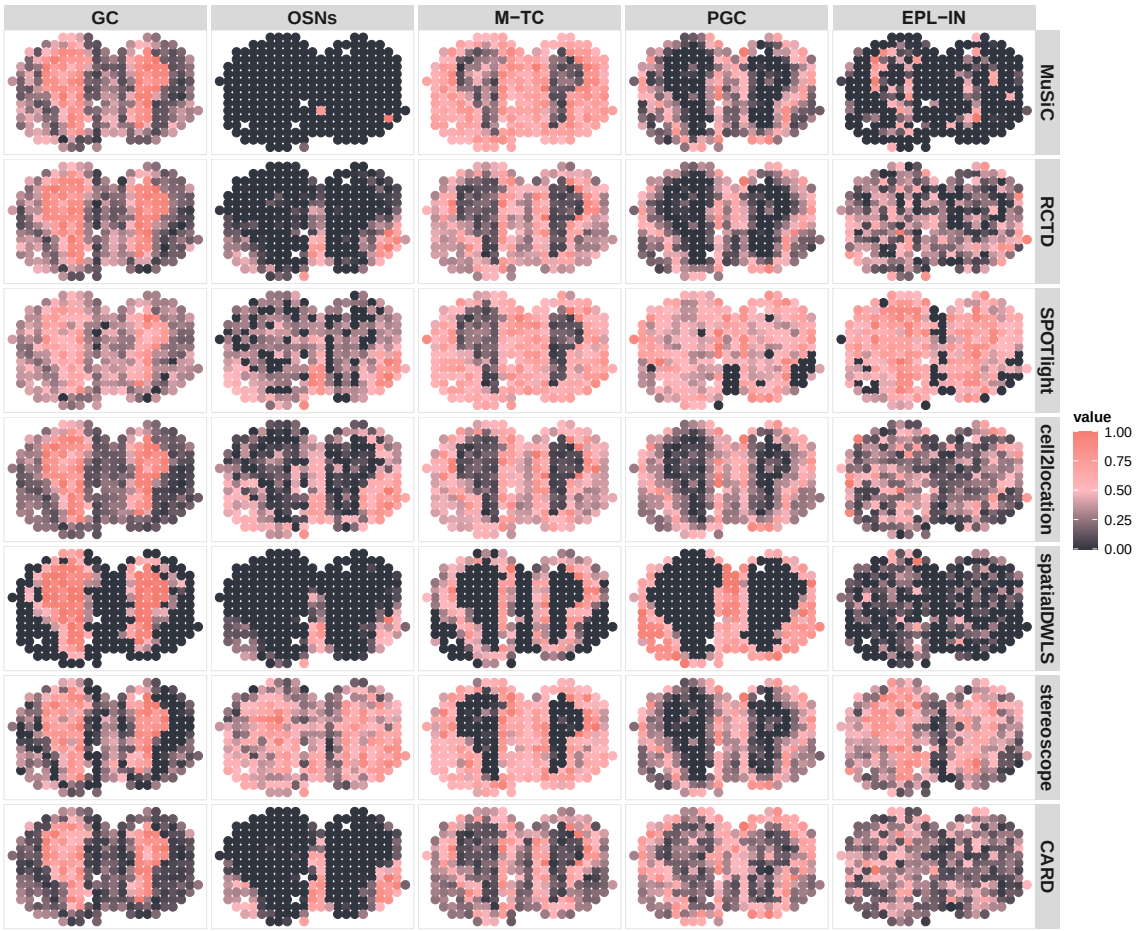


Figure S3.13 Scatter plot of cell type proportion distributions across spatial locations in the mouse olfactory bulb Spatial Transcriptomics data.

Specifically, cell type proportions estimated by CARD can accurately depict the layered structure of mouse olfactory bulb. For example, the granule cells inferred by CARD are highly enriched in the granule layer while other methods show diffused pattern outward from the granule layer towards other layers. CARD also distinguished correctly the adjacent mitral cell layer and glomerular layer, with distinct enrichment of mitral/tufted cells and periglomerular cells in the two layers, respectively, despite the similarity between these two cell types. In contrast, other methods were unable to clearly distinguish the mitral cell layer or glomerular layer with the nerve layer. Here, for each cell type, the cell type proportion was scaled to 0-1 range. Color was shown to represent the 0-1 range of cell type proportions correspondingly.

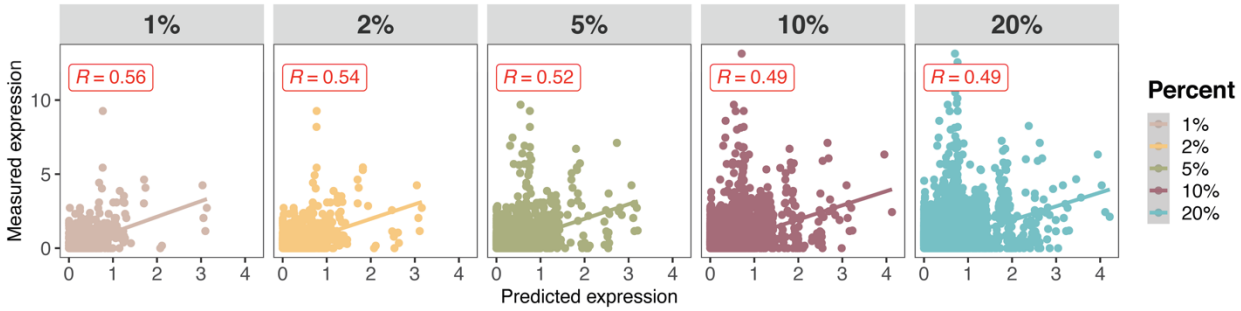


Figure S3.14 Accuracy of CARD imputation in the masking analysis in the mouse olfactory bulb data.

A fixed percentage of locations are masked as missing (1%, 2%, 5%, 10%, and 20%) and CARD is used to impute the gene expression on the masked locations. Scatterplot displays the relationship between the estimated gene expression for the masked spot and the true gene expression. Here, each dot represents one masked spatial location in one simulation replicate setting.

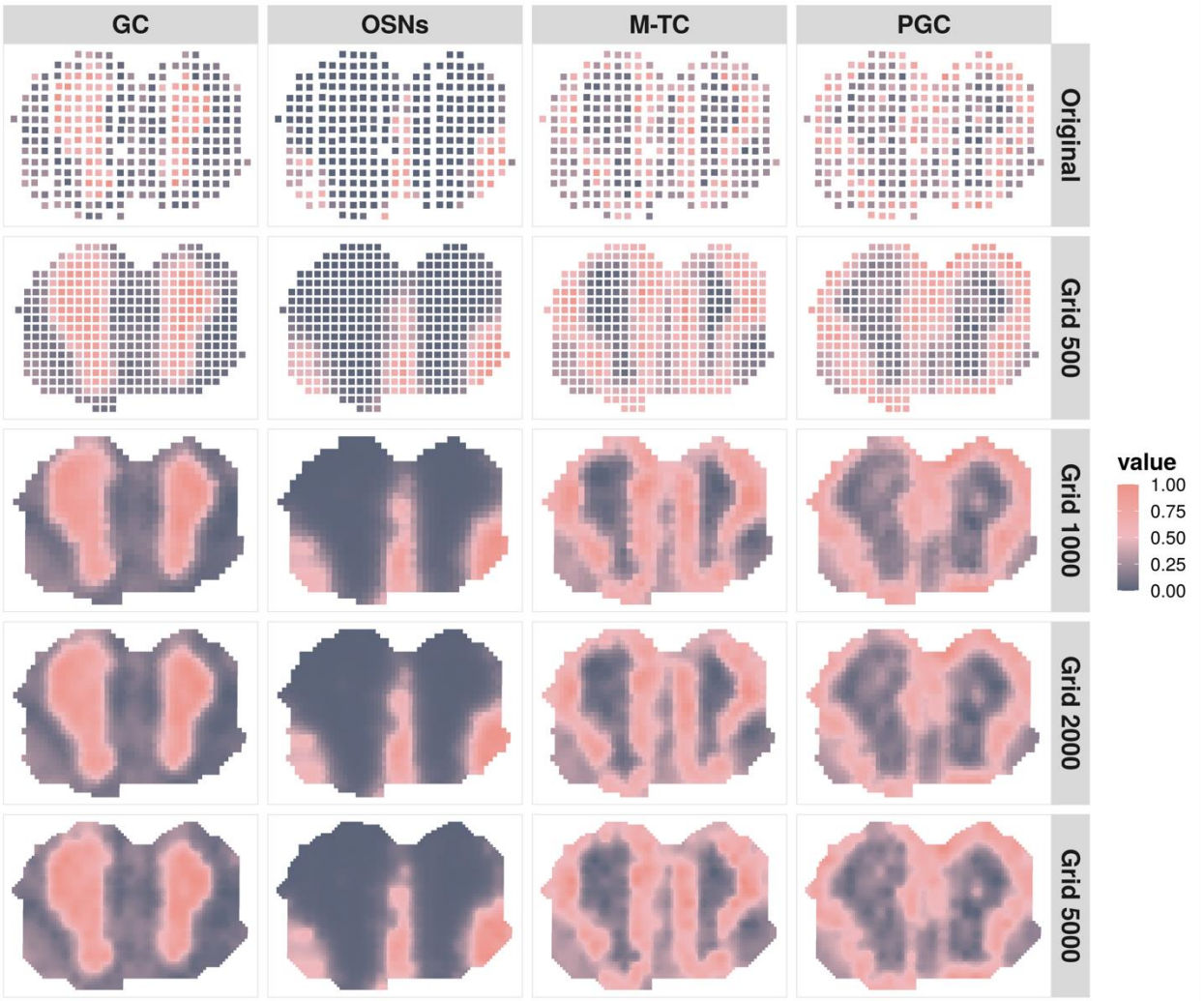


Figure S3.15 The refined spatial map of cell type composition constructed by CARD.

The refined spatial map of cell type composition captures fine grained details of the layered structure in the olfactory bulb with enhanced resolutions. The spatial pattern is shown for the distribution of granule cells (GC), olfactory sensory neurons (OSNs), mitral/tufted cells (M-TC), and periglomerular cells (PGC) at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of cell type proportions, which are not shown obviously in the original dataset at lower resolution.

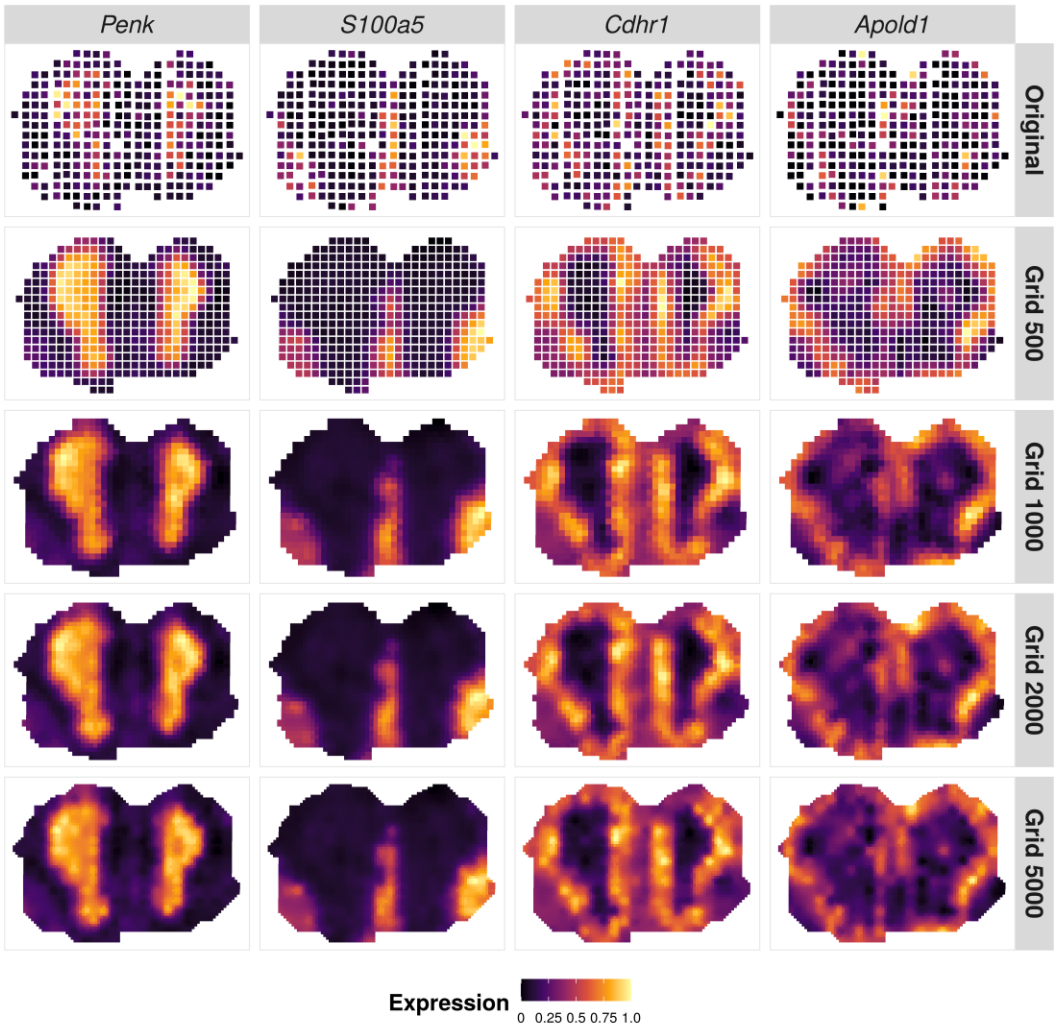


Figure S3.16 The refined spatial map of gene expression constructed by CARD.

The refined spatial map of gene expression captures fine grained details of the layered structure in the olfactory bulb with enhanced resolutions. The spatial pattern is shown for granule cells (GC) marker gene: *Penk*, olfactory sensory neuron's marker gene: *S100a5*, mitral/tufted cell marker gene: Cdhr1 and periglomerular cell's marker gene: Apold1 at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of previously known marker genes, which are not shown obviously in the original dataset at lower resolution.

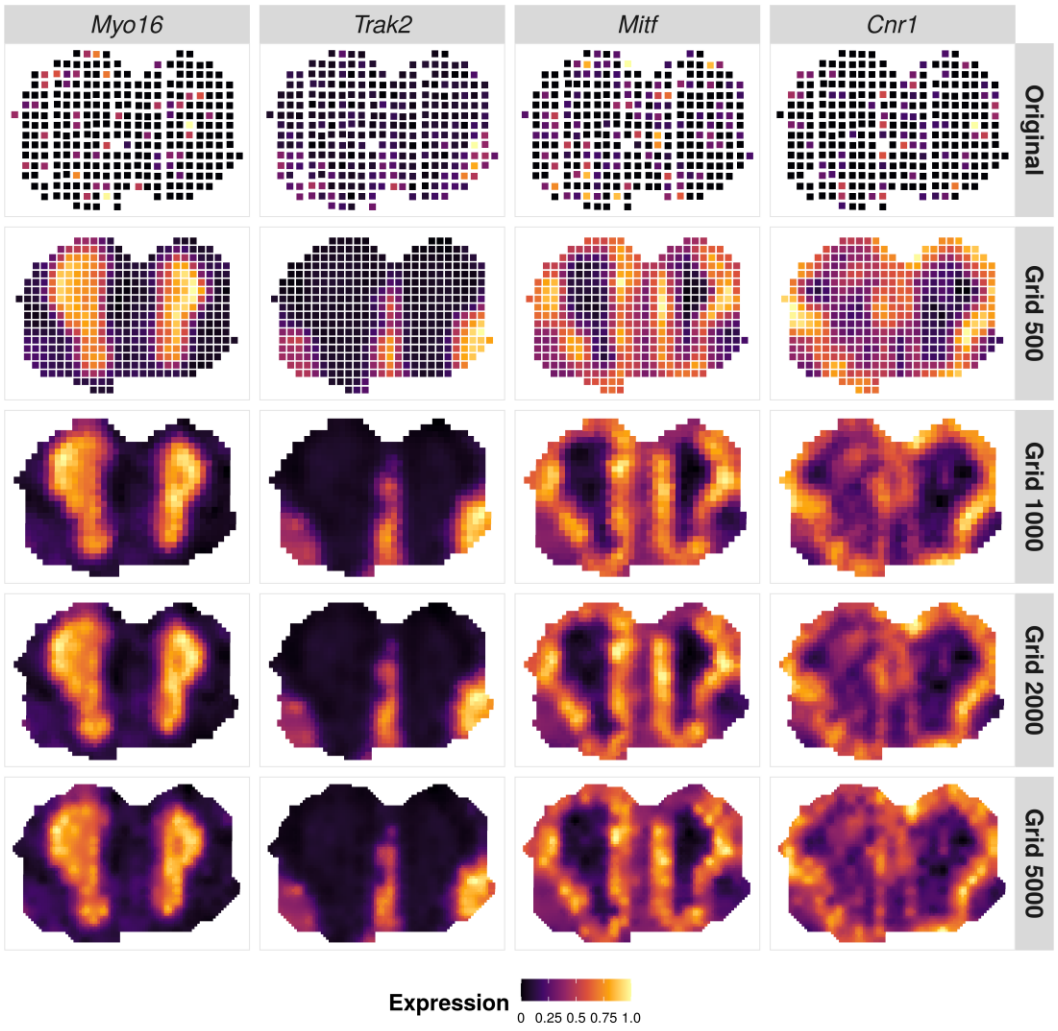


Figure S3.17 The refined spatial map of gene expression constructed by CARD helps to reveal spatial patterns of genes.

The refined spatial map of gene expression captures fine grained details of the layered structure in the olfactory bulb with enhanced resolutions. The spatial pattern is shown for the non-marker genes *Myo16*, *Trak2*, *Mitf* and *Cnr1* at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of lowly expressed genes, which are not shown obviously or differently in the original dataset at lower resolution.

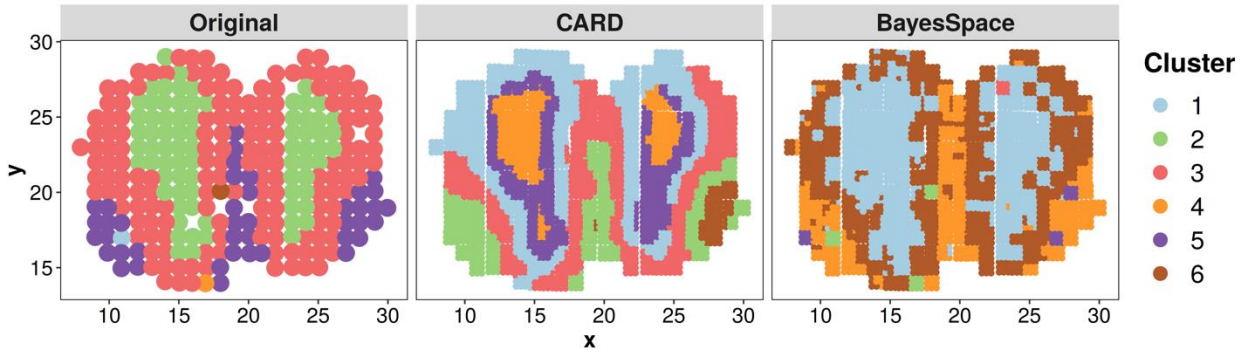


Figure S3.18 Clustering results on the original mouse olfactory bulb ST data (n = 282), CARD and BayesSpace imputed data at a higher resolution (n = 2538).

Here, we directly used CARD to impute gene expression on the fixed sub-spots created by BayesSpace. We then performed clustering analysis on the imputed data by either CARD or BayesSpace on the same set of sub-spots. Specifically, clustering analysis was performed by K-means clustering algorithm on the first 20 PCs of all three data. The clustering results based on CARD displayed a clear inside-out layered structure that resembles the anatomic organization of the olfactory bulb, much more so than that obtained with the original scale data and that by BayesSpace.

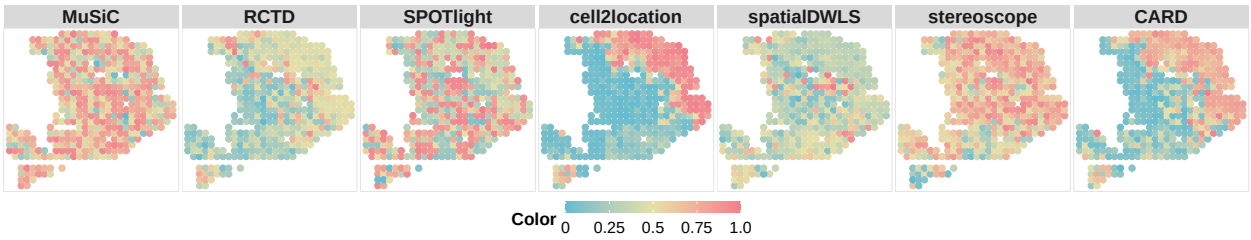


Figure S3.19 Scatterplot of the first principal component of the estimated cell type compositions matrix.

Specifically, the first principal component of the estimated cell type compositions by CARD and cell2location clearly capture a gross regional segregation between cancer and non-cancer regions in human pancreatic ductal adenocarcinomas data. In contrast, for MuSiC, SPOTlight, and spatialDWLS, their PC1 shows almost completely random spatial pattern. Here, each dot represents one location and is colored by the first principal component correspondingly.

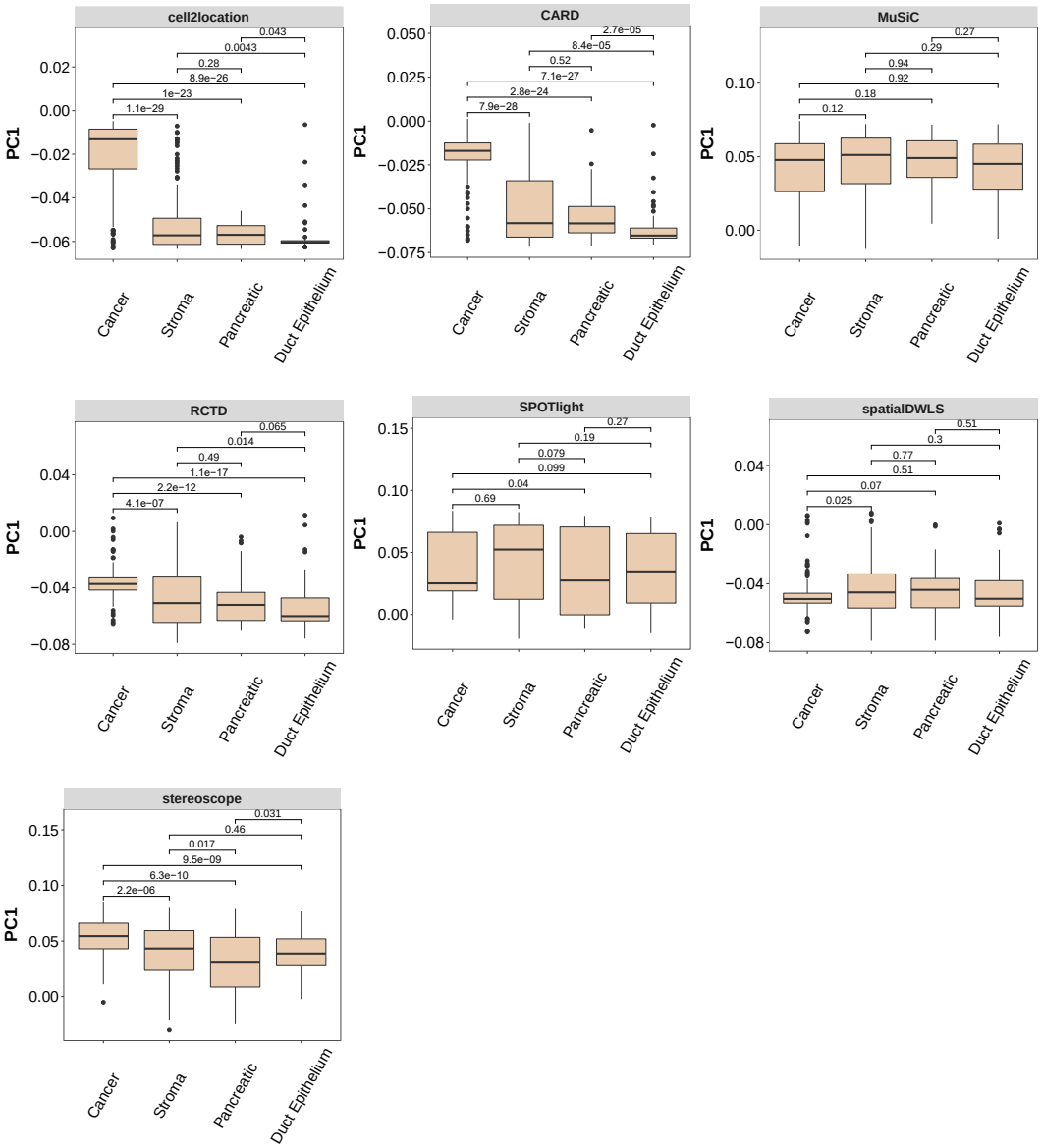


Figure S3.20 Boxplot of the first principal component score (PC1) of the estimated cell type proportions by CARD and other methods at different regions in the human PDAC dataset respectively.

The two-sided Wilcoxon Rank Sum test was used to pair wisely test the difference between PC1 score in different regions. Specifically, the PC1 score of CARD's inferred cell type proportion (p-value = 2.7e-05) significantly differentiate between the pancreatic and ductal regions, while cell2location cannot (p-value = 0.043 > Bonferroni corrected p-value threshold 0.05/6 =0.008). In contrast, none of the other methods were able to significantly differentiate between ductal and stroma regions or between pancreatic and ductal regions. Here, the sample size for each region is n = 137 for Cancer region, n = 147 for Stromal region, n = 70 for Pancreatic region while n = 72 for Ductal Epithelium region. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

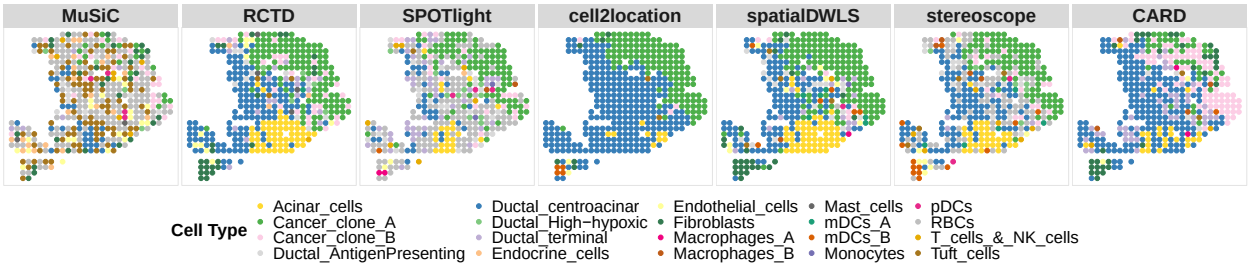


Figure S3.21 Spatial distribution of dominant cell type on each location based on the cell type proportions from each method.

Specifically, the dominant cell types on each location from CARD deconvolution capture the segregation between cancer and non-cancer regions as well as distinguish two sub cancer regions that were missed by other methods. Here, each dot represents one spatial location colored by the dominant cell type.

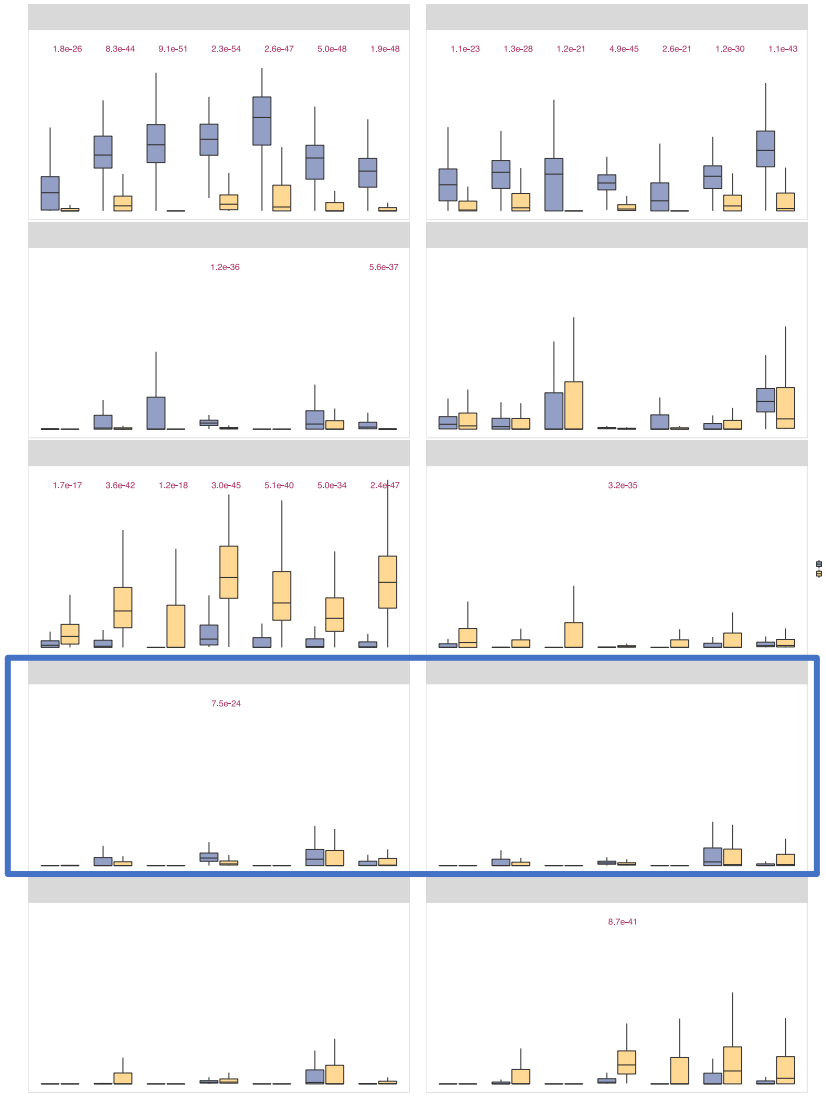


Figure S3.22 Comparisons of cell type proportions in cancer region versus non-cancer region.

For each cell type, a two-sided Wilcoxon Rank Sum test was conducted for each method's deconvoluted cell type proportions to compare the cancer vs non-cancer region. Here, y axis represents the estimated proportion and x axis represents each method while color represents either cancer or non-cancer regions. CARD reveals distinct distribution of two macrophage subpopulations between the cancer (n = 137) and non-cancer regions (n = 289), which are missed by the other methods. Specifically, macrophage B is enriched in the non-caner region and such spatial enrichment pattern underlies its function as tissue resident macrophages (Zhu et al. 2017, Moncada et al. 2020). In contrast, macrophage A is enriched in both cancer and non-cancer regions, likely reflecting a high inflammatory state of this macrophage subpopulation (Moncada et al. 2020, Rőszer 2015). The distinct distribution of macrophage subpopulations detected by CARD represents a key functional signature of the regional compartmentalization of cancer tissues. Here, each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

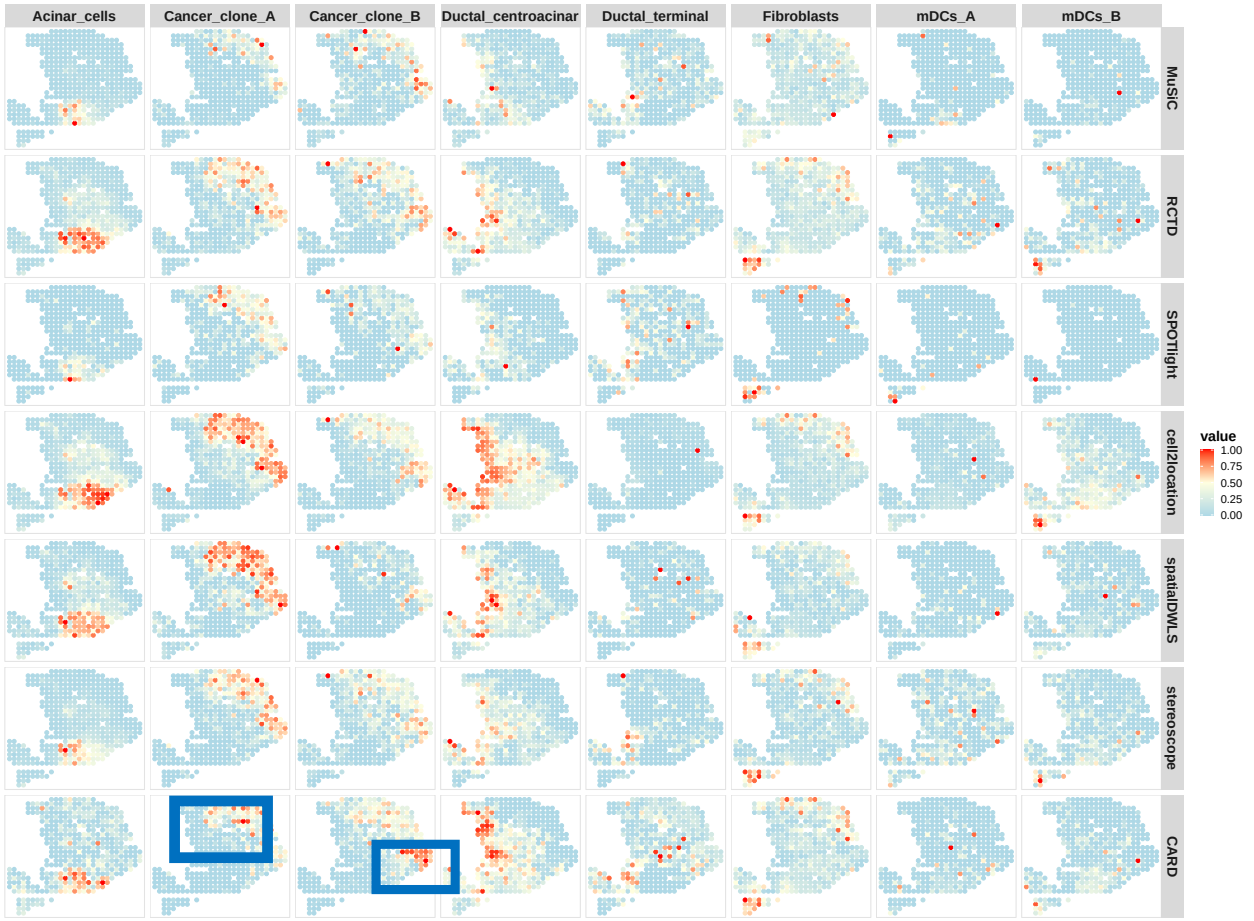


Figure S3.23 Scatter plot of cell type proportion distributions across spatial locations in the human pancreatic ductal adenocarcinomas data.

Specifically, cell type proportions estimated by CARD can accurately localize cell types into the biologically meaningful tissue region. For example, cancer clone A and cancer clone B cells inferred by CARD reside in different regions of cancer regions. Ductal centroacinar cells inferred by CARD are enriched in the ductal epithelium region and partially in the normal pancreatic tissue region, consistent with the spatial expression pattern of the marker gene *Crisp3* and early literature . The terminal ductal cells inferred by CARD mainly reside around the ductal epithelium area and normal pancreatic tissue region, consistent with the expression pattern of the marker gene *Tff3* and early literature[7]. In contrast, none of the other methods capture the expected spatial localization of both ductal centroacinar and terminal ductal cells. In addition, acinar cells inferred by CARD are mainly enriched in the normal pancreatic tissue region but are either absent in the pancreatic region or are diffused outward from the pancreatic region towards the stroma region and cancer region when inferred by the other methods. Here, for each cell type, the cell type proportion was scaled to 0-1 range. Color was shown to represent the 0-1 range of cell type proportions correspondingly.

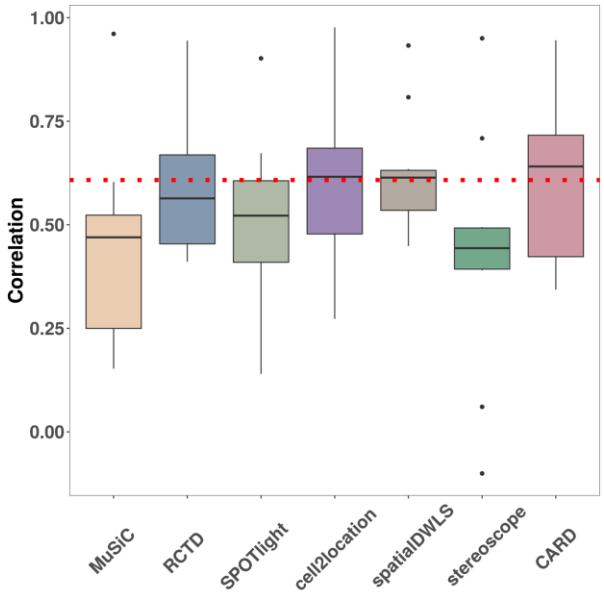


Figure S3.24 CARD generated consistent deconvolution results across different scRNASeq references.

Boxplot displaying the pairwise correlation of estimated cell type proportions from different scRNA-seq references (n = 10) for the human PDAC data. Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box. The median correlation across all different scRNA-seq references for CARD is 0.64, higher than all other methods (the median correlation is 0.47 for MuSiC, 0.56 for RCTD, 0.52 for SPOTlight, 0.61 for cell2location, 0.61 for spatialDWLS, and 0.44 for stereoscope). Here, the cell type proportions are pair wisely compared between all scRNA-seq references, including acinar cells, cancer clone A cells/ductal 2 cells, ductal centroacinar cells/ductal 1 cells, endocrine cells, and endothelial cells. We focused on these cell types because they exist in all matched and unmatched scRNA-seq references.

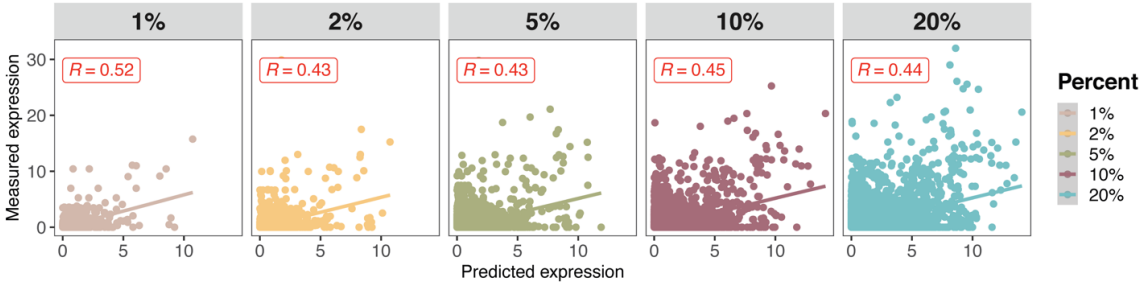


Figure S3.25 Accuracy of CARD imputation in the masking analysis in the human pancreatic ductal adenocarcinoma (PDAC) data.

A fixed percentage of locations are masked as missing (1%, 2%, 5%, 10%, and 20%) and CARD is used to impute the gene expression on the masked locations. Scatterplot displays the Pearson's correlation between the estimated gene expression for the masked spot and the true gene expression. Here, each dot represents one masked spatial location in one simulation replicate setting.

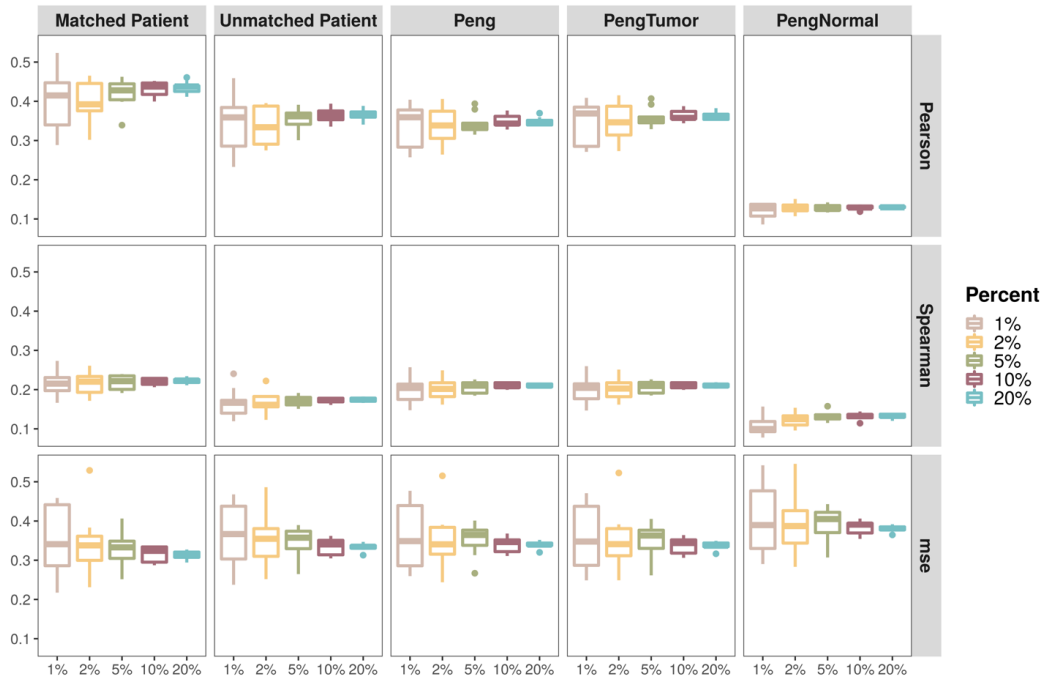


Figure S3.26 Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10) when using different scRNA-seq as references.

A fixed percentage of locations are masked as missing (x-axis) and CARD is used to impute the gene expression on the masked locations. Three different metrics (y-axis) are used to evaluate imputation accuracy in terms of the similarity between the imputed expression and true expression on masked locations: Pearson's correlation, Spearman's correlation and mean square error (MSE). Using the scRNA-seq from the same patient displays the expected better performance in terms of prediction correlation and error and the external scRNA-seq dataset from the normal samples displays the worst performance. Here, each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

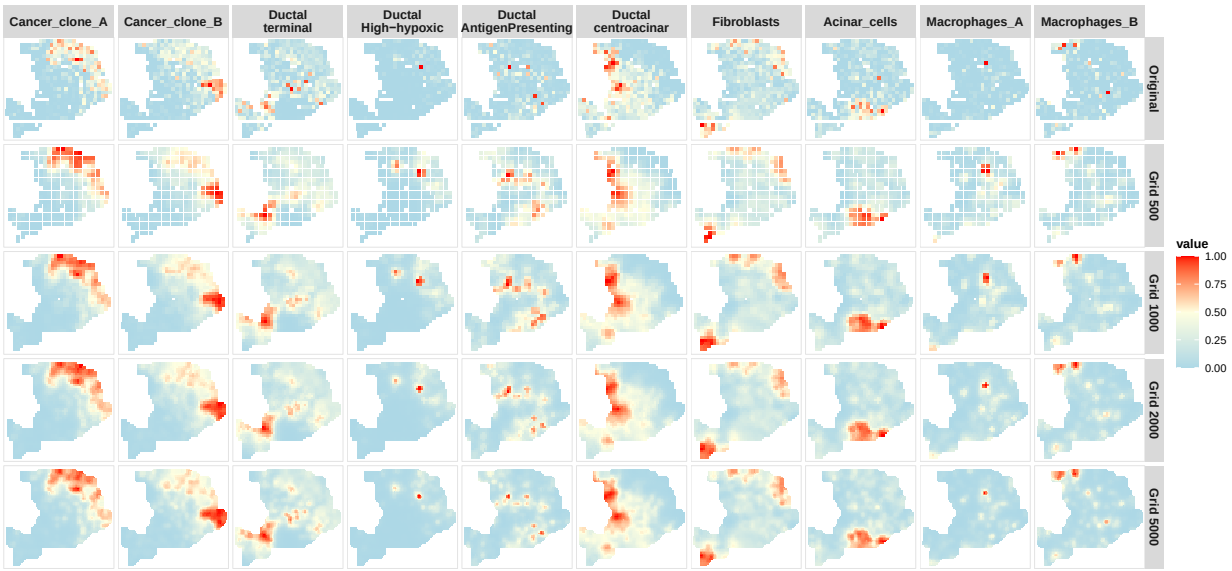


Figure S3.27 The refined spatial map of cell type composition constructed by CARD in the human pancreatic ductal adenocarcinoma (PDAC) tissue.

The refined spatial map of cell type composition captures fine grained details of the regional structure of PDAC tissue with enhanced resolutions. The spatial pattern is shown for the distribution of cancer clone A cells, cancer clone B cells, ductal terminal cells, ductal high hypoxic cells, ductal antigen presenting cells, ductal centroacinar cells, fibroblast cells, acinar cells, microphage A cells and macrophage B cells at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of cell type proportions, which are not shown obviously in the original dataset at lower resolution.

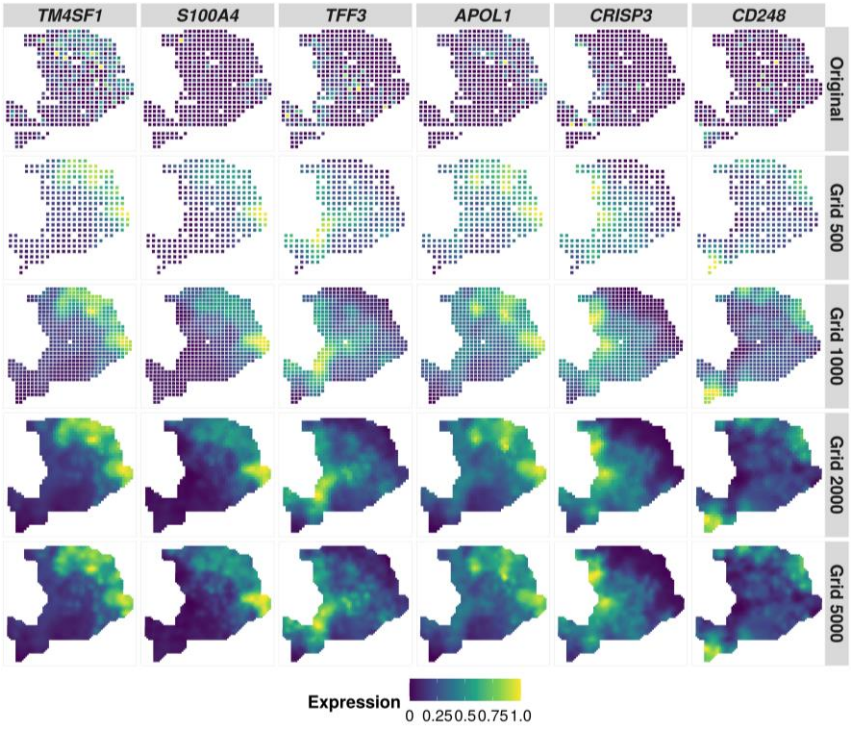


Figure S3.28 The refined spatial map of gene expression constructed by CARD in the human pancreatic ductal adenocarcinoma (PDAC) tissue.

The refined spatial map of gene expression captures fine grained details of the regional structure of PDAC tissue with enhanced resolutions. The spatial pattern is shown for cancer clone A cell marker gene *TM4SF1*, cancer clone B cell marker gene *S100A4*, ductal terminal cells marker gene *TFF3*, ductal high hypoxic cell marker gene *APOL1*, ductal centroacinar cell marker gene *CRISP4* and fibroblast cell marker gene *CD248* at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of previously known marker genes, which are not shown obviously in the original dataset at lower resolution.

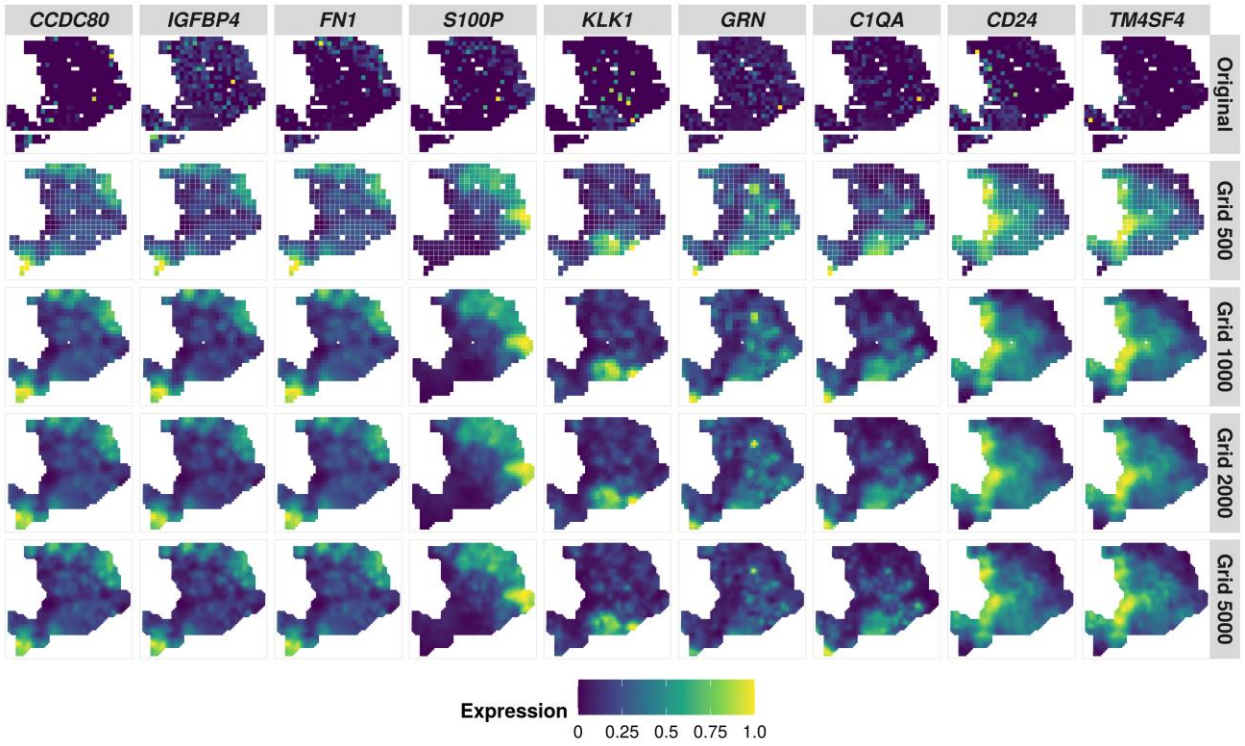


Figure S3.29 The refined spatial map of gene expression constructed by CARD in the human pancreatic ductal adenocarcinoma (PDAC) tissue.

The refined spatial map of gene expression captures fine grained details of the regional structure of PDAC tissue with enhanced resolutions. The spatial pattern is shown for non-marker genes *CCDC80*, *IGFBP4*, *FN1*, *S100P*, *KLK1*, *GRN*, *C1QA*, *CD24*, *TM4SF4* at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of new genes, which are not shown obviously or different in the original dataset at lower resolution.

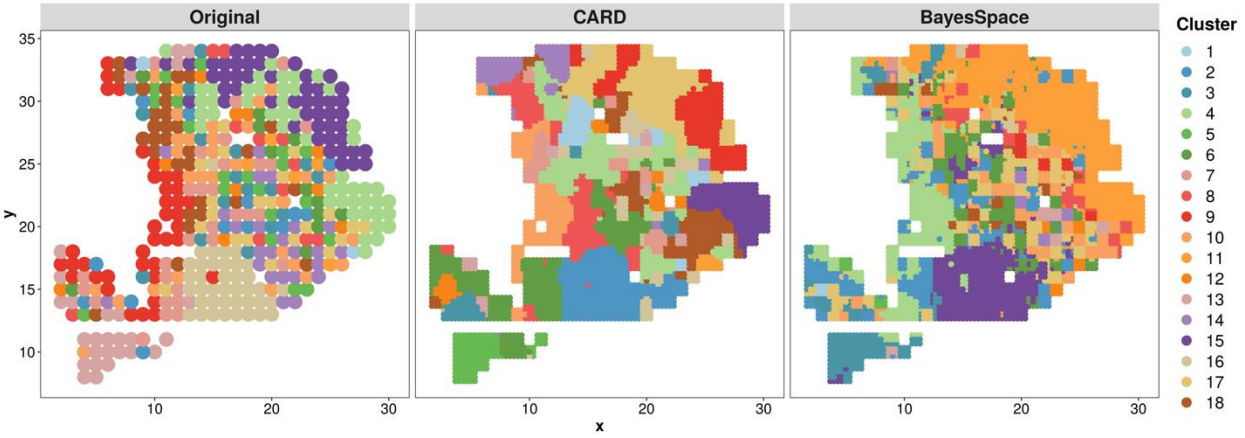


Figure S3.30 Clustering results on the original human PDAC ST data (n = 428), CARD and BayesSpace imputed data at a higher resolution (n = 3852).

Here, we directly used CARD to impute gene expression on the fixed sub-spots created by BayesSpace. We then performed clustering analysis on the imputed data by either CARD or BayesSpace on the same set of sub-spots. Specifically, clustering analysis was performed by K-means clustering algorithm on the first 20 PCs of all three data. Clustering analysis on CARD imputed high resolution data also segregated the two cancer sub-regions, the normal pancreatic region, and the ductal region, more clearly than the original data and refined data by BayesSpace.

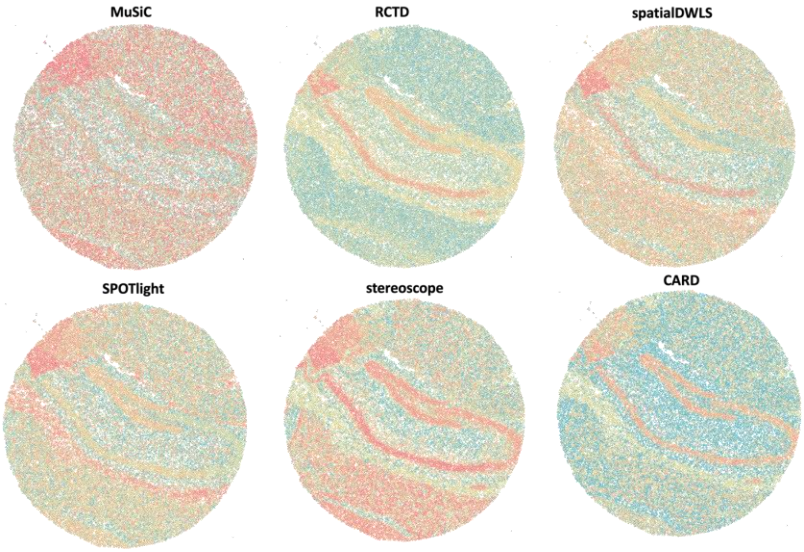


Figure S3.31 Scatterplot of the first principal component of the estimated cell type compositions matrix.

Specifically, the first principal component of the estimated cell type compositions by CARD clearly captures the curved shape of hippocampus accurately that is consistent with the UMI counts displayed in Figure 5A, more so than the other methods. Here, each dot represents one location and is colored by the first principal component correspondingly.

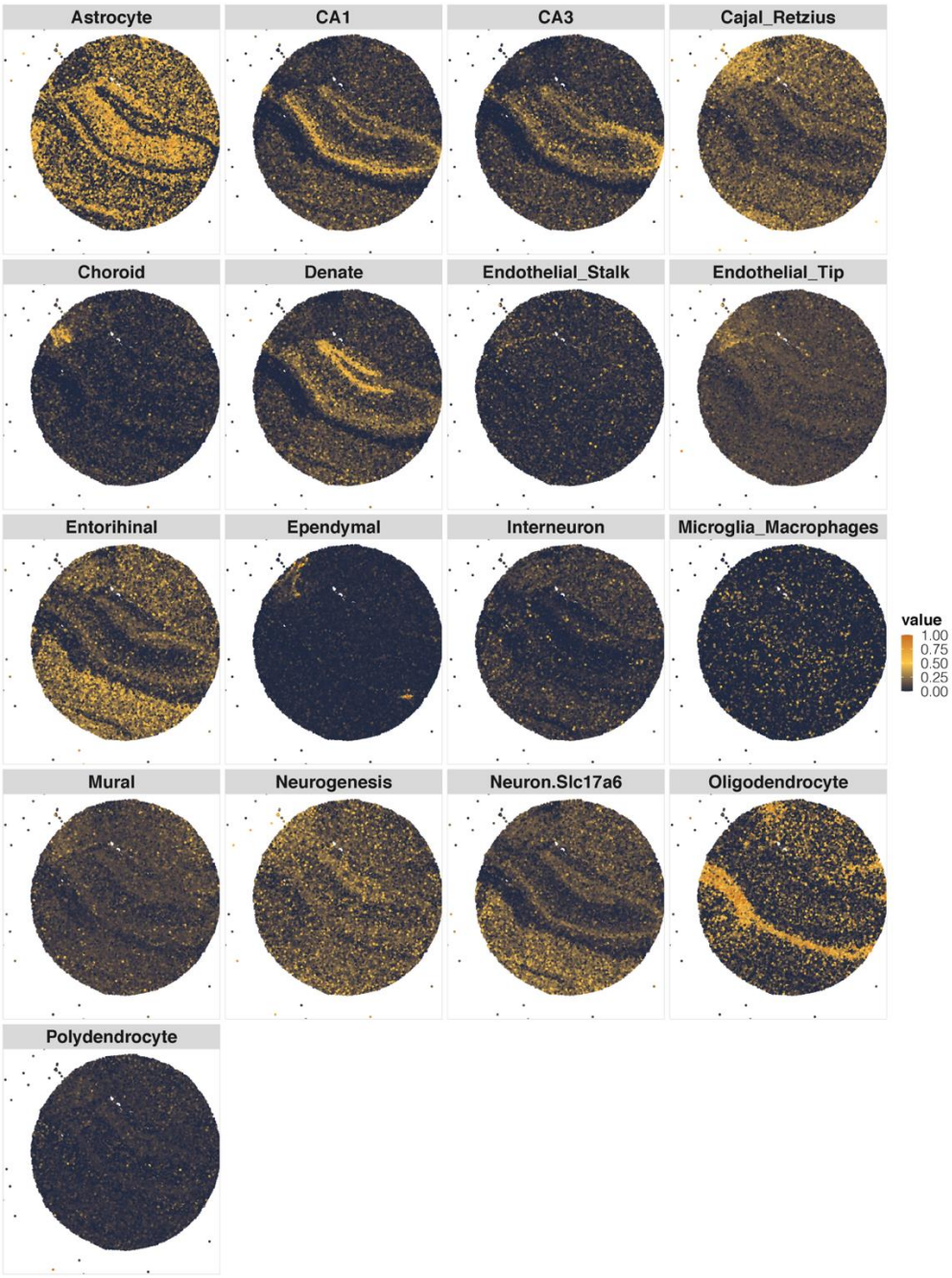


Figure S3.32 Scatter plot of cell type proportion distributions across spatial locations in the mouse hippocampus Slide-seq V2 data.

Specifically, cell type proportions estimated by CARD can accurately localize cell types into the biologically meaningful tissue region. For example, CA1 cells are highly enriched in CA1; CA3 cells mainly localize in CA3; dentate cells reside in a C-shaped ring region of dentate gyrus. Here, for each cell type, the cell type proportion was scaled to 0-1 range. Color was shown to represent the 0-1 range of cell type proportions correspondingly.

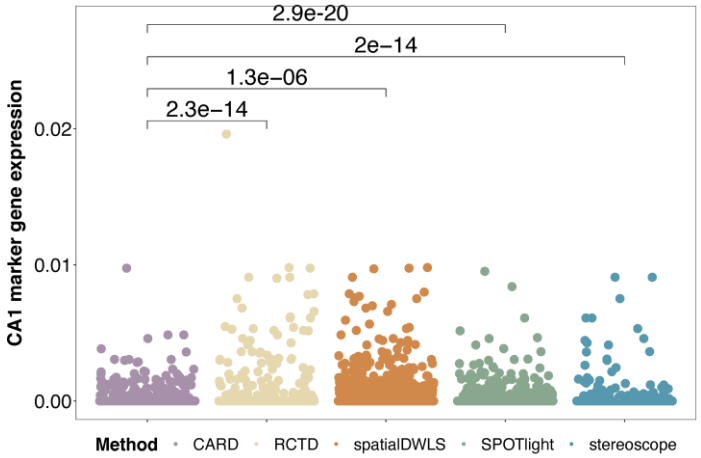


Figure S3.33 RCTD, spatialDWLS, stereoscope, and SPOTlight incorrectly locate CA3 cells into CA1 regions more so than CARD.

The dot plot displays the gene expression of CA1 cells *Wfs1* in their inferred CA3 spatial locations versus in the CARD inferred CA3 spatial locations. RCTD, spatialDWLS, and stereoscope perform similarly, all inferring CA3 cells incorrectly in CA1 region based on the spatial distribution of the inferred dominant cell type and the spatial distribution of the inferred CA3 cells, more so than CARD. Specifically, we quantify the gene expression of the marker of CA1 cells *Wfs1* in the spatial locations that are dominated by CA3 cells inferred by each method. We expect that if the other methods incorrectly locate CA3 cells into CA1 regions more so than CARD, the marker gene expression of other methods' inferred CA3 locations should be significantly higher than that inferred by CARD. Consistent with our expectation, we observed that the Wfs1 marker gene expression is statistically significantly higher in other methods' inferred CA3 regions than CARD inferred CA3 regions, indicating that other methods incorrectly locate more CA3 cells in CA1 region, more so than CARD. Pairwise differences of the gene expression between CARD and other methods were assessed by two-sided Wilcoxon Rank-Sum test.

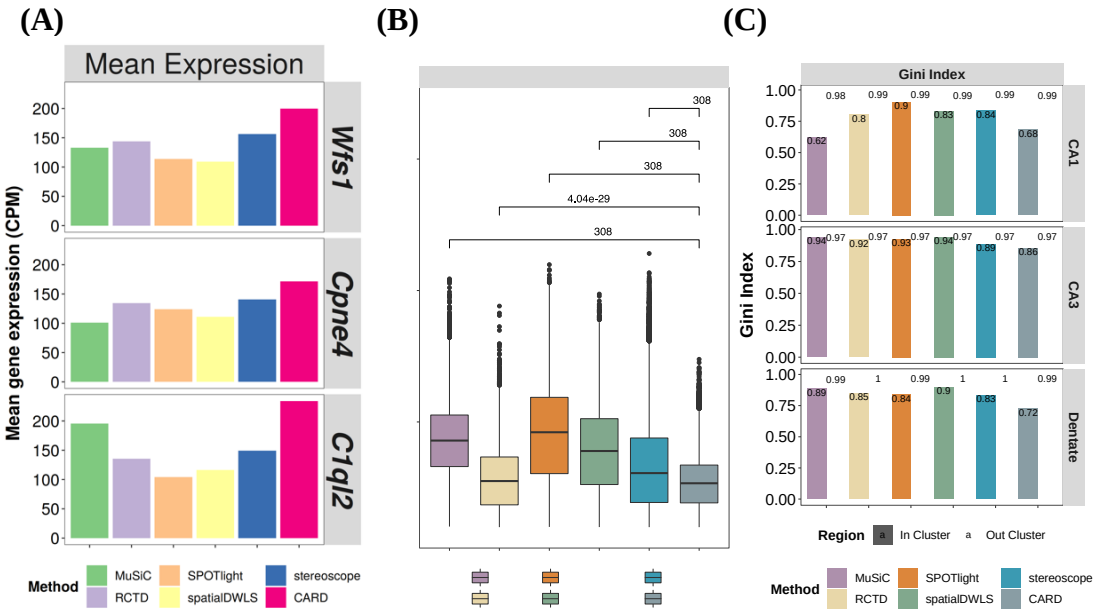


Figure S3.34 Comparisons of the specificity of inferred major regions in the Slide-seq V2 mouse hippocampus data by different deconvolution methods.

(A) Bar plots display the comparisons of the mean gene expression level in the major regions inferred by different deconvolution methods (same as Figure 5D); **(B)** Boxplot displays the local inverse Simpson's index (LISI) for each method (n = 41758) while pairwise differences of the LISI value between CARD and other methods were assessed by one-sided Wilcoxon Rank-Sum test; Each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box. **(C)** Gini index of marker genes in the major regions inferred by different deconvolution methods versus outside of the major region. We reasoned that a good deconvolution method would yield accurate cell type composition estimates and subsequently accurate tissue structures and would thus capture the expected structure specific expression pattern for the marker genes well. Therefore, for each marker gene, we calculated the three metrics on the inferred tissue structures from a given method to serve as quantifications for its deconvolution performance. The three metrics include: (1) mean gene expression in the tissue structure where the marker gene is expected to be enriched, where a high value is desirable; (2) local inverse Simpson's index, where a lower value indicates a better segregation between the hippocampus structures; (3) Gini index for the marker gene within the specific tissue structure, where a lower Gini index indicates higher expression homogeneity within the structure. Here the marker gene for CA1 is *Wfs1*, for CA3 is *Cpne4* and for dentate cells is *C1ql2*.

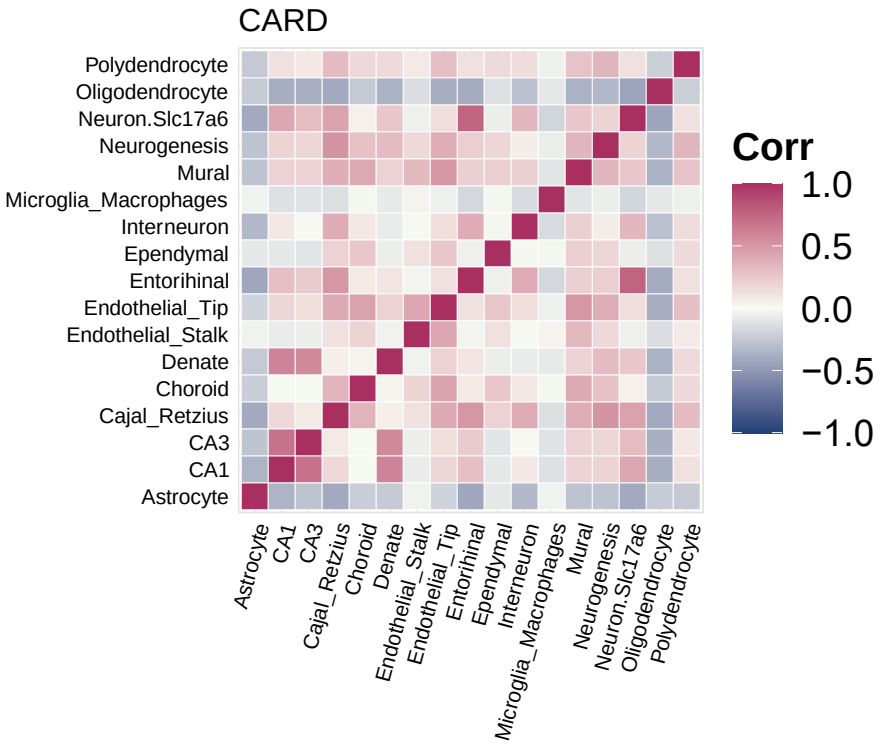


Figure S3.35 Correlations in cell type proportion across spatial locations between pairs of cell types inferred by CARD in the mouse hippocampus Slide-seq V2 data.

Color is scaled by the correlation value.

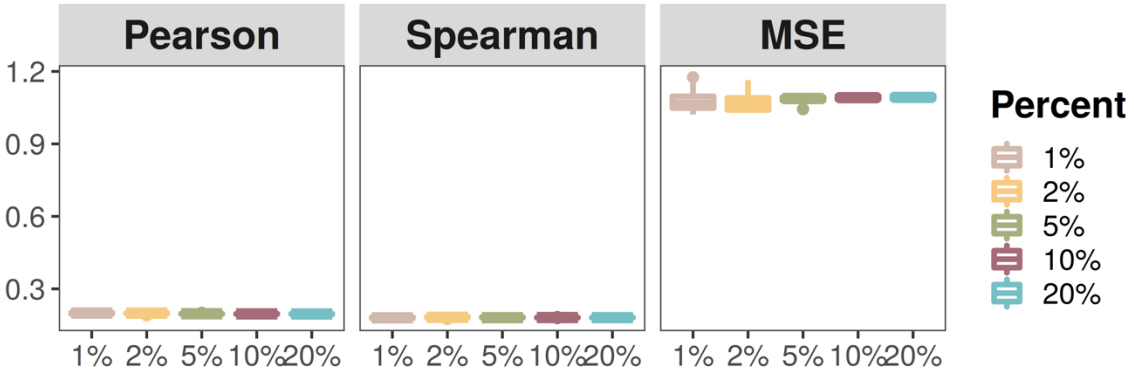


Figure S3.36 Accuracy of CARD imputation in the masking analysis across 10 replicates (n = 10) in the mouse hippocampus Slide-seq V2 data.

A fixed percentage of locations are masked as missing (x-axis) and CARD is used to impute the gene expression on the masked locations. Three different metrics (y-axis) are used to evaluate imputation accuracy in terms of the similarity between the imputed expression and true expression on masked locations: Pearson's correlation, Spearman's correlation and mean square error (MSE). Here, each boxplot ranges from the first and third quartiles with the median as the horizontal line while whiskers represent 1.5 times the interquartile range from the lower and upper bounds of the box.

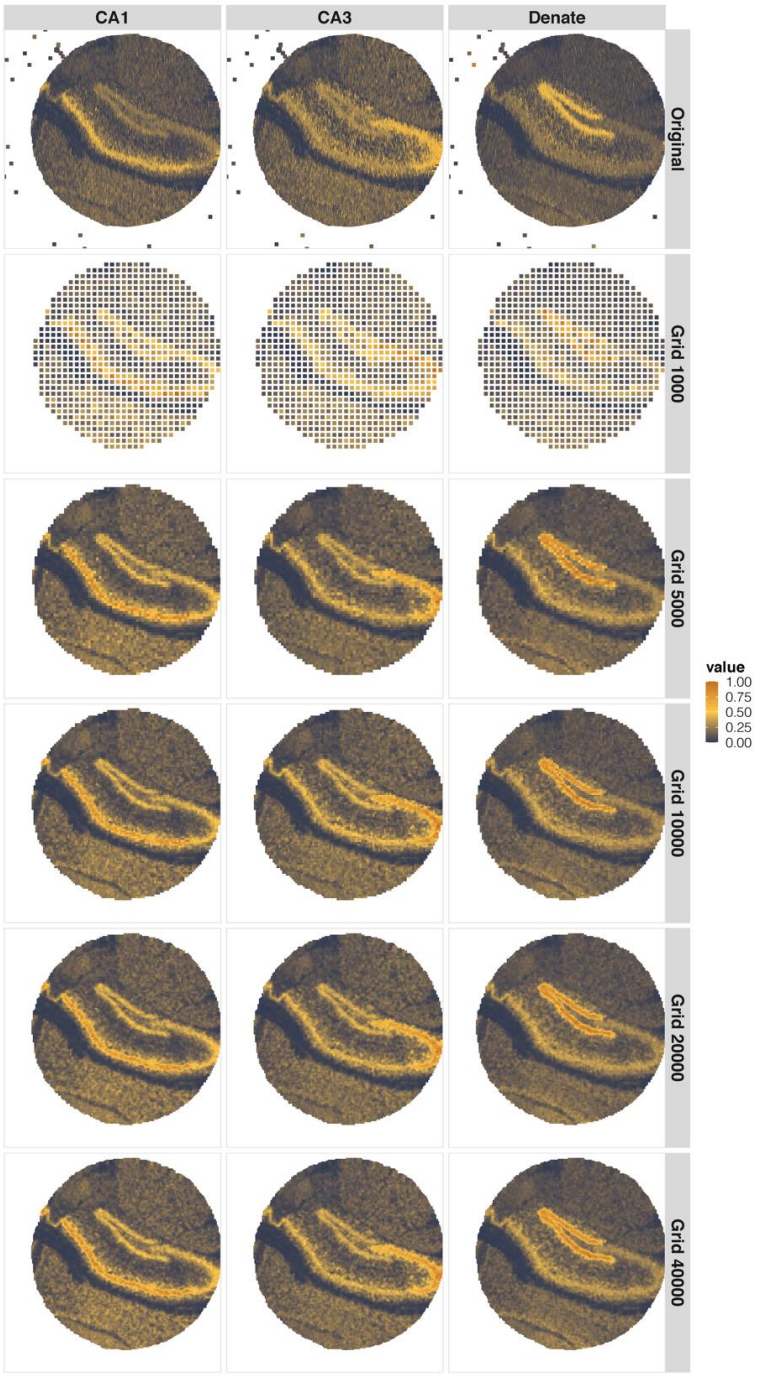


Figure S3.37 The refined spatial map of cell type composition constructed by CARD in the mouse hippocampus tissue.

The refined spatial map of cell type composition captures fine grained details of the regional structure of mouse hippocampus tissue with enhanced resolutions. The spatial pattern is shown for the distribution of CA1 cells, CA3 cells and dentate cells at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of cell type proportions, which are not shown obviously in the original dataset at lower resolution.

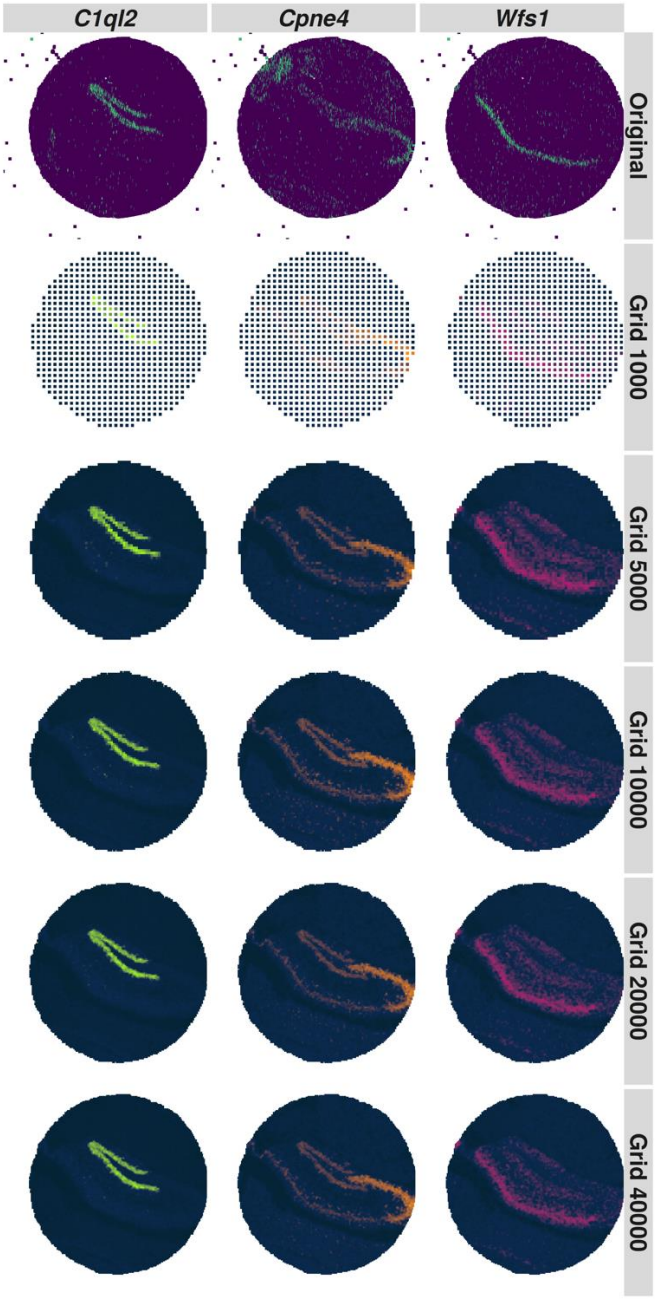


Figure S3.38 The refined spatial map of gene expression constructed by CARD in the mouse hippocampus tissue.

The refined spatial map of gene expression captures fine grained details of the regional structure of mouse hippocampus tissue with enhanced resolutions. The spatial pattern is shown for CA1 cell marker gene *Wfs1*, CA3 cell marker gene *Cpne4*, and dentate cell marker gene *C1ql2* at different resolution represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of previously known marker genes, which are not shown obviously in the original dataset at lower resolution. Here, color was scaled to 0-1 range by the specific marker gene expression. Due to the large sparsity of the original Slide-seq V2 data, we set the color for the zero expression as the lowest color in the color palette to visualize it clearly.

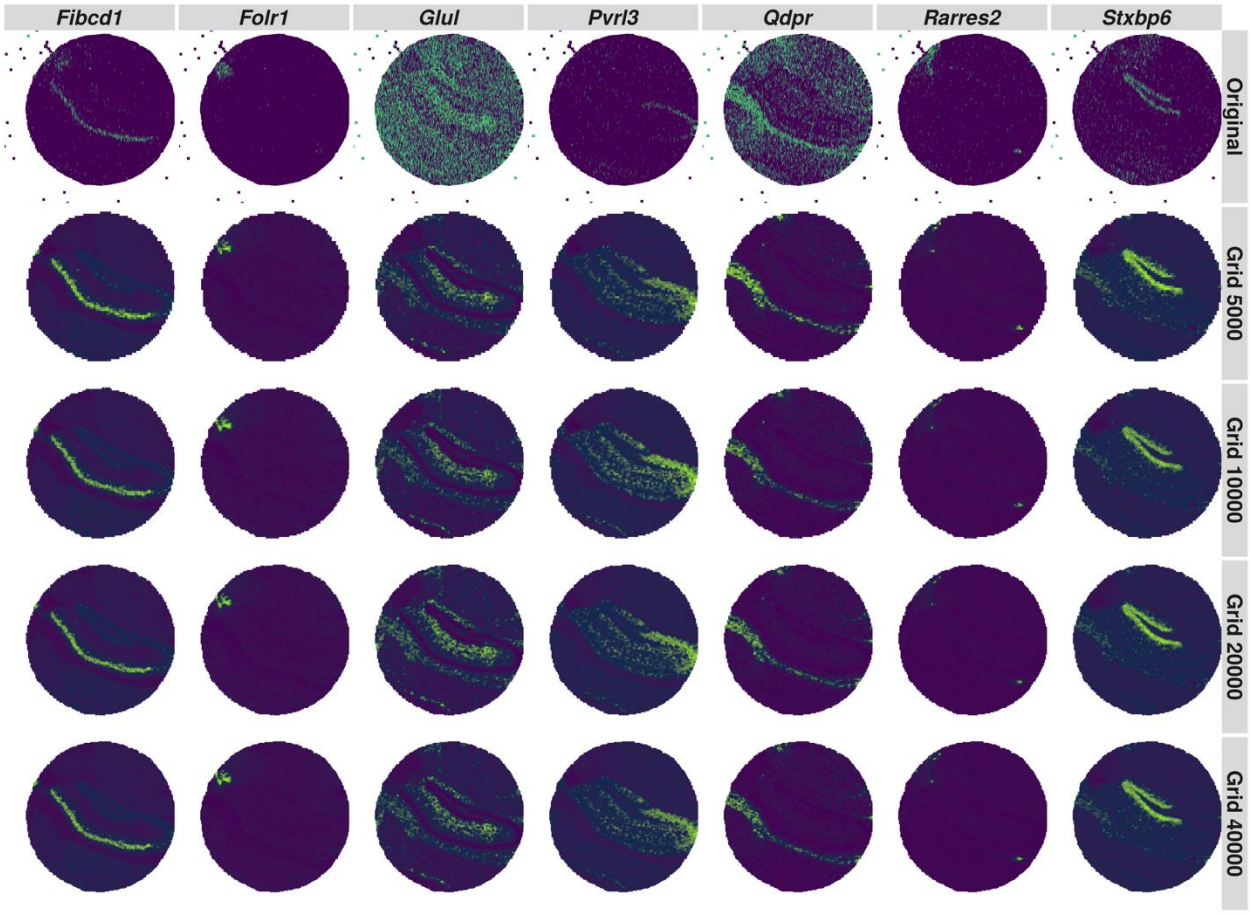


Figure S3.39 The refined spatial map of gene expression constructed by CARD in the mouse hippocampus tissue.

The refined spatial map of gene expression captures fine grained details of the regional structure of mouse hippocampus tissue with enhanced resolutions. The spatial pattern is shown for the non-marker genes at different resolutions represented by the number of gridded spatial locations. CARD can generate an enhanced spatial pattern of non- marker genes, which are not shown obviously or shown differently in the original dataset at lower resolution. Here, color was scaled to 0-1 range by the specific marker gene expression. Due to the large sparsity of the original Slide-seq V2 data, we set the color for the zero expression as the lowest color in the color palette to visualize it clearly.

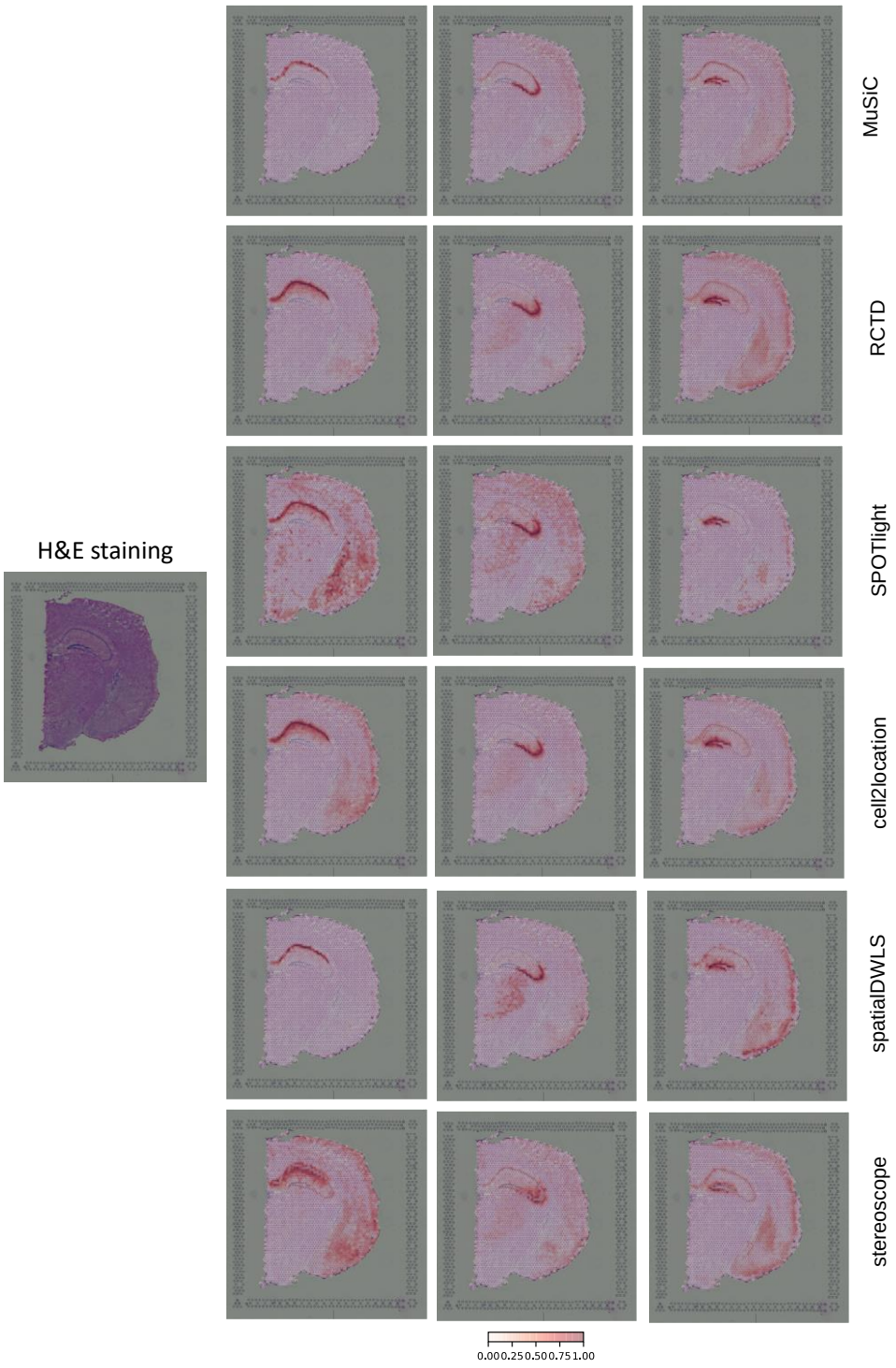


Figure S3.40 Scatter plot of cell type proportion distributions across spatial locations in the mouse hippocampus 10x Visium data when overlayed on top of H&E staining.

Left panel displays the original H&E staining figure. Specifically, cell type proportions estimated by MuSiC, RCTD, SPOTlight, cell2location, spatialDWLS, and stereoscope correspondingly. Color was shown to represent the 0-1 range of cell type proportions correspondingly.

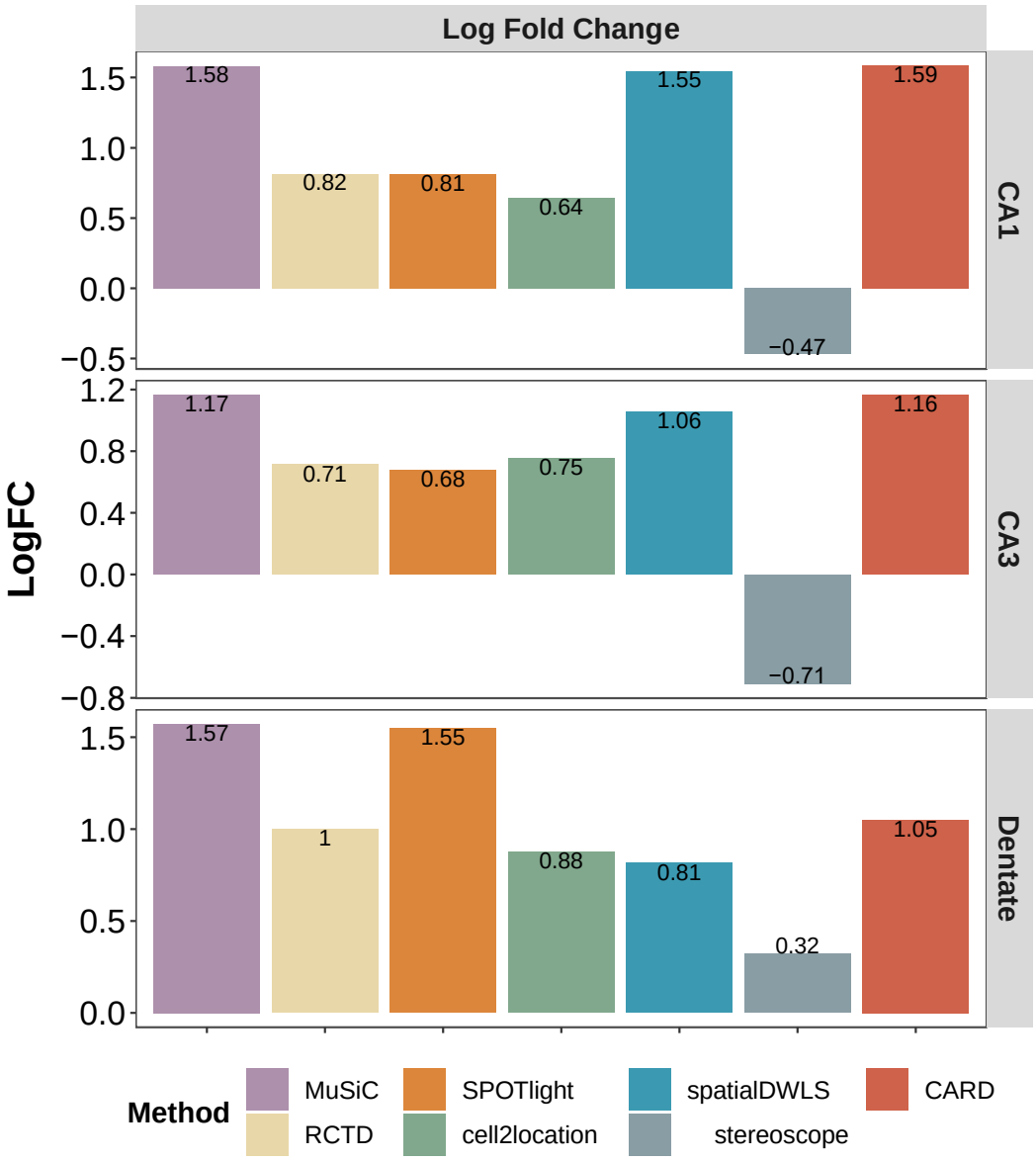
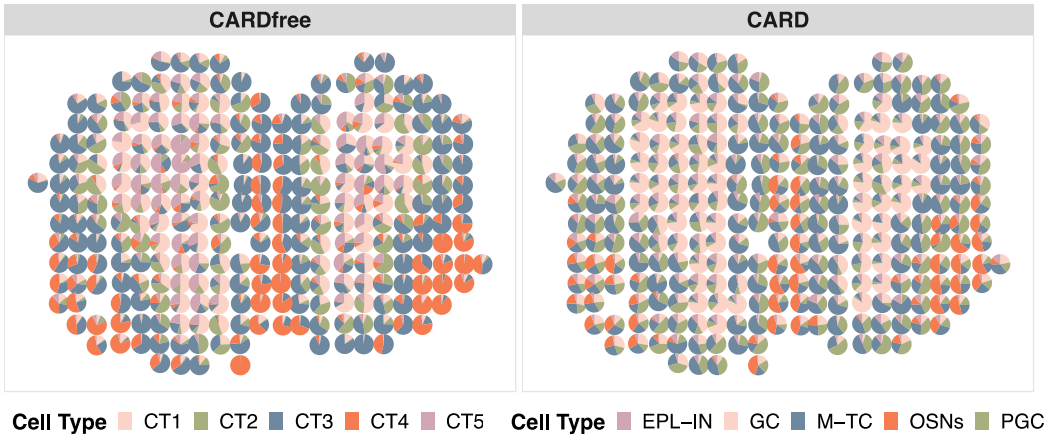


Figure S3.41 Specificity in the cell type proportion in each region compared with its corresponding boundary for all methods.

The estimated cell type compositions on CA1, CA3 and dentate gyrus from both CARD and MuSiC matched the corresponding structures on the H&E image, while those from the other methods appear to also occupy regions outside the expected structure boundaries (details see Supplementary Figure 73). For quantification, we calculated the mean cell type proportion within each structure, the mean cell type proportion on the boundary locations right adjacent to each structure (as shown in Supplementary Figure 74) and contrasted the two mean values by computing a ratio to serve as the location specificity measurement for the cell types. Consistent with visualization, we found that both CARD and MuSiC generated higher location specificity for all region-specific cell types as compared to the other methods.

(A)



(B)

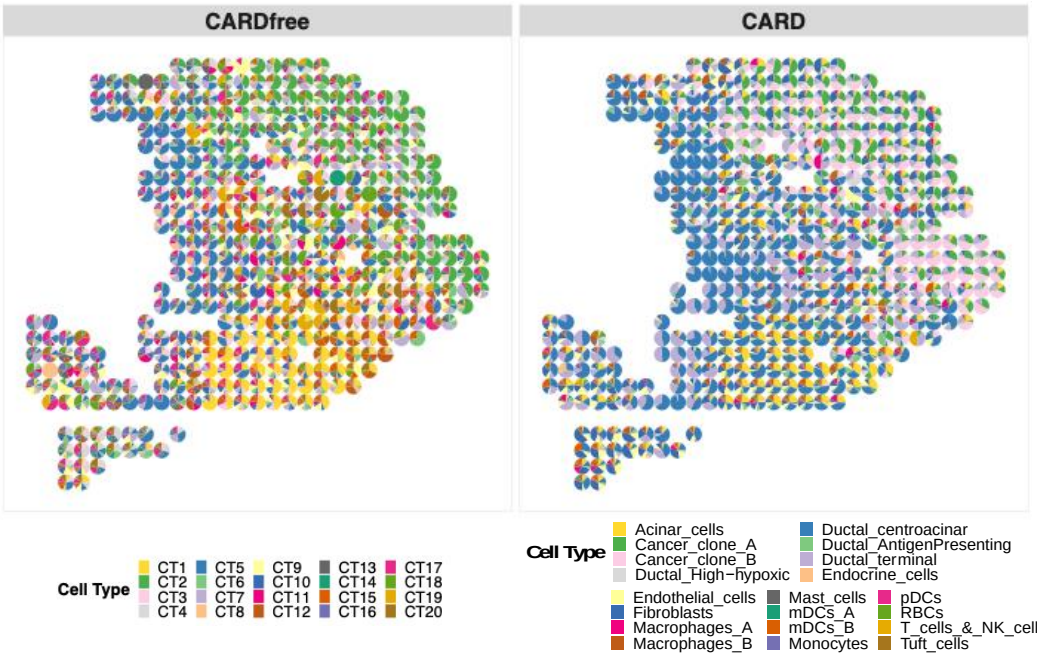


Figure S3.42 CARDfree generates comparable deconvolution results with CARD.

(A) when using known marker genes from [13] to deconvolute the mouse olfactory bulb spatial transcriptomics dataset. Spatial scatter pie plot displays inferred cell type composition on each spatial location from different deconvolution methods by CARDfree and CARD. The cell type CT1, CT2, CT3, CT4, and CT5 represent five clusters in the reference-free framework. **(B)** when using DE genes calculated from Seurat pipeline as the marker genes to deconvolute the human pancreatic ductal adenocarcinoma (PDAC) dataset. Spatial scatter pie plot displays inferred cell type composition on each spatial location from different deconvolution methods by CARDfree and CARD. The cell type CT1 to CT20 represents twenty clusters in the reference-free framework.

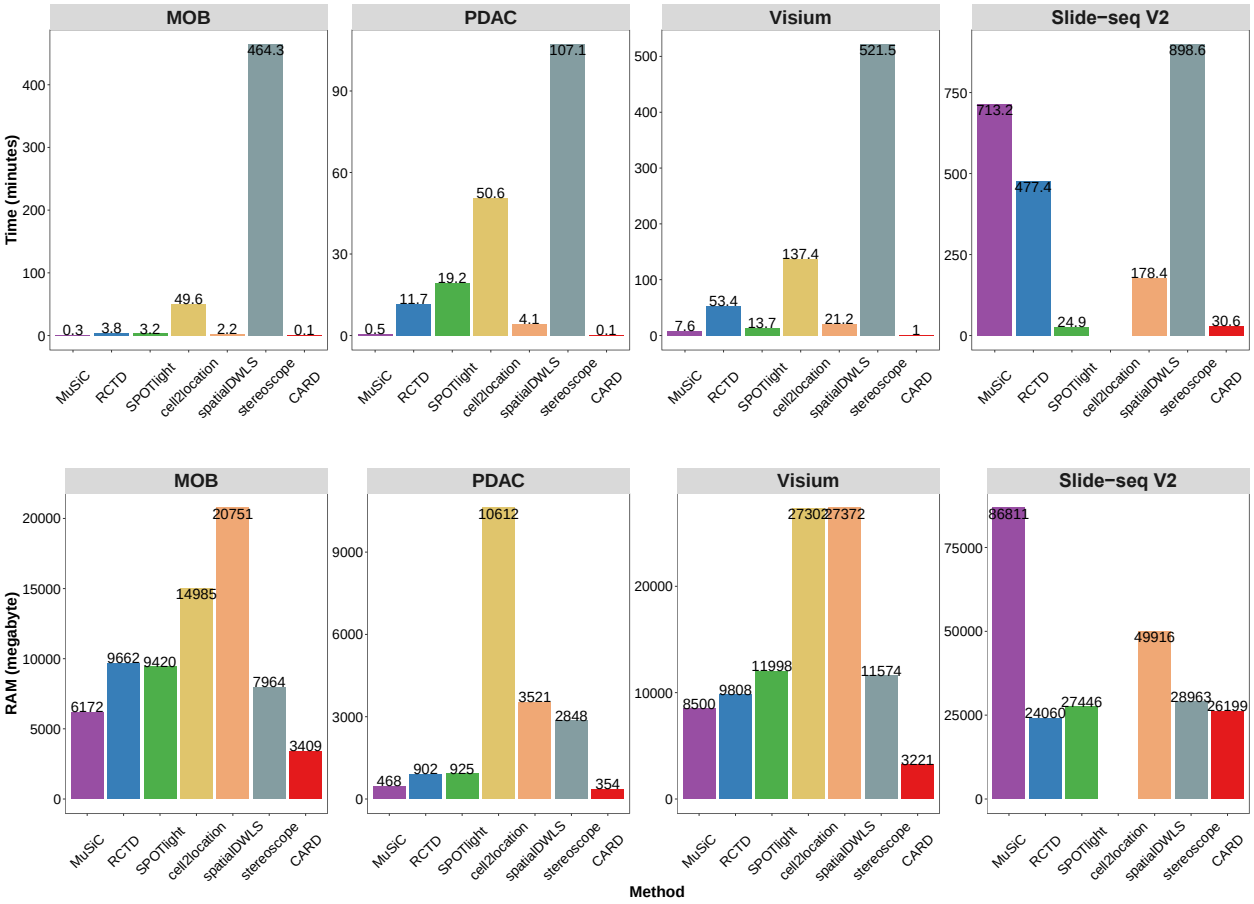


Figure S3.43 Computation time (minutes top panel) and peak memory usage (MB, bottom panel) for each deconvolution method on four real spatial transcriptomics datasets.

Computation of CARD, MuSiC and RCTD was performed on a single core of an Intel(R) Xeon(R) CPU E5-2683 v3 2.00GHz processor while computation of cell2location was performed on the GTX 1080 GPU processor. For better visualization of the time difference between CARD and other methods, the GPU time for the cell2location and stereoscope are plotted here. Note that, GTX 1080 GPU processor provides a total of 9 TFLOPS of performance while Intel(R) Xeon(R) CPU E5-2683 v3 provide a total of 0.448 TFLOPS of performance per core. So, the CPU time for cell2location and stereoscope should be calculated as the GPU time of cell2location * 9 / 0.448, almost 20 times slower than the GPU time shown in the figure. CARD is computationally fast and efficient compared with other methods in four real data applications. Note that we did not apply cell2location to the Slide-seq V2 data due to its heavy computational burden.

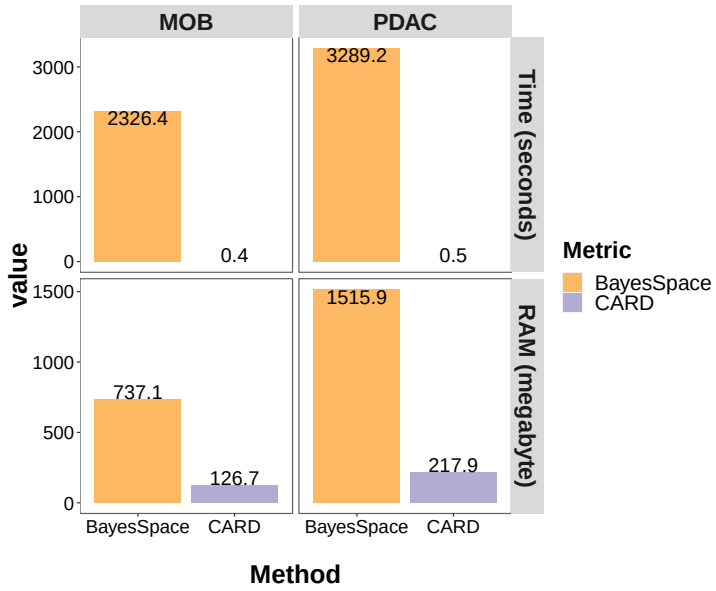


Figure S3.44 Computation time (seconds) and peak memory usage (MB) for CARD on constructing a refined spatial map.

Computation was performed on a single core of an Intel(R) Xeon(R) CPU E5-2683 v3 2.00GHz processor. grid = number of newly grided spatial locations. **(A)** Performance of CARD and BayesSpace on computational time (the first row) and peak memory usage (the second row) across different spatial-resolved transcriptomics. Here, the number of new locations for CARD is fixed to be 5000 while the number of refined locations for BayesSpace is dependent on the structure.

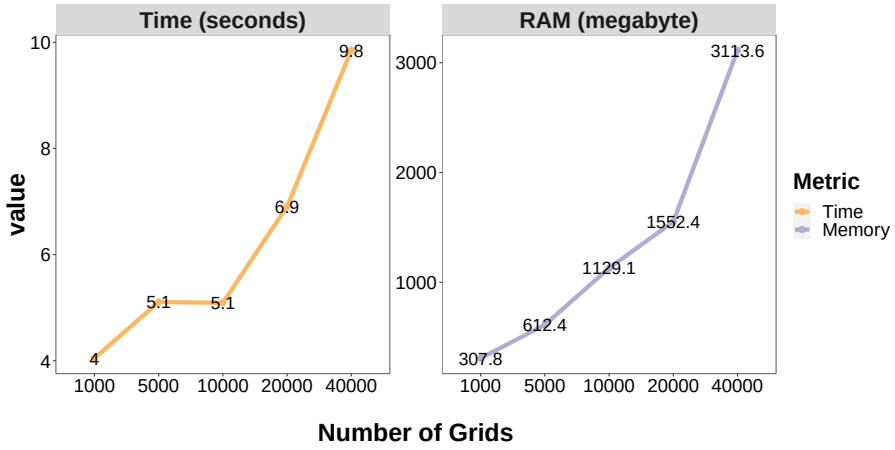


Figure S3.45 Computation time (seconds) and peak memory usage (MB) for CARD on constructing a refined spatial map on different number of new locations (new grid).

Computation was performed on a single core of an Intel Xeon L5420 2.50 GHz processor. grid = number of newly grided spatial locations. Performance of CARD on computational time (left panel) and peak memory usage (right panel) on the same datasets when the number of newly grided spatial locations = 1000, 5000, 10,000, 20,000, 40,000. Here, we used the largest dataset mouse hippocampus Slide-seq V2 data as an example.

3.7 Supplementary Tables

Dataset	Gene	Moran's I		Geary's C	
		P-value	Adjusted P-value	P-value	Adjusted P-value
MOB	*Penk*	*5.149e-57*	*2.060e-56*	*7.314e-03*	*2.926e-02*
	Apold1	*2.261e-03*	*9.046e-03*	*2.019e-02*	*8.075e-02*
	Cdhr1	*1.593e-01*	*6.372e-01*	*9.329e-01*	*1.000e+00*
	S100a5	*1.770e-98*	*7.082e-98*	*2.138e-12*	*8.554e-12*
	TM4SF1	*3.002e-201*	*1.801e-200*	*2.068e-11*	*1.241e-10*
	S100A4	*2.244e-40*	*1.346e-39*	*8.798e-02*	*5.279e-01*
PDAC	*TFF3*	*1.627e-62*	*9.761e-62*	*1.313e-02*	*7.879e-02*
	APOL1	*5.824e-168*	*3.495e-167*	*8.428e-05*	*5.057e-04*
	CRISP3	*7.914e-48*	*4.748e-47*	*2.852e-05*	*1.711e-04*
	CD248	*1.192e-03*	*7.152e-03*	*8.081e-01*	*1.000e+00*
	Wfs1	*<2.2e-308*	*<2.2e-308*	*7.452e-05*	*2.236e-04*
10X Visium	*Cpne4*	*<2.2e-308*	*<2.2e-308*	*4.813e-06*	*1.444e-05*
	C1ql2	*<2.2e-308*	*<2.2e-308*	*8.273e-01*	*1.000e+00*

Table S3.1 Results of Moran's I and Geary's C spatial statistical tests in real data applications.

P-value was determined by the one-sided Moran's I and one-sided Geary's C test and was adjusted by Bonferroni procedure.

Data set	Protocol	Year	# Genes	#Spots	scRNAseq /Spatial	Data Type	H&E Staining
Mouse Olfactory Bulb (Replicate 12)	ST	2016	16034	282	GSE121891	Spatial	*Link1
Human PDAC (PDAC-A)	ST (Different version)	2020	25753	428	GSE111672 Peng	Spatial	*Link2
Mouse Hippocampus Slide-seqV2	Slide-seqV2	2020	23265	53208	DropViZ	Spatial	NA

Mouse Hippocampus 10x Visium	10x Visium	2020	21143	2698	DropViZ	Spatial	*Link3
Mouse Brain Cortex	seqFISH+	2019	10000	523	GSE102827	Spatial	NA
Zeisel	10x Chromium	2018	27933	20585	Simulation (Scenario 1 2 3 4)	scRNA seq	NA
GSE109447	microwell-seq + Drop-seq	2019	32104	32684	Simulation (Scenario 5)	scRNA seq	NA
GSE121891	10x Chromium	2018	18560	21746	Mouse Olfactory Bulb	scRNA seq	NA
GSE111672 (PDAC-A)	inDrop	2020	19736	1926	Human PDAC	scRNA seq	NA
GSE111672 (PDAC-B)	inDrop	2020	19736	1733	Human PDAC	scRNA seq	NA
Peng	10x Chromium	2019	24005	57530	Human PDAC	scRNA seq	NA
Peng_Normal	10x Chromium	2019	24005	15544	Human PDAC	scRNA seq	NA
Peng_Tumor	10x Chromium	2019	24005	41986	Human PDAC	scRNA seq	NA
DropViZ (RCTD processed)	Drop-seq	2018	27953	1000	Mouse Hippocampus	scRNA seq	NA
GSE102827	inDrops	2017	25187	65539	Mouse Brain Cortex	scRNA seq	NA

Table S3.2 List of 5 spatially resolved transcriptomics datasets and 10 scRNA-seq datasets we used in our analysis.

Specifically, the links are provided in ref (Ma and Zhou 2022)

Chapter 4 Accurate and Efficient Integrative Reference-Informed Spatial Domain Detection for Spatial Transcriptomics

4.1 Abstract

Spatially resolved transcriptomics (SRT) studies are becoming increasingly common and increasingly large, offering unprecedented opportunities to characterize the spatial and functional organization of complex tissues. Here, we introduce a computational method, IRIS, that characterizes the spatial organization of complex tissues through accurate and efficient detection of spatial domains. IRIS uniquely leverages the widespread availability of single-cell RNA-seq data for reference-informed spatial domain detection, integrates multiple SRT tissue slices jointly while explicitly considering correlation both within and across slices, produces biologically interpretable spatial domains, and benefits from multiple algorithmic innovations for highly scalable computation. We demonstrate the advantages of IRIS through in-depth analysis of four SRT datasets from different technologies across various tissues, species, and spatial resolutions. IRIS attains an unprecedented 58% ~ 1,083% accuracy gain over existing methods in a gold standard dataset with known ground truth. Furthermore, IRIS is 8.5 ~ 134.7 times faster than existing methods in moderate-sized datasets and is the only method applicable to large-scale SRT datasets, including the recent stereo-seq and 10x Xenium. As a result, IRIS uncovers the fine-scale structures of brain regions, reveals the spatial heterogeneity of distinct tumor microenvironments, and characterizes the structural changes of the seminiferous tubes in the testis associated with diabetes, all at a speed and accuracy unachievable by existing approaches.

4.2 Introduction

Spatially resolved transcriptomics (SRT) are a set of recently developed technologies that enable the profiling of gene expression on a tissue with spatial localization information. These SRT technologies include both imaging-based approaches, which rely on either single molecular *in situ* hybridization (e.g., MERFISH (Chen et al. 2015, Vizgen 2021), seqFISH (Lubeck et al. 2014), seqFISH+ (Lubeck et al. 2014), 10x Xenium (Janesick et al. 2022)) or *in situ* sequencing (e.g., STARmap (Wang et al. 2018a), FISSEQ (Lee et al. 2015a)), and next-generation sequencing based approaches, which include Spatial Transcriptomics (ST) (Ståhl et al. 2016), 10x Visium (10XGenomics), Slide-seq (Rodriques et al. 2019) (Stickels et al. 2021), Stereo-seq (Chen et al. 2022), and Seq-Scope (Cho et al. 2021), to name a few. All together, these SRT technologies have provided unprecedented opportunities for investigating and characterizing the transcriptomic and cellular landscape of complex tissues (Tian et al. 2022).

A major analytic task of SRT studies is to characterize the spatial organization of complex tissues in the form of spatial domain detection (Moses and Pachter 2022, Tian et al. 2022, Rao et al. 2021). Tissues are complex cellular ecosystems that consist of many spatially organized and functionally distinct anatomical domains and microenvironments, each characterized by unique local features with varying cell type compositions and transcriptomic heterogeneity. The spatial organization of tissues in the form of local domains facilitates how different cell types coordinate with each other in carrying out tissue functions in development, homeostasis, communication, repair, and signaling responses. Consequently, detecting spatial domains on the tissue in SRT studies can facilitate our understanding of the spatial and functional organization of a normal tissue and reveal how alterations in the tissue structure may underlie disease etiology. Several computational methods have been recently developed for detecting spatial domains in SRT (Hu et

al. 2021, Zhao et al. 2021, Fu et al. 2021, Dries et al. 2021, Moses and Pachter 2022, Palla et al. 2022, Zhu et al. 2018, Tian et al. 2022, Rao et al. 2021, Li and Zhou 2022, Shang and Zhou 2022). Examples include the graph convolutional network method spaGCN (Hu et al. 2021), the Potts model based methods such as hidden Markov random field (HMRF) (Zhu et al. 2018), BayesSpace (Zhao et al. 2021) and BASS (Li and Zhou 2022), the autoencoder based method SEDR (Fu et al. 2021), and the hybrid deep learning and Bayesian modeling framework Maple (Allen et al. 2022). Unfortunately, almost all existing methods directly rely on transcriptomic heterogeneity to disentangle the spatial domains. However, transcriptomic heterogeneity across spatial domains is only a secondary feature of the spatial domains, as it is the direct consequence of the unique cell type composition underlying each spatial domain. Modeling the secondary feature of transcriptomic heterogeneity instead of the primarily feature of cell type composition for spatial domain detection is not ideal, as such approach makes it difficult to characterize the cellular landscape of the tissue, reduces the accuracy and interpretability of the detected spatial domains, and as will be shown here, often leads to the identification of biologically irrelevant structures.

Here, we present an alternative strategy for detecting spatial domains in SRT studies. Specifically, we directly model the primary feature of cell type compositional heterogeneity across spatial locations and use it to segment the tissue into multiple biologically relevant spatial domains, each of which is now characterized by a distinct composition of cell types. This alternative strategy has four important benefits. First, it allows us to directly characterize the cellular landscape of the tissue and identify biologically interpretable spatial domains, thus facilitating the understanding of the cellular mechanism underlying tissue function. Second, it offers a framework for integrating cell type specific transcriptomic profiles (Biancalani et al. 2021, Moriel et al. 2021, Ma and Zhou 2022, Li et al. 2023) obtained from readily available single-cell RNA-seq (scRNA-seq) data into

SRT. This enables us to leverage the vast amount of data gathered in scRNA-seq studies, which are almost always available for the tissue samples used in the SRT study, leading to potentially substantial accuracy gain for spatial domain detection. Third, by focusing on the primary feature of cell type composition, the alternative strategy naturally provides an anchoring point for integrating SRT data across multiple tissue slices or multiple samples that are commonly collected in recent SRT studies but not yet analyzable by most existing domain detection methods. In particular, because multiple slices from the same tissues often contain a similar set of spatial domains characterized by similar cell type compositions, anchoring them on the shared domain-specific cell type composition allows us to borrow the similarity information in the spatial domain characteristics from multiple tissue slices to further enhance the performance of spatial domain detection. Finally, by transforming the task of spatial domain detection based on transcriptomic heterogeneity into a conceptually simpler task of characterizing domains based on domain-specific cell type compositions, our strategy also paves the way for various computationally efficient algorithms, making it feasible to detect spatial domains in very large-scale SRT datasets that are currently intractable for almost all existing domain detection methods.

While the above alternative strategy is biologically intuitive, efficiently implementing it, however, proves to be non-trivial. As we will show below, a naive application of this alternative strategy, that first infers cell type compositions on the tissue and then conducts tissue segmentation via clustering on the inferred compositions, does not produce spatial domains as accurate as the current state-of-the-art methods. Consequently, to fully harness the potential of this alternative strategy, we have developed a novel computational method called the Integrative and Reference-Informed tissue Segmentation (IRIS). IRIS is accurate, scalable, and robust for spatial domain detection. We demonstrate the advantages of IRIS by analyzing four SRT datasets from different

tissues and species that were sequenced using various techniques including 10x Visium, Slide-seq, Stereo-seq, and 10x Xenium that just came out months ago. Our results show that IRIS considerably surpassed the state-of-the-art methods for spatial domain detection with significantly greater computational efficiency. The unparalleled accuracy and computational gains delivered by IRIS make it an indispensable tool for integrated tissue segmentation in large-scale SRT datasets that are rapidly accumulating.

4.3 Results

4.3.1 Method overview

IRIS is described in **Methods**, with its method schematic shown in **Figure 4.1**. Briefly, IRIS is a reference-informed integrative method for detecting spatial domains on multiple tissue slices from spatial transcriptomics with spot-level, single-cell level, or subcellular level resolutions. IRIS is based on the idea that each spatial domain on the tissue is characterized by a unique composition of cell types and that similar composition is observed for the same spatial domain across different slices of the same tissue (Stoltzfus et al. 2020, Bove et al. 2017). Consequently, IRIS integrates a reference scRNA-seq data to inform and characterize the cell type composition on the tissue of spatial transcriptomics for accurate and interpretable spatial domain detection. In the process, IRIS accommodates cell type compositional similarity across locations within the same slice and across different slices on the same domain to borrow information both within and between tissue slices for integrative and accurate spatial domain detection. Importantly, IRIS comes with an efficient optimization framework with multiple algebraic innovations for scalable computation and can easily handle multiple spatial transcriptomics datasets with millions of spatial locations and tens of thousands of genes. IRIS is implemented as an open-source R package, freely available at www.xzlab.org/software.html.

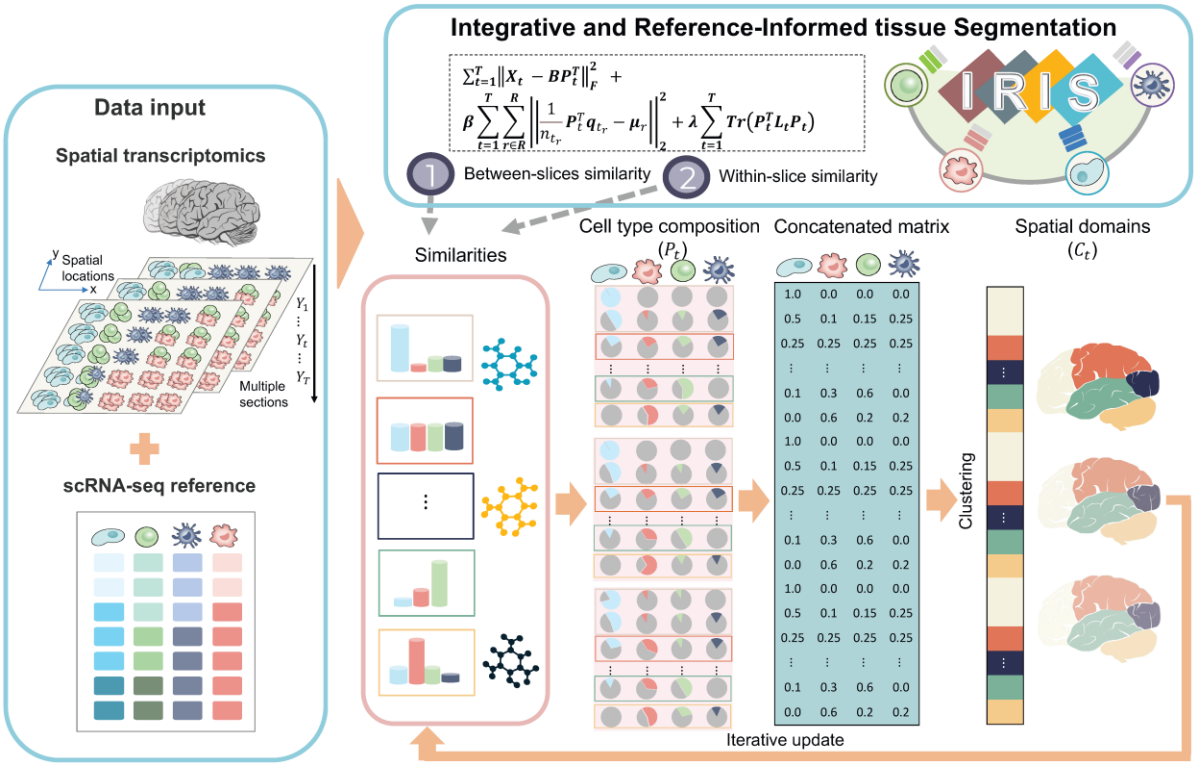


Figure 4.1 Schematic overview of IRIS.

IRIS is an accurate and efficient integrative reference-informed segmentation method for detecting spatial domains on multiple tissue slices across a range of SRT technologies. As it is shown in the left box, IRIS requires two types of input: a SRT data measured on multiple tissue slices with spatial localization information, and a scRNA-seq reference data measured on the same tissue with cell type specific gene expression information. With these two data inputs, IRIS builds upon the fact that multiple slices from the same tissues often contain a similar set of spatial domains characterized by similar cell type compositions and that neighboring locations within the same tissue usually share similar cell type compositions. In the process, IRIS accommodates cell type compositional similarity across locations within the same slice and across different slices on the same domain to borrow information both within and between tissue slices. By encouraging these two similarities, IRIS first updated the cell type compositions on each slice, then concatenate the cell type composition into a concatenated matrix, and then a K-means clustering algorithm is performed on the concatenated cell type composition matrix to update the spatial domains, then iteratively update the composition matrix and the spatial domain label to achieve optimal performance.

4.3.2 Human dorsolateral prefrontal cortex 10x Visium data

To benchmark the performance of IRIS and other spatial domain detection methods, we first examined the gold standard DLPFC data by 10x Visium, which contains the ground truth spatial domains (Maynard et al. 2021, Pardo et al. 2022). Specifically, this data contains 12 tissue slices from three neurotypical adult donors, with transcriptomic measurements on 33,538 genes and 3,460 ~ 4,789 spatial locations. We obtained a scRNA-seq data from 10x Chromium on the post-mortem brain tissue with 44 cell types to serve as the reference (Mathys et al. 2019). Note that the scRNA-seq reference is from an external study with samples unmatched to the SRT data and with cell types and states potentially different from that in SRT. The DLPFC contains seven spatial domains that include six cortical layers and white matter. We used the domain annotations provided by the histologist in the original study as the ground truth and evaluated the accuracy of the spatial domains detected by different methods using adjusted Rand index (ARI) following (Zhao et al. 2021, Hu et al. 2021) (**APPENDIX C.1**). In particular, we compared IRIS with state-of-the-art spatial domain detection methods that belong to three distinct categories (details in **Methods**): (1) methods that analyze a single slice one at a time (single-slice methods): spaGCN (with or without image), BayesSpace, and SEDR; (2) methods that first infer cell type compositions on the tissue and then conduct tissue segmentation via clustering on the inferred compositions (deconvolution-based methods): CARD_kmeans, RCTD_kmeans; and (3) methods that jointly analyze multiple slices (multi-slice methods): BASS, Maple, and BayesSpaceJoint. We first examined a simple setting where we analyzed tissue slices that come from the same donor, with similar tissue structures shared across slices. In the analysis, IRIS correctly detects the layered structures of the prefrontal cortex (**Figure 4.2**), with much higher accuracy than the other methods. Specifically, IRIS achieved a median ARI of 0.71 across slices, representing 58%-1,083%

accuracy gain compared to the other methods (0.45 for the second-best method spaGCN with H&E image; **Figure 4.2B**). In addition, IRIS clearly captures the sandwich-like structure of the cortical layers 1-3 that is missed by all the other methods and detects the deep layers 4-6, for which none of the other methods detect. Regardless of the pre-specified number of spatial domains, IRIS always performs the best and the comparative results remain consistent (**Figure 4.2C**). Note that the accuracy achieved by IRIS in this gold standard data has never been attained in any previous studies.

Next, we evaluated more challenging analytic settings. First, we examined settings where there are missing or misclassified cell types (details in **Methods**) in the scRNA-seq reference. In this setting, IRIS remains superior compared to the other methods (median ARI = 0.66 across all settings), with a 47% -1,000% accuracy gain compared to the other methods, regardless of whether there are missing cell types or mis-classified cell types (**Figure S4.1**) in the scRNA-seq reference. Second, we analyzed tissue slices from different donors, where the tissue structures from different slices display distinct shapes. Here, IRIS again detects the spatial domains accurately (median ARI = 0.71), representing 51%-1,083% accuracy gain compared to the other methods (0.47 for the second-best method spaGCN without H&E image; **Figure S4.2**). The relative performance of different methods remains the same in both baseline and challenging settings in terms of ARI (**Figure 4.2C**, **Figure S4.1 - Figure S4.2**).

We examined the layer specific marker genes to validate the detected spatial domains and provide additional evidence to further quantify the accuracy of different methods. We reasoned that a good spatial domain detection method would yield accurate tissue structures that captured the expected domain-specific expression pattern for the marker genes. Therefore, for each layer-

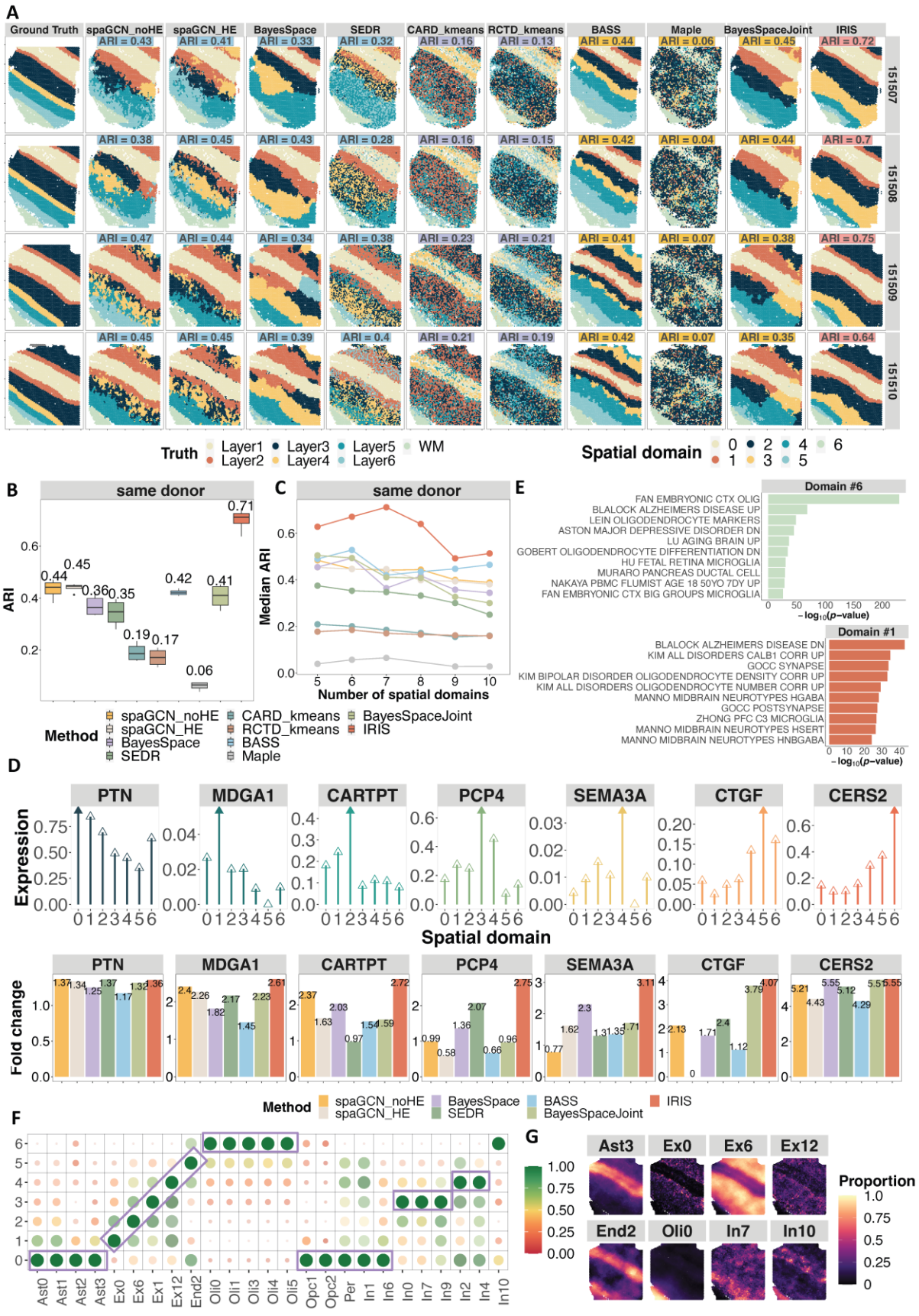

A
Ground Truth
spaGCN_noHE
spaGCN_HE
BayesSpace
SEDR
CARD_kmeans
RCTD_kmeans
BASS
Maple
BayesSpaceJoint
IRIS
151507
151508
151509
151510
Truth
Layer1
Layer2
Layer3
Layer4
Layer5
Layer6
WM
Spatial domain
B
same donor
ARI
Method
C
same donor
Median ARI
Number of spatial domains
E
Domain #6
FAN EMBRYONIC CTX OLIG
BLALOCK ALZHEIMERS DISEASE UP
LEIN OLIGODENDROCYTE MARKERS
ASTON MAJOR DEPRESSIVE DISORDER DN
LU AGING BRAIN UP
GOBERT OLIGODENDROCYTE DIFFERENTIATION DN
HU FETAL RETINA MICROGLIA
MURARO PANCREAS DUCTAL CELL
NAKAYA PBMC FLUMIST AGE 18 50YO 7DY UP
FAN EMBRYONIC CTX BIG GROUPS MICROGLIA
Domain #1
BLALOCK ALZHEIMERS DISEASE DN
KIM ALL DISORDERS CALB1 CORR UP
GOCC SYNAPSE
KIM BIPOLAR DISORDER OLIGODENDROCYTE DENSITY CORR UP
KIM ALL DISORDERS OLIGODENDROCYTE NUMBER CORR UP
MANNO MIDBRAIN NEUROTYPES HGABA
GOCC POSTSYNAPSE
ZHONG PFC C3 MICROGLIA
MANNO MIDBRAIN NEUROTYPES HSERT
MANNO MIDBRAIN NEUROTYPES HNBGABA
D
PTN
MDGA1
CARTPT
PCP4
SEMA3A
CTGF
CERS2
Expression
Spatial domain
Fold change
F
G
Ast3
Ex0
Ex6
Ex12
End2
Oli0
In7
In10
Proportion

Figure 4.2 Analyzing the human DLPFC 10x Visium data.

(A) Spatial domains detected by IRIS, spaGCN without HE, spaGCN with HE, BayesSpace, SEDR, CARD_kmeans, RCTD_kmeans, BASS, Maple, and BayesSpaceJoint. Results are shown in the baseline analysis setting where the tissue slices are from the same donor (denoted as same donor). Ground truth tissue regions of the human prefrontal cortex are obtained from the original DLPFC study (left panel). Clustering accuracy of different methods in recapitulating the true tissue domains is measured by ARI, with a higher ARI indicating higher accuracy. **(B)** Boxplots display the clustering accuracy measured in the format of ARI on the tissue slices applied in the baseline setting (same donor). Compared spatial domain detection methods (x-axis) include (1) single slice method: spaGCN without HE image (yellow), spaGCN with HE image (beige), BayesSpace (purple), and SEDR (green); (2) deconvolution-based method: CARD_kmeans (blue), RCTD_kmeans (tape); (3) multiple-slice methods: BASS (lake blue), Maple (grey), BayesSpaceJoint (matcha), and IRIS (red). **(C)** Clustering performance of different methods when varying the pre-specified number of spatial domains. The median ARI across all slices in either setting was calculated to measure spatial domain detection accuracy. **(D)** Lollipop plots (top) display the mean expression of seven important marker genes (y-axis) in each spatial domain identified by IRIS (x-axis). Solid arrow indicates the highest enrichment value. Bar plots (bottom) display the fold change of each marker gene in the expected domain versus other domains, where a higher value indicates higher domain detection accuracy. **(E)** Top enriched gene sets in the selected spatial domains detected by IRIS in gene set enrichment analysis. **(F)** Heatmap plot displays the estimated mean cell type proportion for each cell type in each spatial domain detected by IRIS. The color scale was normalized to 0-1 range. **(G)** Spatial scatter plot displays the spatial distribution of IRIS estimated cell type proportion for each cell type across spatial locations. For D - G, the results are shown for example slice 151509 in the baseline analysis.

specific marker gene in turn, we first calculated its mean expression in the inferred spatial domain where the marker gene is expected to be enriched and then contrasted it to the mean expression in the other domains by calculating an enrichment fold change (**Methods**). In the analysis, we found that the layers identified by IRIS are enriched with known layer marker genes including *PTN* (Mentlein and Held-Feindt 2002) (layer 1; spatial domain 0), *MDGA1* (Maynard et al. 2021, Uchida et al. 2011) (layer 2; spatial domain 1), *CARTPT* (Maynard et al. 2021) (layer 3; spatial domain 2), *PCP4* (Tang et al. 2015) (layer 4; spatial domain 3), *SEMA3A* (Chen et al. 2008) (layer 5; spatial domain 4), *CTGF* (Zeng et al. 2012) (layer 6; spatial domain 5), CERS2 (Sampaio-Baptista et al. 2020) (white matter; spatial domain 6) (**Figure 4.2D**). The enrichment pattern of the marker genes in the corresponding domains for IRIS is 16%-374% stronger than the other methods.

We carefully examined the spatial domains detected by IRIS and performed downstream analysis to characterize the transcriptomic and cellular landscape of the tissue domains (**Methods**). First, we carried out differential expression (DE) analysis to identify genes that are specifically expressed in different spatial domains. We identified a median of 188 DE genes across the seven domains, including both previously known layer-specific marker genes (i.e., *CARTPT* (Maynard et al. 2021), *PCP4* (Tang et al. 2015)) and novel marker genes (i.e., *APOE*, *NRGN*, and *PLP1* (Tabata 2015)) (**Figure S4.3**). For example, *APOE* is an identified DE gene in spatial domain #0 (layer 1) and encodes the apolipoprotein E that plays as a central role in lipid metabolism (Chung et al. 2016). *PLP1* is an identified DE gene in spatial domain #6 (white matter) and is closely related to myelination occurring there (Ocklenburg et al. 2019). Second, we performed gene set enrichment analysis on the detected DE genes. We found that the detected DE genes are highly enriched in synapse signaling pathways, post-synapse signaling, and Alzheimer's disease gene sets, all of which are hallmarks of brain functionality (**Figure 4.2E**, **Figure S4.4**). Third, we carefully examined the domain specific cell type compositions detected by IRIS. We found that a mixture of astrocyte subtypes, oligodendrocyte progenitor subtypes, and pericytes are highly colocalized in the spatial domain (#0) corresponding to the cortical layer 1 (**Figure 4.2F**) while different excitatory neuron subtypes are enriched in the spatial domains (# 1 to #4) corresponding to the cortical layers 2 to 5. In addition, we found that various oligodendrocyte subtypes are enriched in the white matter (#6) while inhibitory neuronal subtypes are mainly enriched in the deeper layers 4-5 (Swanson and Maffei 2019) (#3 to #4). Similar results are observed by examining the spatial distribution of the representative cell types (**Figure 4.2G**).

4.3.3 Mouse spermatogenesis Slide-seq data

Next, we analyzed the mouse spermatogenesis Slide-seq data (Chen et al. 2021) collected on testis, which consists of well-defined tissue structures in the form of seminiferous tubes that can be easily visualized to evaluate method performance. This study sequenced six tissue slices, one from each of the three diabetic (ob/ob) mice and three wildtype (WT) mice, containing 23,515 ~ 24,450 genes and 27,194 ~ 42,776 spatial locations. We obtained the scRNA-seq reference from an external study by Drop-seq technology (Green et al. 2018) which measured gene expression on six batches of adult mice. We followed the original publication to primarily focus on the analysis of one slice from a WT mouse and one slice from an ob/ob mouse, with 81,982 spatial locations in total, to investigate the structural changes of the seminiferous tubes underlying diabetes-induced male infertility. Besides the primary analysis, we also performed joint analysis on three slices that come from either the three WT mice or the three diabetic mice as supplementary examples.

Spermatogenesis is the biological process that involves five successive stages: spermatogonia (SPG), primary spermatocytes (SPC), secondary SPC, spermatids that include both round spermatids (RS) and elongating spermatids (ES), and spermatozoa (Linn et al. 2021), all organized by their spatial localization with respect to the seminiferous tubules (**Figure 4.3A**). In the analysis of the spermatogenesis data, we found that the spatial domains detected by IRIS accurately depict the expected structure of the seminiferous tubules in the testis (**Figure 4.3B**). Specifically, in the WT mouse (WT3_Puck7 slice), IRIS captured a spatial domain (#0) that is primarily located in the center of the seminiferous tubules where RS resides; a domain (#2) that is located in the peripheral of the seminiferous tubules where ES resides; two domains (#3 & #4) that are colocalized with domain #2 and likely represent the primary and secondary SPC; and a domain (#1) that captures the interstitial space where spermatogonia, leydig cells, sertoli cells, endothelial cells, and macrophage cells all locate. In contrast, the spatial domains detected by

spaGCN, BayesSpace, and BayesSpaceJoint all display a chaotic pattern that does not resemble the known seminiferous tubular structures of the testis including RS and SPC. Both SEDR and BASS can capture the general structure of ES colocalized with SPC but were unable to identify the circular shaped RS domain and incorrectly detected an ES domain that is much smaller than expected (**Figure 4.3B**). The superior performance of IRIS also applies to the ob/ob mouse (Diabetes2_Puck10 slice), where only IRIS can accurately identify the position of the ES domain and capture its reduced size as compared to the WT mice, supporting disrupted spatial cellular architecture of the seminiferous tubules in diabetic mice (Alves et al. 2013, Chen et al. 2021, Ballester et al. 2004). Quantification by spatial CHAOS score (details see **APPENDIX C.1**) shows that the spatial domains detected by IRIS and BASS (median CHAOS = 0.014 for IRIS and 0.014 for BASS) display better spatial continuity and compactness than the other methods (**Figure 4.3C;** 0.015 for the second-best method SEDR), regardless of the pre-specified number of spatial domains (**Figure 4.3C**) and regardless of which combinations of slices (i.e., slices from all WT mice, or from all ob/ob mice) were used (**Figure S4.5**).

We examined the expression pattern of several known spermatogenesis-related genes (Chen et al. 2021) to further validate and quantify the identified spatial domains by IRIS (**Figure 4.3D**). We found that the ES and RS related domains (#2 and #0) are enriched with the corresponding marker genes including *Prm1* (Ren et al. 2021), *Tnp1* (Yan et al. 2010) for ES and *Tekt2* (Lehtiniemi and Kotaja 2017), and *Tex36* (Wang et al. 2022b) for RS. Two domains identified by IRIS, likely corresponding to the primary spermatocytes (#3) and secondary spermatocytes (#4), are enriched with the marker gene *Mllt10* (ProteinAtlas) and *Aurka* (Nguyen and Schindler 2017), respectively. Specifically, gene *Mllt10* was previously found to be highly ex-

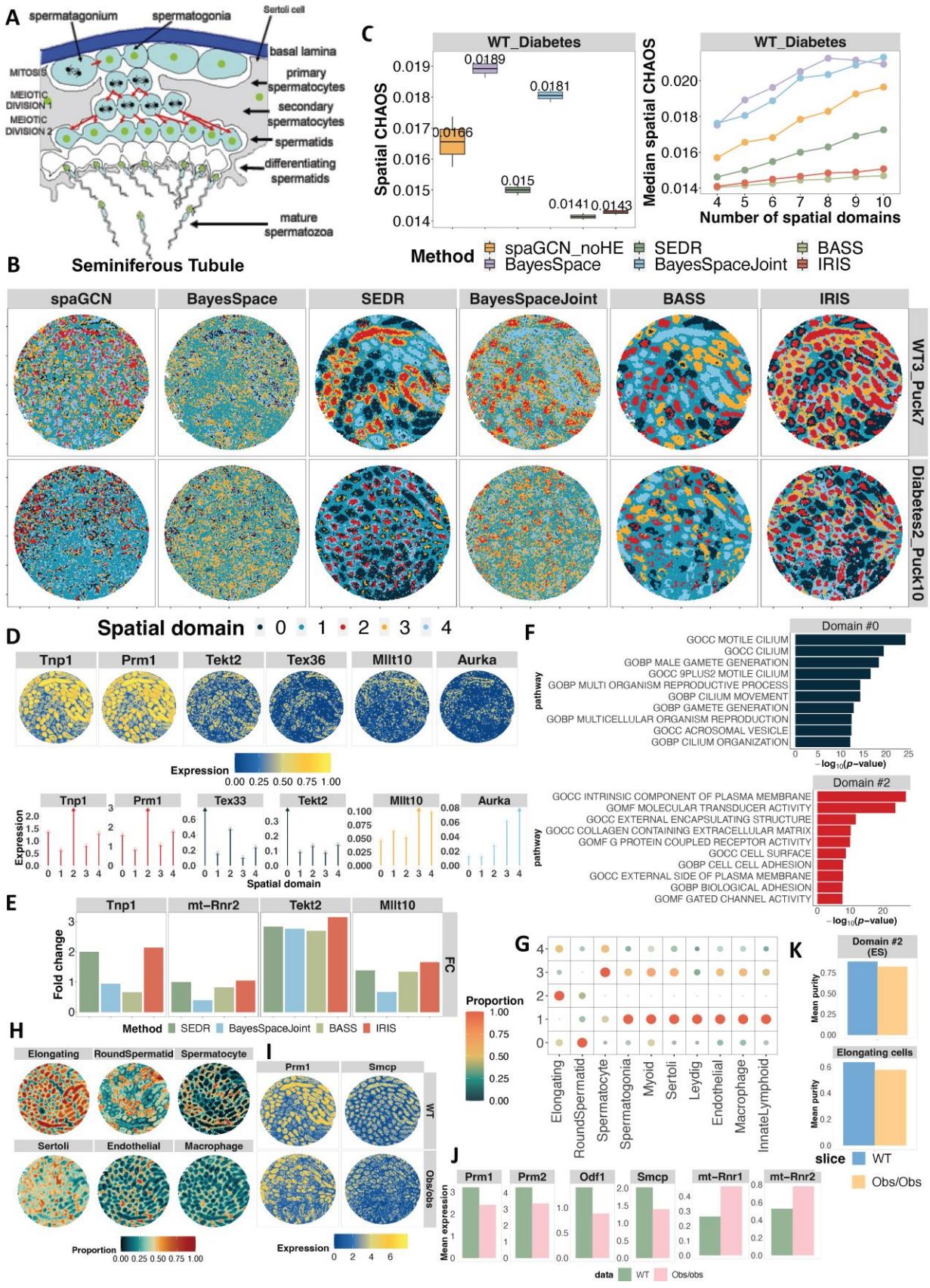

A
spermatagonium
spermatogonia
Sertoli cell
basal lamina
MITOSIS
primary spermatocytes
MEIOTIC DIVISION 1
secondary spermatocytes
MEIOTIC DIVISION 2
spermatids
differentiating spermatids
mature spermatozoa
B
Seminiferous Tubule
C
WT_Diabetes
Spatial CHAOS
0.0189
0.0181
0.0166
0.015
0.0141
0.0143
Median spatial CHAOS
Number of spatial domains
Method
spaGCN_noHE
BayesSpace
SEDR
BayesSpaceJoint
BASS
IRIS
spaGCN
BayesSpace
SEDR
BayesSpaceJoint
BASS
IRIS
WT3_Puck7
Diabetes2_Puck10
D
Spatial domain 0 1 2 3 4
Tnp1
Prm1
Tekt2
Tex36
Mllt10
Aurka
Expression
Tnp1
Prm1
Tex33
Tekt2
Mllt10
Aurka
Expression
Spatial domain
E
Tnp1
mt−Rnr2
Tekt2
Mllt10
Fold change
FC
Method SEDR BayesSpaceJoint BASS IRIS
F
Domain #0
GOCC MOTILE CILIUM
GOCC CILIUM
GOBP MALE GAMETE GENERATION
GOCC 9PLUS2 MOTILE CILIUM
GOBP MULTI ORGANISM REPRODUCTIVE PROCESS
GOBP CILIUM MOVEMENT
GOBP GAMETE GENERATION
GOBP MULTICELLULAR ORGANISM REPRODUCTION
GOCC ACROSOMAL VESICLE
GOBP CILIUM ORGANIZATION
pathway
−log10(p−value)
Domain #2
GOCC INTRINSIC COMPONENT OF PLASMA MEMBRANE
GOMF MOLECULAR TRANSDUCER ACTIVITY
GOCC EXTERNAL ENCAPSULATING STRUCTURE
GOCC COLLAGEN CONTAINING EXTRACELLULAR MATRIX
GOMF G PROTEIN COUPLED RECEPTOR ACTIVITY
GOCC CELL SURFACE
GOBP CELL CELL ADHESION
GOCC EXTERNAL SIDE OF PLASMA MEMBRANE
GOBP BIOLOGICAL ADHESION
GOMF GATED CHANNEL ACTIVITY
pathway
−log10(p−value)
G
Proportion
Elongating
RoundSpermatid
Spermatocyte
Spermatogonia
Myoid
Sertoli
Leydig
Endothelial
Macrophage
InnateLymphoid
H
Elongating
RoundSpermatid
Spermatocyte
Sertoli
Endothelial
Macrophage
Proportion
I
Prm1
Smcp
WT
Obs/obs
Expression
J
Prm1
Prm2
Odf1
Smcp
mt−Rnr1
mt−Rnr2
Mean expression
data WT Obs/obs
K
Domain #2 (ES)
Mean purity
Elongating cells
Mean purity
slice WT Obs/Obs

Figure 4.3 Analyzing the mouse spermatogenesis Slide-seq data.

(A) The diagram displays the biology process underlying spermatogenesis, which contains five successive stages: spermatogonia (SPG), primary spermatocytes (SPC), secondary spermatocytes (SPC), spermatids which include round spermatids (RS) and elongating spermatids (ES), and spermatozoa**. (B)** Spatial domains detected by IRIS, spaGCN, BayesSpace, SEDR, BASS, and BayesSpaceJoint in the integrative analysis of slices from wildtype and diabetes mice (denoted as "WT_Diabetes" analysis). **(C)** Boxplots display CHAOS values for different methods, which measure the spatial continuity and compactness of the detected spatial domains from different methods (left panel). Compared spatial domain detection methods (x-axis) include spaGCN (yellow), BayesSpace (purple), and SEDR (green), BASS (lake blue), BayesSpaceJoint (matcha), and IRIS (red). Line plots display CHAOS values when varying the pre-specified number of spatial domains. The median CHAOSs across all slices was used. **(D)** Scatter plots (top) display the spatial distribution of important spermatogenesis related marker genes. Lollipop plots (bottom) display the mean expression of important marker genes (y-axis) in each spatial domain identified by IRIS (x-axis). Solid arrow indicates the highest enrichment value. **(E)** Bar plots display the fold change of the marker gene expression in the expected domain versus other domains, where a higher value indicates better spatial domain detection accuracy. **(F)** Top enriched gene sets in selected spatial domains (e.g., #0 & #2) **(G)** Heatmap plot displays the estimated mean cell type proportion for each cell type in each spatial domain detected by IRIS. The color scale was normalized to 0-1 range. **(H)** Spatial scatter plot displays the spatial distribution of IRIS estimated cell type proportion for each cell type across spatial locations. **(I)** Comparison of the spatial pattern of ES marker genes in the WT mice and ob/ob (diabetic) mice. **(J)** Bar plot displays the mean expression of ES marker genes in the ES domain (#2; the first four panels) and the mean expression of mitochondrial genes in domain #1 (the latter two panels) detected by IRIS in WT (olive) and ob/ob (pink) slices in the "WT_Diabetes" analysis. **(K)** Purity score of ES domain (top panel) and ES cells (bottom panel) in the WT (blue) and ob/ob (yellow) mice. For D – G, the results are shown for the example WT slice (WT3_Puck7) in the main analysis.

-pressed in early spermatids and pachytene/diplotene spermatocytes and gene *Aurka* plays an important role in the secondary spermatocyte stage during the metaphase II in meiosis II (Nguyen and Schindler 2017). The domain #1 from IRIS is enriched with several mitochondrial genes and Sertoli cells marker genes (*e.g., mt-Rnr2, and Clu* (Zhang et al. 2015)), thus likely representing the germinal epithelium structure (**Figure S4.7**). Importantly, quantifications with marker gene enrichment again highlight the accuracy of IRIS, which achieved enrichment pattern 4%-225% stronger compared to the other methods (**Figure 4.3E**).

We carefully examined the molecular and cellular signatures of the spatial domains detected by IRIS. First, we performed domain specific DE analysis. DE analysis identified a

median of 4,077 DE genes across five domains, revealing both known region-specific marker genes (e.g., *Tekt2, Tex36, Tnp1, Prm1, Aurka,* and *Clu*) and novel DE genes (e.g., *Kif2b* (Lin et al. 2019), *Sycp1* (Nabi et al. 2022), *Gsg1* (Malcher et al. 2013), *Lyar* (Chen et al. 2021), and *Ldhc* (Tang, Kung and Goldberg 2008), **Figure S4.7**). For example, *Kif2b*, a kinesin family member gene that is involved in metaphase plate congression and chromosome segregation during meiosis (Lin et al. 2019), is identified as the RS domain (#0) specific gene. *Ldhc*, a testis specific gene that is involved in energy metabolism during the middle and later stages of spermatogenesis (Tang et al. 2008), is identified as the secondary spermatocyte domain (#4) specific gene. GSEA analysis further revealed that the domain specific DE genes are highly enriched in cilium cellular component related pathways, male gamete generation, and cell-cycle associated pathways (**Figure 4.3F, Figure S4.8**). Next, we investigated the inferred cell type compositions in each detected domain and found that each domain is often characterized by domain specific cell types (**Figure 4.3G - Figure 4.3H**, details in **Methods**). For example, the spatial domain #2 is dominated by ES cells; domain #0 is dominated by RS cells; domains #3 & #4 are dominated by spermatocytes; and domain #1 is composed of multiple cell types including sertoli cells, leydig cells, endothelial cells, macrophage cells, etc. The distinct cellular composition of the seminiferous tubules is clearly visualizable by the spatial distribution of different cell types (**Figure 4.3H**).

Importantly, IRIS revealed critical changes in the spatial organization of testicular microenvironment under diabetic conditions. Specifically, the ES region (#2) is highly concentrated in the center of the seminiferous tubes in the WT but displays much diffused pattern in the ob/ob mice with frequent intermingling with the other spatial domains (**Figure 4.3B**, from top to the bottom). Purity analysis (details in **Methods**) revealed reduced ES domain purity and reduced ES cell type purity under diabetic conditions (**Figure 4.3K**), suggesting a loss of

ES/spermatozoon in the ob/ob testes (Ding et al. 2015, Chen et al. 2021). The structural change of the ES region under diabetic conditions is also accompanied by downregulation of ES marker genes such as *Prm1, Prm2, Odf1,* and *Smcp* (**Figure 4.3I**, **Figure S4.9**), with 30% - 59% reduction in expression levels after diabetes (**Figure 4.3J**). In addition, several mitochondrial genes (i.e., *mt-Rnr1, mt-Rnr2*) are upregulated in the spatial domain #1 in the ob/ob mice, consistent with mitochondrial dysfunction in the pathogenesis of diabetes (Antonetti, Reynet and Kahn 1995, Aly 2021, Al-Kafaji, Sabry and Bakhiet 2016).

4.3.4 High resolution mouse olfactory bulb Stereo-seq data

We further applied IRIS to a subcellular resolution spatial transcriptomics data collected from a recent technology Stereo-seq (Chen et al. 2022). This data consists of gene expression measurements on 23,815 or 26,145 genes across 104,931 or 107,416 spatial locations on two adjacent mouse olfactory bulb tissue slices. For the reference, we used the scRNA-seq sequenced through 10x Genomics Chromium technology, consisting of 18 cell subpopulations (Tepe et al. 2018). In this data, we were only able to apply IRIS as all the other methods failed to run due to heavy computational burden.

The mouse olfactory bulb consists of multiple layered structures organized in an inside out fashion that include the subependymal–ependymal layer (SEL) (Nagayama et al. 2014), granule cell layer (GCL), internal plexiform layer (IPL), mitral cell layer (MCL), external plexiform layer (EPL), glomerular layer (GL), olfactory nerve layer (ONL), and meninges (Gudjohnsen et al. 2015) (**Figure 4.4A**). IRIS accurately and clearly depicted the layered structure of the olfactory bulb (**Figure 4.4B**, **Figure S*4.10***) regardless of the specified number of spatial domains (**Figure 4.4C**). The identified layered structure is supported by the enrichment of known marker genes (**Figure 4.4D - Figure 4.4E**). For instance, the meninge marker gene *Ptgds* (DeSisto et al. 2020)

is enriched in the spatial domain #1. The ONL layer marker *S100a5* (Tepe et al. 2018, Tan, Li and Xie 2015) is highly expressed in domain #2 (**Figure 4.4D**). The olfactory ensheathing cell (OEC) marker gene *Mgst1* (Chen et al. 2021) is enriched in a peripheral sublayer of the ONL (#3). The periglomerular cells (PGCs) marker in GL, *Apold1* (Li et al. 2021), is highly concentrated in domain #6. IRIS also differentiates the outer/superficial (#4) and inner/deep (#10) layer of EPL, as is evident by the distinct spatial patterns of *Cbln4* (Nagayama et al. 2014) and Ly6g6e (Zeppilli et al. 2021). The MCL interneuron marker *Vip* (Wang et al. 2022a) is highly enriched in domain #7. The retina IPL marker gene *Slc17a7* is enriched in domain #5 (Johnson et al. 2007), suggesting that domain #5 is the IPL of the olfactory bulb that shares transcriptomic similarity with the IPL in the retina. IRIS also divides GCL into GCL-external (#9) and GCL-inner (#8) with distinct enrichment of the granule cell marker *Tpbg* and a new marker gene *Penk* (Hawrylycz et al. 2014a, Malvaut et al. 2017). Finally, *Sox11* is a marker for immature neuron (Haslinger et al. 2009) and is enriched in the detected SEL (#0). Quantifications confirm these enrichment patterns (**Figure 4.4E**).

We performed additional analysis to further examine the molecular and cellular signatures of the spatial domains detected by IRIS. First, we performed domain specific DE analysis and identified a median of 1,376 genes across 11 domains. The identified DE genes include known marker genes (e.g., *Sox11, Ptgds, S100a5, Ly6g6e, Slc17a7, Apold1, Vip, Tpbg, and Cbln4*) and novel DE genes (e.g., Fabp7 (Young, Heinbockel and Gondré-Lewis 2013), *Clca3a1* (Gwon, Rhee and Sung 2018), *Meis2* (Fujiwara and Cave 2016), Cck (Liu and Liu 2018, Sun et al. 2020b), **Figure S*4.11***). For example, the identified ONL (#2) specific gene *Fabp7* is heavily expressed in the ensheathing glial cells of the olfactory nerve, promoting the establishment and regenerative growth in sensory neurons in ONL (Young et al. 2013). The identified inner/deep EPL (#10)

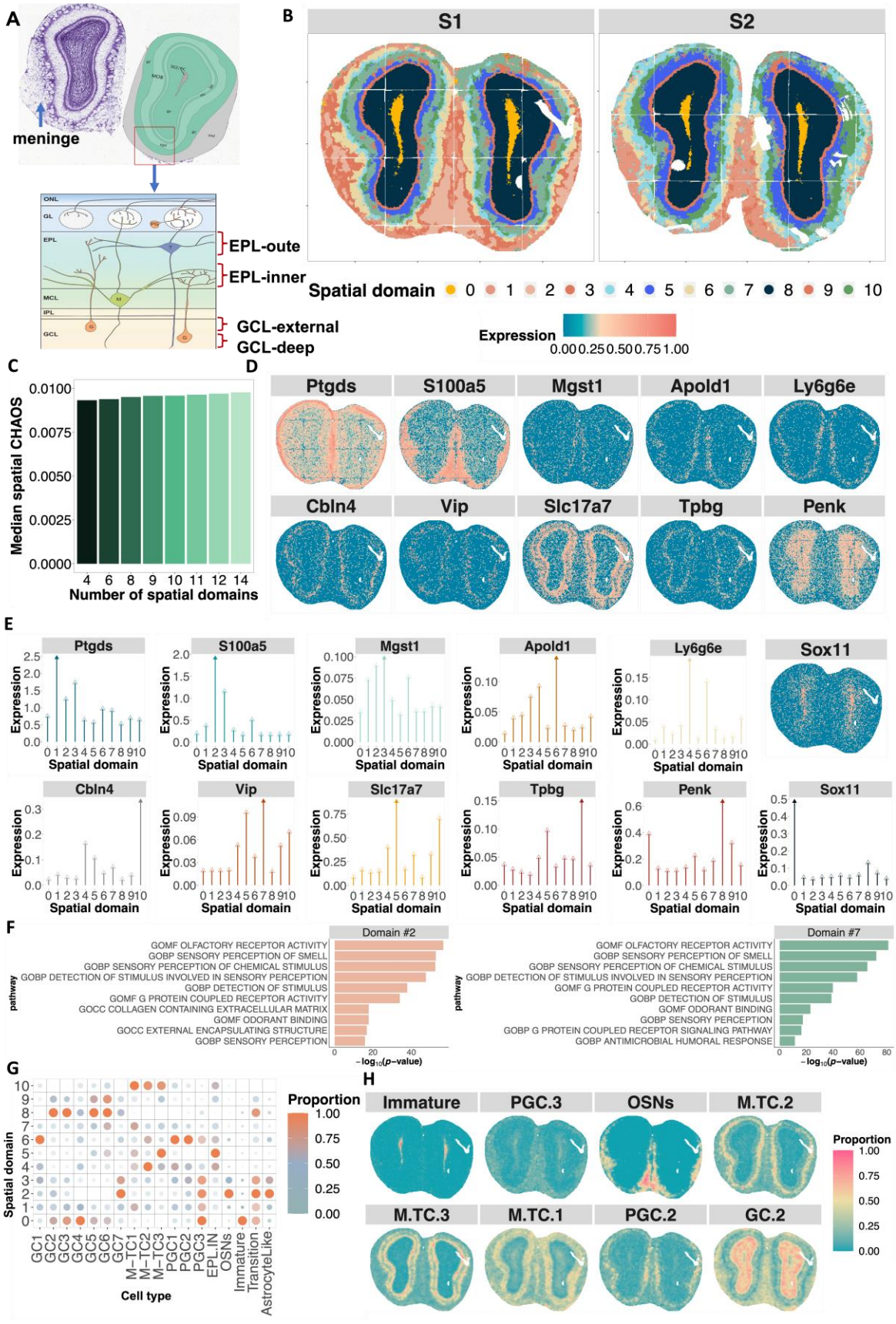

A
meningine
ONL
GL
EPL
MCL
IPL
GCL
EPL-oute
EPL-inner
GCL-external
GCL-deep
B
S1
S2
Spatial domain 0 1 2 3 4 5 6 7 8 9 10
Expression
0.00 0.25 0.50 0.75 1.00
C
Median spatial CHAOS
0.0100
0.0075
0.0050
0.0025
0.0000
4 6 8 9 10 11 12 14
Number of spatial domains
D
Ptgds
S100a5
Mgst1
Apold1
Ly6g6e
Cbln4
Vip
Slc17a7
Tpbg
Penk
E
Ptgds
S100a5
Mgst1
Apold1
Ly6g6e
Sox11
Cbln4
Vip
Slc17a7
Tpbg
Penk
Sox11
Expression
Spatial domain
F
Domain #2
GOMF OLFACTORY RECEPTOR ACTIVITY
GOBP SENSORY PERCEPTION OF SMELL
GOBP SENSORY PERCEPTION OF CHEMICAL STIMULUS
GOBP DETECTION OF STIMULUS INVOLVED IN SENSORY PERCEPTION
GOBP DETECTION OF STIMULUS
GOMF G PROTEIN COUPLED RECEPTOR ACTIVITY
GOCC COLLAGEN CONTAINING EXTRACELLULAR MATRIX
GOMF ODORANT BINDING
GOCC EXTERNAL ENCAPSULATING STRUCTURE
GOBP SENSORY PERCEPTION
pathway
−log10(p−value)
Domain #7
GOMF OLFACTORY RECEPTOR ACTIVITY
GOBP SENSORY PERCEPTION OF SMELL
GOBP SENSORY PERCEPTION OF CHEMICAL STIMULUS
GOBP DETECTION OF STIMULUS INVOLVED IN SENSORY PERCEPTION
GOMF G PROTEIN COUPLED RECEPTOR ACTIVITY
GOBP DETECTION OF STIMULUS
GOMF ODORANT BINDING
GOBP SENSORY PERCEPTION
GOBP G PROTEIN COUPLED RECEPTOR SIGNALING PATHWAY
GOBP ANTIMICROBIAL HUMORAL RESPONSE
pathway
−log10(p−value)
G
Spatial domain
GC1 GC2 GC3 GC4 GC5 GC6 GC7 M−TC1 M−TC2 M−TC3 PGC1 PGC2 PGC3 EPL_IN OSNs Immature Transition AstrocyteLike
Cell type
Proportion
1.00
0.75
0.50
0.25
0.00
H
Immature
PGC.3
OSNs
M.TC.2
M.TC.3
M.TC.1
PGC.2
GC.2
Proportion
1.00
0.75
0.50
0.25
0.00

Figure 4.4 Analyzing the mouse olfactory bulb stereo-seq subcellular data.

(A) The structure of the mouse olfactory bulb (MOB) with the layers annotated based on both the Allen Brain Atlas and previous literature. The mouse olfactory bulb is organized in a layered structure and consists of subependymal–ependymal layer (SEL) (Nagayama et al. 2014), granule cell layer (GCL), internal plexiform layer (IPL), mitral cell layer (MCL), external plexiform layer (EPL), glomerular layer (GL), olfactory nerve layer (ONL), and meninges (Gudjohnsen et al. 2015). The membrane surrounding the mouse olfactory bulb is called meninges**. (B)** Spatial domains detected by IRIS in both slices. **(C)** Barplots display CHAOS values on the tissue slices of MOB when varying the pre-specified number of spatial domains. The median CHAOSs across all slices in the analysis was used. **(D)** Scatter plots display the spatial distribution of important MOB related marker genes. **(E)** Lollipop plots display the mean expression of important marker genes (y-axis) in each spatial domain identified by IRIS (x-axis). The solid arrow indicates the highest enrichment in the corresponding domain. **(F)** Top enriched gene sets in selected spatial domains detected by IRIS (e.g., #2, #7) from the gene set enrichment analysis. **(G)** Heatmap plot displays the estimated mean cell type proportion for representative cell types in each spatial domain detected by IRIS. The color scale was normalized to 0-1 range. **(H)** Spatial scatter plot displays the spatial distribution of IRIS estimated cell type proportion for representative cell types across spatial locations. For D – H, the results are shown for example S1 slice in the main analysis.

specific DE gene *Cck* is expressed preferentially in tufted cells mainly located in deep EPL (Liu and Liu 2018, Sun et al. 2020b). GSEA further revealed that the domain specific DE genes detected by IRIS are highly enriched in the olfactory receptor activity related pathway, sensory perception of smell pathway, and odorant binding pathway (**Figure 4.4F, Figure S4.12**), all related to the functional activation of the olfactory bulb. Next, we examined the cell type compositions inferred by IRIS and found that the spatial domains identified by IRIS consist of unique cell type characteristics (**Figure 4.4FG - Figure 4.4H**, **Figure S4.13**). For example, mitral cell subtypes are enriched in three adjacent layers EPL (#4 & #10), IPL (#5) and MCL (#7). The majority of the granule cell subpopulations (GC2, GC3, GC5, GC6) reside mainly in the GCL-inner (#8) while GC5 and GC6 are also located in the GCL-external (#9), implying the distinct functions of the two GCL sublayers (**Figure 4.4G**). Immature cells are only located in domain #0, supporting domain #0 being the SEL.

4.3.5 High resolution human breast cancer 10x Xenium data

Finally, we applied IRIS to a high resolution spatial transcriptomics dataset generated by the most recent technology 10x Xenium (Janesick et al. 2022). This data consists of gene expression measurements on 313 genes across either 118,708 or 167,782 spatial locations on two adjacent human breast cancer tissue slices. For the reference, we used the scRNA-seq sequenced through 10x Genomics Chromium technology, consisting of 29 cell subpopulations (Wu et al. 2021). In this data, again we were only able to apply IRIS as all the other methods failed to run due to heavy computational burden.

Breast cancer is a heterogeneous disease with high intratumoral and intertumoral variation in histological and molecular features. In the 10x Xenium data, IRIS clearly identified four distinct tumor domains, including two ductal carcinoma *in situ* (DCIS) domains that represent non-invasive forms of breast cancer (#10 & #16) and two invasive ductal carcinoma (IDC) domains (domain #5 & #8) (**Figure 4.5A, Figure S4.14**), along with multiple other domains that belong to part of tumor microenvironment (TME), including the immune-related regions (#0, #4, & #13), tumor stroma region (#1), and myoepithelial layer (#14). The spatial domains detected by IRIS are spatially smooth and compact, regardless of the pre-specified number of spatial domains (**Figure 4.5B**). The spatial domains identified by IRIS are supported by the enrichment of known marker genes (**Figure 4.5C - Figure 4.5D**). For instance, *TCIM* (*TC-1*), a candidate breast cancer oncogene associated with β-catenin pathway that elevates the invasive behaviour of cancer cells (Su et al. 2013, Yang et al. 2007), is highly enriched in the IDC domains #5 & #8. *FASN (Janesick et al. 2022),* an invasive tumor marker gene, is highly enriched in the IDC domain #8. *CEACAM6,* which acts as a cell adhesion molecular with expression negatively correlated with cell differentiation (Han et al. 2008, Janesick et al. 2022), is highly enriched in the DCIS domains #10 & #16. Besides the tumor domains, the multiple immune related domains detected by IRIS are also

distinguished from each other by their unique cellular compositions characterized by different marker genes (**Figure 4.5D**). For example, *FGL2*, a marker gene mainly expressed in macrophages and dendric cells, is highly enriched in the immune-related domain #0 but not tumor related domains (#5, #8, #10, & #16) (Feng et al. 2020, Yu et al. 2021a). In contrast, the tumor-infiltrating B cells marker gene *BANK1* (Janesick et al. 2022) is enriched in the immune-related domain #4, while the lymphocyte marker gene *IL7R* (Janesick et al. 2022) is enriched in the immune-related domain #13. In addition, the stromal cell marker gene *POSTN* is enriched in domain #1 and the myoepithelial cell marker gene *KRT6B* is enriched in domain #14, supporting the complexity of cellular composition of the TME. Quantifications confirm these enrichment patterns.

We performed additional analysis to examine the molecular and cellular signatures of the spatial domains detected by IRIS. First, we performed domain specific DE analysis and identified a median of 141 genes across 17 domains. The identified DE genes include known marker genes (e.g., *TCIM, FASN, CEACAM6, FGL2, BANK1, IL7R, POSTN, and KRT6B*) and novel DE genes (e.g., *NDUFA4L2 (Yuan et al. 2021), ABCC11* (Toyoda and Ishikawa 2010), *EPCAM* (Soysal et al. 2013), *CXCR4* (Mukherjee and Zhao 2013), **Figure S4.15**). For example, the identified IDC (#5) specific gene *NDUFA4L2* can promote trastuzumab resistance in HER2+ breast cancer, supporting it as a potential therapeutic target (Yuan et al. 2021). GSEA further revealed that the tumor related domain specific DE genes (#5, #8, #10, & #16) detected by IRIS are highly enriched in multiple tumor cell lines such as luminal-like breast cancer cell lines and HER2/ERBB2 breast cancer cell line; the immune-related domain specific DE genes (#0, #4, & #13) are highly enriched in immune cells pathways, lymphocyte activation pathway, monocyte macrophage cells, NK cell, T cells and immune response related pathways (**Figure 4.5E**, **Figure S4.16**); and the stromal domain specific DE genes (#1) are highly enriched in fibroblast cells, stem cells, stellate cells, and

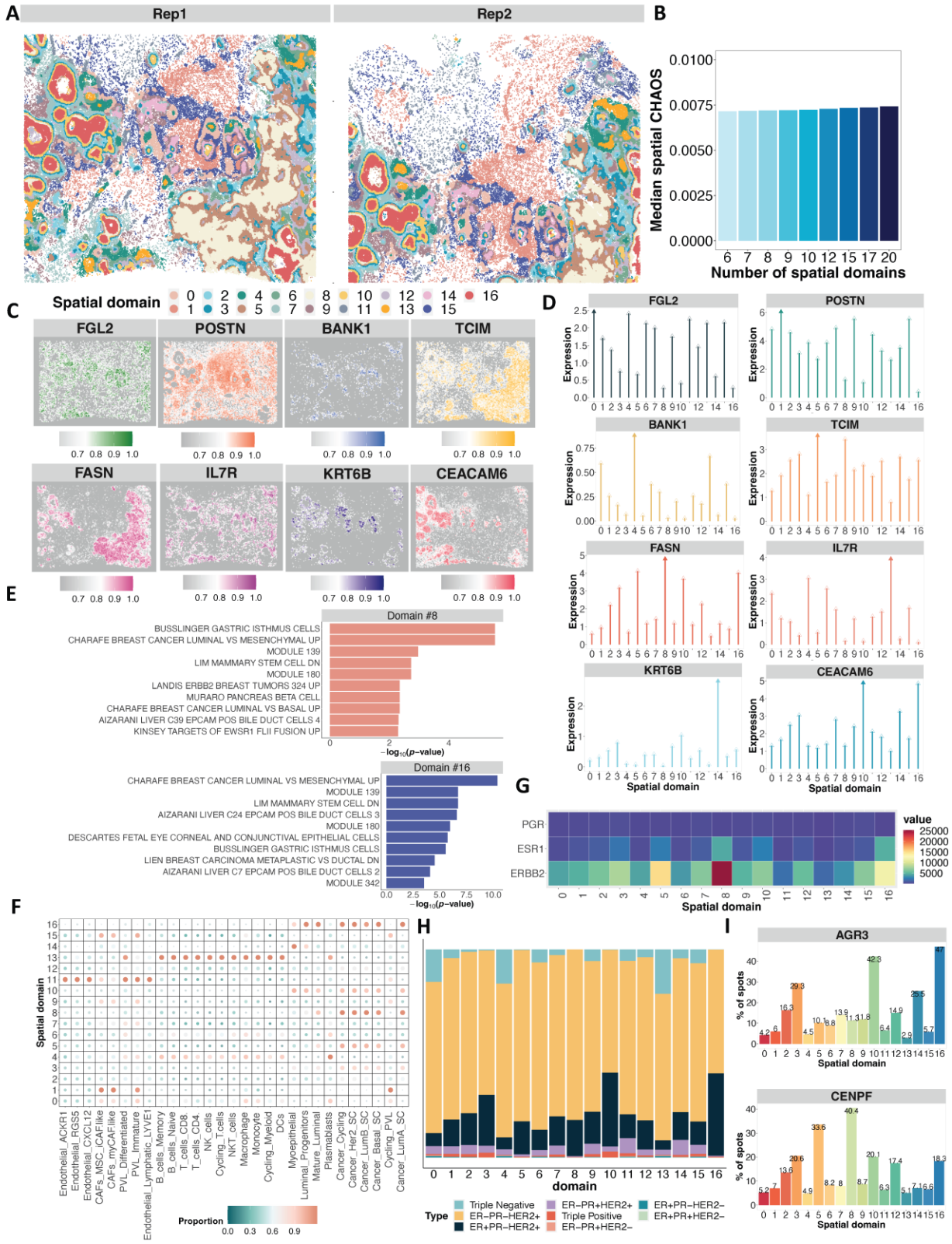

A
Rep1
Rep2
B
Median spatial CHAOS
Number of spatial domains
C
Spatial domain
FGL2
POSTN
BANK1
TCIM
FASN
IL7R
KRT6B
CEACAM6
D
Expression
Spatial domain
E
Domain #8
BUSSLINGER GASTRIC ISTHMUS CELLS
CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL UP
MODULE 139
LIM MAMMARY STEM CELL DN
MODULE 180
LANDIS ERBB2 BREAST TUMORS 324 UP
MURARO PANCREAS BETA CELL
CHARAFE BREAST CANCER LUMINAL VS BASAL UP
AIZARANI LIVER C39 EPCAM POS BILE DUCT CELLS 4
KINSEY TARGETS OF EWSR1 FLII FUSION UP
−log10(p−value)
Domain #16
CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL UP
MODULE 139
LIM MAMMARY STEM CELL DN
AIZARANI LIVER C24 EPCAM POS BILE DUCT CELLS 3
MODULE 180
DESCARTES FETAL EYE CORNEAL AND CONJUNCTIVAL EPITHELIAL CELLS
BUSSLINGER GASTRIC ISTHMUS CELLS
LIEN BREAST CARCINOMA METAPLASTIC VS DUCTAL DN
AIZARANI LIVER C7 EPCAM POS BILE DUCT CELLS 2
MODULE 342
−log10(p−value)
G
PGR
ESR1
ERBB2
value
Spatial domain
F
Spatial domain
Endothelial_ACKR1
Endothelial_RGS5
Endothelial_CXCL12
CAFs_MSC_iCAFlike
CAFs_myCAFlike
PVL_Differentiated
PVL_Immature
Endothelial_Lymphatic_LYVE1
B_cells_Memory
B_cells_Naive
T_cells_CD8
T_cells_CD4
NK_cells
Cycling_T.cells
NKT_cells
Macrophage
Monocyte
Cycling_Myeloid
DCs
Myoepithelial
Luminal_Progenitors
Mature_Luminal
Plasmablasts
Cancer_Cycling
Cancer_Her2_SC
Cancer_LumB_SC
Cancer_Basal_SC
Cycling_PVL
Cancer_LumA_SC
Proportion
H
domain
Type
Triple Negative
ER−PR−HER2+
ER+PR−HER2+
ER−PR+HER2+
Triple Positive
ER−PR+HER2−
ER+PR−HER2−
ER+PR+HER2−
I
AGR3
CENPF
% of spots
Spatial domain

Figure 4.5 Analyzing the human breast cancer 10x Xenium data.

(A) Spatial domains detected by IRIS in both slices. **(B)** Barplots display the CHAOS values on both tissue slices when varying the pre-specified number of spatial domains. The median CHAOSs across all slices in the main analysis was used. **(C)** Scatter plots display the spatial distribution of important breast cancer related marker genes. **(D)** Lollipop plots display the mean expression of important marker genes (y-axis) in each spatial domain identified by IRIS (x-axis). Solid arrow indicates the highest enrichment value. **(E)** Top enriched gene sets in the selected spatial domains identified by IRIS (e.g., IDC domain #8, DCIS #16) in the gene set enrichment analysis. **(F)** Heatmap plot displays the estimated mean cell type proportion for representative cell types in each spatial domain detected by IRIS. Color scale was normalized to 0-1 range. **(G)** Heatmap plot displays the number of spatial locations that express the three hormone receptor marker genes (*ERBB2*, *ESR2*, and *PGR*) in each spatial domain detected by IRIS. **(H)** Bar plot displays the proportion of breast cancer subtypes based on the hormone receptor status in each spatial domain detected by IRIS. **(I)** Bar plot displays the percentage of spatial locations that have high expression (greater than median expression across all spatial locations) of two important tumor invasiveness marker gene *AGR3* and *CENPF* in each spatial domain. For C – I, the results are shown for the example Rep1 slice in the main analysis.

stromal cells related pathways. Next, we examined the cell type compositions inferred by IRIS and found that the spatial domains identified by IRIS consist of unique cell type characteristics (**Figure 4.5F**, **Figure S4.17**). For example, cancer cell subpopulations (cycling, Her2, luminal like, basal) are highly enriched in tumor domains #5, #8, #10, and #16, with the latter two DCIS regions also enriched with myoepithelial, luminal progenitors, and mature luminal cells. The majority of immune cells (e.g., B cells, T cells, NK cells, macrophages, monocytes, cycling myeloid cells, dendric cells) are highly enriched in immune-related domains #13 and #4, though the enrichment is lower in the latter domain. And the cancer-associated fibroblast (CAF) subpopulations (Von Ahrens et al. 2017) are highly enriched in the stroma domain #1.

Finally, we carefully examined the hormonal receptor status of the tumor regions to characterize the invasiveness and intratumor heterogeneity of the tissue (details in **Methods**). We found that the hormonal receptor *ERBB2* is highly expressed in the IDC domains #8 and #5, more so than that in the DCIS domains #10 and #16, supporting the advanced metastasis-related properties (Yu and Hung 2000) of the IDC domains (**Figure 4.5G**). We also found that the

hormonal receptor *ESR1* is mainly expressed in the IDC and DCIS regions while the hormonal receptor *PGR* is lowly expressed in all domains. Classification based on the expression of these three hormonal receptors thus confirmed that most regions are either *ERBB2+* (HER2+) or double positive *ERBB2+/ESR1+* (HER2+/ER+) (**Figure 4.5H**, **Figure S4.18**). The proportion of *ERBB2+/ESR1+* breast cancer in DCIS regions (#10 & #16) is much higher than that in the IDC regions (#5 & #8). In addition, the proportion of triple negative breast cancer (TNBC) is higher in immune-related domains #0, #4, and #13, consistent with the fact that TNBC is associated with a high density of tumor-immune cells infiltration (Lv et al. 2021). Quantifications of the proportion of spatial locations in each domain that display a high expression of the tumor invasiveness marker genes (*AGR3* and *CENPF*) further confirmed that domain #8 is more invasive than domains #5, #10 and #16 (**Figure 4.5I**, details in **Methods**). Specifically, *AGR3* is associated with low histological grade breast tumors (Jian et al. 2020) and the proportion of spatial locations marked with high *AGR3* expression is higher in the DCIS domains #16 and #10. In contrast, *CENPF* is associated with tumor aggressive features and poor prognosis of patients with breast cancer (O'Brien et al. 2007), and the proportion of spatial locations marked with high *CENPF* expression is higher in IDC domains #8 and #5.

4.4 Discussion

We have presented a new computational method, IRIS, for accurate and scalable spatial domain detection in SRT studies via integrated reference-informed segmentation. Different from existing methods (Hu et al. 2021, Zhao et al. 2021, Fu et al. 2021, Dries et al. 2021, Moses and Pachter 2022, Palla et al. 2022, Zhu et al. 2018, Tian et al. 2022, Rao et al. 2021, Li and Zhou 2022, Shang and Zhou 2022), IRIS leverages the cell type specific gene expression information from scRNA-seq data to detect the spatial domains on multiple tissue sections in the SRT study,

while simultaneously accounting for both within and between tissue slice cell type compositional similarities. In addition, IRIS takes advantage of multiple algorithmic innovations to achieve scalable computation. In the moderate-sized datasets (i.e., 10x Visium human DLPFC, Slide-seq mouse spermatogenesis) with 3,460 ~ 88,884 spatial locations, IRIS is 8.5 - 134.7 times faster than the existing methods while using only 1.5% - 58.7% of the physical memory required by these methods. IRIS is also the only method that is scalable to the very recent large mouse MOB stereo-seq and human breast cancer 10x Xenium datasets with 104,931-167,782 spatial locations and can finish analysis there in 26 ~ 28 minutes with 7.6 - 26.1 GB physical memory (**Figure S4.19**). We have demonstrated the benefits of IRIS through in-depth analysis of four SRT datasets generated from different technologies across distinct tissues and species.

While IRIS represents the first attempt to integrate a reference scRNA-seq with the SRT study to detect spatial domains, scRNA-seq has been previously used in two other analytic settings in SRT studies (Li et al. 2022, Sun et al. 2022, Li et al. 2023, Mages et al. 2023). In particular, integrating scRNA-seq data with SRT has been shown to improve the prediction of the spatial distribution of transcripts (Shengquan et al. 2021, Nitzan et al. 2019, Cang and Nie 2020, Abdelaal et al. 2020, Lopez et al. 2019, Biancalani et al. 2021) and improve the estimation of cell type proportions across spatial locations on a single tissue slice (Ma and Zhou 2022, Cable et al. 2022, Kleshchevnikov et al. 2022, Andersson et al. 2020, Dong and Yuan 2021, Lopez et al. 2021). The results in the present study thus dovetail these recent findings and highlight the benefits of integrating reference scRNA-seq data to improve the analytics of SRT studies. Besides integrating scRNA-seq data, IRIS also provides a flexible framework for integrating information from neighboring spatial locations in each tissue slice as well as that from multiple tissue slices to enhance spatial domain detection in spatial transcriptomics. The flexible modeling framework of

IRIS in principle can be extended to performing other analytic tasks in SRT studies. For example, we can extend the current optimization framework of IRIS into a matrix tri-factorization to directly map the single cells from the scRNA-seq study onto each measured spatial location in the SRT data. Exploring the benefits of integrating scRNA-seq data with SRT and the integrative modeling of multiple SRT datasets in the future will likely yield fruitful results for many other SRT analytic tasks.

We have performed a series of post-domain detection analyses to further validate and quantify the identified spatial domains by IRIS. One analysis task we performed was domain-specific differential expression (DE) analysis. However, when the domains are not known as a priori but inferred from the same expression data, the clustering analysis will introduce a "selection bias" that would result in false discoveries. The DE results after the domain detection analysis might contain false signals with an enrichment of small DE p-values under the null. Indeed, our focus is to find biological evidence from domain-specific genes to support the domains detected by IRIS when there is no ground truth. Several methods have been proposed to address and correct the "selection bias" in DE analysis for scRNA-seq studies. For instance, Gao and Witten (Gao, Bien and Witten 2022) introduced a selective inference framework that tests for mean differences in clusters. Neufeld et al. (Neufeld et al. 2022) developed a count splitting framework to control the type I error in scRNA-seq DE p-values. Other methods (Vandenbon and Diez 2020, Missarova et al. 2023), such as singleCellHaystack (Vandenbon and Diez 2020), rely on a clustering-free framework to detect DE genes in an unbiased manner. Although these methods have been proposed and applied in the context of scRNA-seq studies, specific methods targeting the selection bias in SRT studies have not been extensively explored. It is crucial for future efforts to adapt and evaluate

the performance of these existing methods or develop novel approaches to correct the selection bias in domain-specific DE analysis within the spatial transcriptomics context.

There are several important future extensions for IRIS. Firstly, IRIS accommodates the cell type compositional similarity across neighboring locations on each tissue slice through a graph Laplacian regularization, which is constructed based on an adjacency matrix. An important benefit of the adjacency matrix is that it can be easily formulated in a sparse form to facilitate computation. However, we note that our method and software can easily incorporate other types of similarity or kernel matrices such as the Gaussian kernels and the periodic kernels to capture the diverse and complex spatial correlation patterns that may be encountered across datasets. IRIS can also incorporate a weighted combination of multiple kernel matrices to further improve performance. Secondly, while we have primarily focused on using the K-means penalty function as part of the IRIS model for spatial domain detection, IRIS can be coupled with different clustering penalty functions such as those used in the Louvain or Leiden clustering algorithms to take advantage of their distinct benefits. Thirdly, while we have mainly focused on using the transcriptomics measurements from both scRNA-seq and SRT studies, IRIS's modeling framework is general and can in principle be extended to integrate other data modalities such as histological images that are often collected alongside SRT. In particular, we can introduce an additional penalty term to encourage the similarity between domains in terms of image intensity levels and/or cell morphological features extracted from the images. We note, however, that such extension of IRIS requires both the segmentation of the image to identify cells on the tissue and aligning the cells segmented on the image with the SRT data, both of which remain technically challenging and likely require future methodological development.

4.5 Methods

4.5.1 IRIS method overview

We present an overview of IRIS here. IRIS is a reference-informed integrative computational method for spatial domain detection in spatial transcriptomics. A unique feature of IRIS is its ability to integrate cell type specific expression profiles from a reference scRNA-seq data to facilitate the mapping of spatial domains on the same tissue in spatial transcriptomics. In the process, IRIS integrates the spatial transcriptomic profiles from neighboring spatial locations on each single tissue slice as well as that across multiple tissue slices, facilitating accurate and consistent spatial domain detection across slices. Importantly, IRIS performs all these integrative analyses seamlessly in a joint modeling framework and incorporates an efficient iterative optimization algorithm for scalable computation. As a result, IRIS can accurately and rapidly detect spatial domains on complex tissues in large-scale spatial transcriptomics.

IRIS requires two types of data input: a scRNA-seq reference dataset and a spatial transcriptomics dataset that measures the transcriptomic profiles of one or multiple tissue slices. The scRNA-seq reference data consists of K cell types with a set of G cell-type-informative genes, which can be selected based on (Ma and Zhou 2022). We follow (Wang et al. 2019, Ma and Zhou 2022) to extract from the scRNA-seq data a reference basis matrix $\boldsymbol{B}$, which is a G by K matrix that contains the mean expression level of the G cell type informative genes in the K annotated cell types. The spatial transcriptomics data, on the other hand, consists of T different tissue slices ($T \geq 1$). We denote $\boldsymbol{Y}_t$ as the G by N_t gene expression matrix in slice t for the same set of G informative genes measured at N_t spatial locations. We assume that there are total of R distinct spatial domains across tissue slices in spatial transcriptomics. We introduce an $\boldsymbol{N}_t$-vector of spatial domain

indicators $\boldsymbol{c}_t$ to indicate the domain label for each spatial location on tissue slice *t*, where each element of $\boldsymbol{c}_t$ takes values in $\{1, \dots, R\}$.

Our goal is to detect the spatial domains on the tissue slices in spatial transcriptomics by inferring the spatial domain indicators $\boldsymbol{c}_t$. Inferring $\boldsymbol{c}_t$ requires knowing the cell type composition of each spatial location on the tissue slice, since every spatial domain is characterized by a unique composition of cell types. Therefore, to facilitate the inference of $\boldsymbol{c}_t$, we introduce an N_t by *K* cell type composition matrix $\boldsymbol{P}_t$, where each row of $\boldsymbol{P}_t$ represents either the proportions of the *K* cell types (for spot-level spatial transcriptomics) on or the contribution of the *K* cell types (for single-cell or subcellular resolution spatial transcriptomics) to each spatial location on slice *t*. The composition matrix $\boldsymbol{P}_t$ is not only key for inferring $\boldsymbol{c}_t$ but also serves as an important link between the reference basis matrix $\boldsymbol{B}$ in scRNA-seq and the expression matrix $\boldsymbol{Y}_t$ in spatial transcriptomics, thus connecting the two distinct data types. In particular, each element of $\boldsymbol{Y}_t$, which describes the expression level of an informative gene at a spatial location, can be expressed as the product of the gene's expression level in each cell type and the cell type compositions on the location. Consequently, we can infer $\boldsymbol{P}_t$ by minimizing the difference between $\boldsymbol{Y}_t$ and $\boldsymbol{B}\boldsymbol{P}_t^{\boldsymbol{T}}$ in terms of Euclidean distance, also known as the square of the Frobenius norm, through the following cost function:

$$\min_{\substack{0\le \boldsymbol{V}_t\le 1 \\ t=1,2,\dots T}} \sum_{t=1}^{T} \left|\left|\boldsymbol{Y}_t - \boldsymbol{B}\boldsymbol{P}_t^{T}\right|\right|_F^2. \tag{4.1}$$

In the process, we constrain each element of $\boldsymbol{P}_t$ to be non-negative and we accommodate the spatial correlation in the cell type composition among neighboring locations, which are commonly observed on each tissue slice, through the following penalty function:

$$\sum_{t=1}^{T} Tr(\boldsymbol{P}_t^{T}\boldsymbol{L}_t\boldsymbol{P}_t), \tag{4.2}$$

where $Tr(\cdot)$ denotes the trace of a matrix and $\boldsymbol{L}_t$ is the graph Laplacian matrix for slice *t*, expressed as the difference between two matrices $\boldsymbol{L}_t = \boldsymbol{D}_t - \boldsymbol{A}_t$. Here, $\boldsymbol{A}_t$ is a N_t by N_t adjacency matrix its *ij*-th element is one if *i*-th and *j*-th locations on slice *t* are mutual neighbors, defined as being the *k* nearest neighbors of each other (default *k*=10); and is zero otherwise. $\boldsymbol{D}_t$ is a diagonal matrix whose entries are column sums of $\boldsymbol{A}_t$. Minimizing each term $Tr(\boldsymbol{P}_t^T \boldsymbol{L}_t \boldsymbol{P}_t)$ is equivalent to minimizing the summation of the weighted square of the Euclidean norm $\frac{1}{2}\sum_{i,j=1}^{N_t} \left|\left| \boldsymbol{P}_{t_i}^T - \boldsymbol{P}_{t_j}^T \right|\right|_2^2 \boldsymbol{A}_{t_{ij}}$, where $\boldsymbol{P}_{t_i}$ is the *i*-th row of $\boldsymbol{P}_t$ and $\boldsymbol{A}_{t_{ij}}$ is the *ij*-th element of $\boldsymbol{A}_t$. Therefore, minimizing the penalty function in equation (4.2) encourages similarity in the cell type composition in the neighboring locations on each slice, facilitating accurate and robust inference of $\boldsymbol{P}_t$. Through modeling spatial correlations, the penalty function of equation (4.2) also allows us to effectively integrate the spatial transcriptomic profiles across neighboring spatial locations on each tissue slice.

Besides connecting the spatial transcriptomics data with the reference scRNA-seq data, the cell type composition matrix $\boldsymbol{P}_t$ also provides direct evidence for inferring the spatial domain indicators $\boldsymbol{c}_t$. Specifically, we consider the following k-means cost function that connects $\boldsymbol{P}_t$ to $\boldsymbol{c}_t$:

$$\sum_{t=1}^{T}\sum_{r=1}^{R} \left|\left| \frac{1}{n_{t_r}} \boldsymbol{P}_t^T \boldsymbol{q}_{t_r} - \boldsymbol{\mu}_r \right|\right|_2^2, \tag{4.3}$$

where $\boldsymbol{q}_{t_r}$ is an N_t-vector of indicator variables, $\boldsymbol{q}_{t_r}(i) = \begin{cases} 1, & if\ \boldsymbol{c}_{t_i} = r \\ 0, & otherwise \end{cases}$, with each *i*'th element being one when the corresponding location i belongs to the *r*-th spatial domain on tissue slice *t* and

being zero otherwise; n_{t_r} is the total number of locations in the spatial domain *r* on tissue slice *t* (i.e., $n_{t_r} = \sum_i \boldsymbol{q}_{t_r}(i)$); and $\boldsymbol{\mu}_r$ is a *K*-vector of domain-specific cell type composition profile for the *r*-th spatial domain. Equation (4.3) directly relates the average cell type compositions for locations residing on the same spatial domain across tissue slices to the common cell type composition profile parameter $\boldsymbol{\mu}_r$, thus encouraging similarity in the domain-specific cell type compositions across tissue slices. Encouraging such similarity allows us to borrow information across multiple slices for accurate and robust spatial domain inference.

The above equations (4.1) - (4.3) characterize different aspects of the IRIS modeling framework. In particular, equation (4.1) enables the integration of the transcriptomic profiles between scRNA-seq and spatial transcriptomics data. Equation (4.2) encourages within-slice cell type compositional similarities among neighboring locations and allows for integrative transcriptomics analysis across locations on each tissue slice. Equation (4.3) models the consistency of cell type composition on the same spatial domain across slices, allowing for integrative transcriptomics analysis across multiple tissue slices. Importantly, we incorporate all three components together into a joint cost function:

$$\phi(\boldsymbol{P}_t) = \sum_{t=1}^{T} \left|\left|\boldsymbol{Y}_t - \boldsymbol{B}\boldsymbol{P}_t^T\right|\right|_2^2 + \beta \sum_{t=1}^{T} \sum_{r=1}^{R} \left|\left|\frac{1}{n_{t_r}} \boldsymbol{P}_t^T \boldsymbol{q}_{t_r} - \boldsymbol{\mu}_r\right|\right|_2 + \lambda \sum_{t=1}^{T} Tr(\boldsymbol{P}_t^T \boldsymbol{L}_t \boldsymbol{P}_t) \qquad (4.4)$$

where β and λ determine the relative contribution of the three components. The scale of the two parameters β and λ is determined by the dimensionality of the matrices in the three components, which are in turn determined by N_t, *G* and *K*. We set β and λ to fixed values of 1000 and 2000 throughout the present study, with robustness analysis provided in **APPENDIX C.2**.

In the joint model defined in equation (4.4), our primary parameters of interest are the spatial domain labels for the locations on every tissue slice, represented by $\boldsymbol{c}_t$, for $t \in \{1, \ldots T\}$. For model inference, we develop a constrained iterative optimization algorithm, which iteratively updates the spatial domain labels $\boldsymbol{c}_t$, along with the secondary parameters that include the cell type composition matrix $\boldsymbol{P}_t$ and the domain-specific cell type composition profile $\boldsymbol{\mu}_r$. Importantly, the developed inference algorithm makes use of three critical innovations to ensure scalable computation. These innovations include recently developed fast multiplicative updating rules (Lee and Seung 2000, Janecek and Tan 2011), a sparse formulation of the graph Laplacian matrix, and an efficient K-means clustering algorithm. As a result, our algorithm is computationally scalable to large-scale spatial transcriptomics with millions of measured locations and tens of thousands of genes. We refer to our method as IRIS, which is implemented in an R package, with core algorithms written in efficient C++ code that is linked back to the package through Rcpp. The IRIS software is freely available at www.xzlab.org/software.html.

4.5.2 Compared methods for spatial domain detection

In the benchmarking human LIBD dataset, we compared the performance of IRIS with 8 methods belonging to three categories: (1) single-slice based methods: spaGCN (version 1.2.5), BayesSpace (version 1.5.1), and SEDR (downloaded on 04/01/2022); (2) multi-slices based methods: BayesSpaceJoint, which is a variation of BayesSpace (version 1.5.1), BASS (version 1.1.0.16), Maple (version 0.99.1); (3) deconvolution-based methods: CARD (version 1.0), and RCTD (version 1.1.0). Specifically, for the single slice based, and multi-slice based methods, we followed the tutorial on the corresponding GitHub pages and used the recommended default parameter settings for spatial domain detection. For spaGCN, because it can also incorporate histological information whenever available for domain detection, we applied it either with or

without the histological image information for datasets where histological images are available and matched (e.g., the DLPFC data). For SEDR, we followed their tutorial and applied the Leiden clustering algorithm on the low-dimensional latent representation learned by SEDR for domain detection. For the deconvolution-based methods, we followed their tutorial to first perform the deconvolution, then we concatenate the cell type compositional matrix from different slices in the analysis as the concatenated matrix, finally we perform the K-means algorithm on the concatenated composition matrix to detect the spatial domains. We used the term "CARD_kmeans" and "RCTD_kmeans" to represent the deconvolution-based methods correspondingly. Due to the sub-optimal performance of Maple and deconvolution-based methods, we only compare the other methods in the other five datasets.

4.5.3 Real data analysis

We applied IRIS to analyze four published spatial transcriptomics datasets collected by different techniques, with distinct spatial resolutions, and from multiple species and tissues. For each spatial transcriptomics data, we obtained an external scRNA-seq collected on the same type of tissue but with a different sequencing technology to serve as the reference. Details of preprocessing the data are in **APPENDIX C.3**.

Spatial domain detection

We compared the performance of IRIS with the other spatial domain detection methods on four real datasets. In each analysis, we supplied the same spatial transcriptomics data as input for all methods.

In the DLPFC dataset by 10x Visium, we examined four settings in total: (1) baseline setting when the slices are from the same donor that share high similarity; (2) challenging setting

when we use a scRNA-seq reference with one missing cell type information at a time; (3) challenging setting when we use a scRNA-seq reference with mis-classified cell type information by randomly merging two cell types; (4) challenging setting when the slices are from different donors that share low similarity in structures (**APPENDIX C.4**). In each setting, we used the annotated tissue domains as ground truth and evaluated method performance by calculating the adjusted Rand index (ARI; details see **APPENDIX C.1**). In addition to using the ground truth, we also evaluated method performance using marker gene enrichment (details in the next section). For the marker gene enrichment quantification and other downstream analysis, we focus on slice 151509 as the main example as the analysis results for other tissue slices are consistent.

In the high-resolution mouse spermatogenesis data by Slide-seq, we set the number of spatial domains to be 5 and varied it from 4 to 10 to evaluate methods performance. While no histology image nor manual domain annotations were available for this data, the mice testes have well defined tissue architecture in the form of numerous seminiferous tubule structures. Therefore, we evaluated method performance by carefully examining the overall and fine-grained morphology of the tubule structures on the tissue. In addition to carefully examining the inferred tubule structures, we also evaluated method performance using the CHAOS as well as marker gene enrichment. For the marker gene enrichment quantification and other downstream analysis, we focus on the slice WT3_Puck7 as the main example as the analysis results for other tissue slices are consistent.

For the sub-cellular mouse olfactory bulb data by Stereo-seq, we set the number of spatial domains to be 11 based on the domain knowledge of the MOB structure. We varied the number of spatial domains from 4 to 14 and we evaluated the performance of IRIS by calculating CHAOS as

well as marker gene enrichment. For the marker gene enrichment quantification and other downstream analysis, we focus on the slice S1 as the main example as the analysis results for the other tissue slice are consistent. Note that none of the other methods were scalable to this large data.

For the high resolution human breast cancer data by 10x Xenium, we followed the original publication to set the number of spatial domains to be 17 (Janesick et al. 2022). We varied the number of spatial domains from 6 to 20 and we evaluated the performance of IRIS by calculating CHAOS as well as marker gene enrichment. For the marker gene enrichment quantification and other downstream analysis, we focus on the slice Rep1 as the main example as the analysis results for the other slice are consistent. Note that none of the other methods were scalable to this large data.

Domain-specific marker gene quantification across different methods

We quantified the performance of different domain detection methods by comparing the fold change of marker genes in their corresponding spatial domains in the first three datasets. (The two large datasets were excluded as the other methods cannot deal with them). We reasoned that a good spatial domain detection method would yield accurate tissue structures that capture the enrichment of domain-specific marker genes. Because the spatial domain labels inferred from different methods are arbitrary (i.e. domain #1 from one method does not necessarily corresponds to domain #1 from another methods), to ensure fair comparison, we first shuffled the domain labels of each method and mapped them onto a common label system, so that the same domain label from different methods corresponds to the same anatomic region (i.e. domain #1 from all methods now correspond to the same anatomic region). Specifically, for the DLPFC dataset where a ground

truth domain label is available, we manually mapped the inferred domain labels from different methods to the ground truth domains based on the proportion of the spatial locations in each domain that belong to a specific domain in the ground truth. For the other datasets where a ground truth domain label is unavailable, we mapped the domain labels from each of the other methods to the domain labels from IRIS. Specifically, for each method in turn, we first calculated a contingency table for the domain labels from the method and those from IRIS, where the *ij*-th element of the contingency table represents the number of spatial locations in the *i*-th domain of the method that are also inside the *j*-th domain of IRIS. Afterwards, we randomly shuffled the columns of the contingency table, obtained the shuffled contingency table that achieves the maximum trace, and kept the column names in this shuffled contingency table as the new domain labels for the given method. This way, we map the domain labels from each method onto the common IRIS domain label system. Note that the domain label mapping step does not change the spatial clustering results from different methods but simply shuffles and aligns the domain labels to facilitate comparison. Afterwards, with a corresponding set of spatial domain labels, we calculate the fold change of each marker gene in their expected spatial domain to evaluate the performance of different methods. For the human DLPFC dataset, we excluded the deconvolution-based methods (i.e., CARD_kmeans, and RCTD_kmeans) and Maple, as these three methods cannot capture the layered structure of DLPFC tissue. For the Slide-seq mouse testis dataset, we only compared the BayesSpaceJoint, SEDR, and IRIS for marker gene enrichment, as spaGCN and BayesSpace could not capture any tubular structures in the testis at all.

Differential expression (DE) and gene set enrichment analysis (GSEA)

For all the datasets, we performed DE analysis and GSEA analysis in a domain-specific fashion. Specifically, for each spatial domain in turn, we first conducted a Wilcoxon rank sum test by using *Seurat* (Argelaguet et al. 2021) with the function *wilcoxauc* (Argelaguet et al. 2021, Korsunsky et al. 2019). We declared a gene to be a significant domain-specific gene if its Benjamini-Hochberg adjusted p-value < 0.01 and log-fold change > 0 (details see Discussion). We performed GSEA analysis for each spatial domain using fgsea, a method for performing fast pre-ranked GSEA. For the human datasets (i.e., human DLPFC and breast cancer data), we focused on testing all human gene sets downloaded using the R package msigdbr. For the mouse datasets (i.e., mouse spermatogenesis, brain, and MOB data), we focused on testing mouse ontology gene sets (C5) downloaded using the R package msigdbr. We declared significantly enriched gene sets based on a Benjamini-Hochberg adjusted p-value threshold of 0.05.

Purity score analysis

We calculated a purity score either at the domain level or at the cell type level to quantify the percentage of the elongating spermatids in the ES domain. We then compared the two types of purity scores calculated in the control seminiferous tubules (e.g., WT3_Puck7 slice) with those calculated in the diabetic seminiferous tubules (e.g., Diabetes_Puck10 slice). Specifically, for the domain-level purity score, we examined one spatial location on each slice at a time, obtained its three nearest neighbors, and calculated the percentage of its neighbors being in the ES domain (domain #2). We then averaged such a percentage across all locations in the same domain (ES) to obtain the domain-level purity score. The domain-level purity score captures the mean percentage of times a neighboring location of an ES domain location also belongs to the ES domain. For cell type level purity score, we examined one spatial location on each slice at a time, obtained its three

nearest neighbors and the inferred ES cell type proportion on these neighbors, and then averaged such proportion across all locations that belong to the ES domain to obtain the cell type level purity score. The cell type level purity score captures the mean percentage of the ES cell type in a neighboring location of an ES domain location.

Human breast cancer subtype analysis

We classified the human breast cancer 10x Xenium data into eight clinical subtypes based on the expression of three important hormone receptor marker genes: *ESR1* (estrogen receptor), *PGR* (progesterone receptor), and *ERBB2* (human epidermal growth factor receptor 2, a.k.a. *HER2*). Specifically, we first calculated the number of spatial locations in each spatial domain that are positive for the expression of the three marker genes separately. Afterwards, we divided the data into eight subtypes based on whether each of the three genes were expressed or not: Triple Negative, ER+/PR-/HER2-/, ER-/PR+/HER2-, ER-/PR-/HER2+, ER+/PR+/HER2-, ER+/PR-/HER2+, ER-/PR+/HER2+, Triple Positive. We then calculated the percentage of the eight subtypes in each spatial domain detected by IRIS. In addition, we used two important tumor invasiveness marker genes *AGR3* and *CENPF* to quantify the invasiveness of four tumor domains (two DCIS and two IDC domains). Specifically, for each domain in turn, we calculated the percentage of its spatial locations whose marker gene expression is higher than the median level across all locations. Such percentage serves as the evidence for whether the expression of *AGR3* or *CENF* is high in the spatial domain.

4.5.4 Data and code availability

The original public data used in this work can be accessed through the following links: Human dorsolateral prefrontal cortex (DLPFC) data by 10x Visium available at the link:

http://spatial.libd.org/spatialLIBD/, with human post-mortem brain snRNA-seq reference data available at Synapse (https://www.synapse.org/#!Synapse:syn18485175); mouse spermatogenesis data by Slide-seq is available at the link https://www.dropbox.com/s/ygzpj0d0oh67br0/Testis_Slideseq_Data.zip?dl=0, with the mouse testis scRNA-seq reference data available at GEO accession GSE112393; mouse olfactory bulb by Stereo-seq data is available at https://db.cngb.org/stomics/mosta/download/, with the mouse olfactory bulb scRNA-seq data available at GEO accession GSE121891; human breast cancer by 10x Xenium data is available at https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast, with the with the human breast cancer scRNA-seq reference data available at GSE accession GSE176078; The IRIS software package and source code have been deposited at www.xzlab.org/software.html. All scripts used to reproduce all the analysis are also available at the same website.

4.6 Supplementary Figures

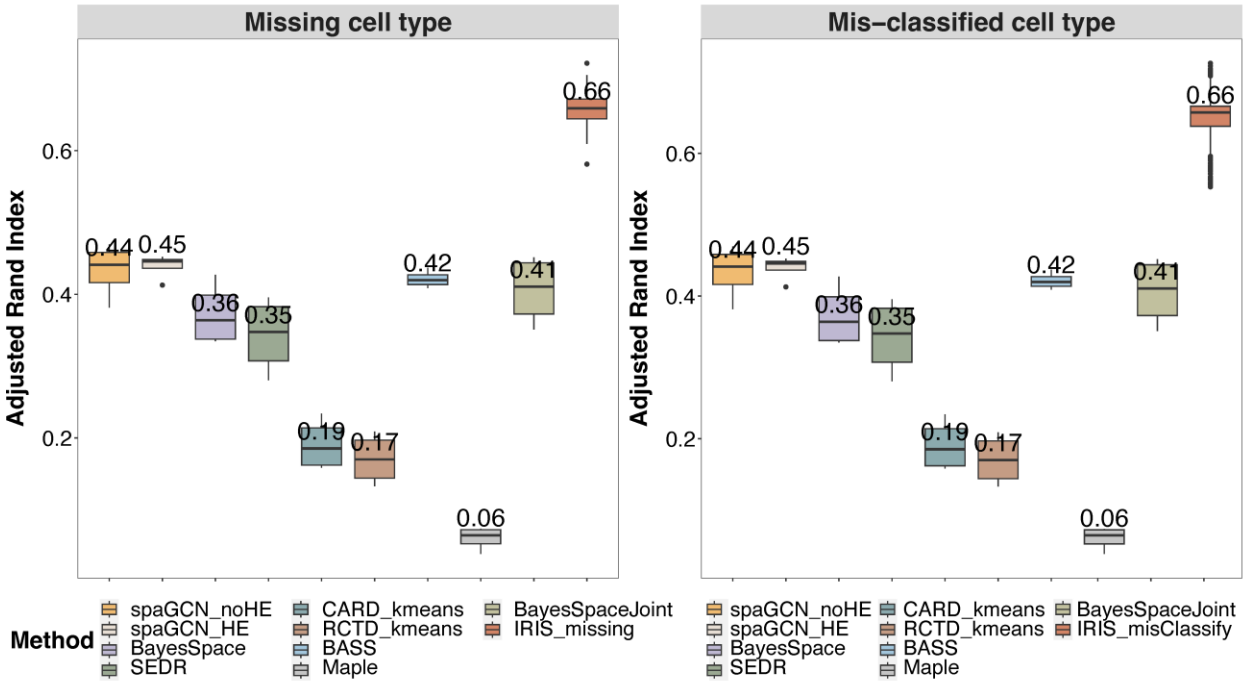


Figure S4.1 Evaluation on the spatial domain detection methods in the human dorsolateral prefrontal cortex (DLPFC) 10x Visium data for challenging settings.

The challenging settings include whether there are **missing cell types or misclassified cell type** in the scRNA-seq reference data. Boxplots display ARI (Adjusted Rand Index) values for different methods across all scenarios in each challenging setting. For example, in the missing cell type setting, we miss one cell type at a time in the scRNA-seq reference, and the boxplot for IRIS_missing calculates the median value of ARI across 44 missing cell type scenarios across 4 slices. In the mis-classified cell type setting, we randomly merge two cell types at a time as the misclassified cell type in the scRNA-seq reference, and the boxplot for IRIS_misClassify calculate the median value of ARI across 946 mis-classified cell type scenarios across 4 slices. Compared spatial domain detection methods (x-axis) include (1) single slice method: spaGCN without HE image (yellow), spaGCN with HE image (beige), BayesSpace (purple), and SEDR (green); (2) deconvolution-based method: CARD_kmeans (blue), RCTD_kmeans (tape); (3) multiple-slice methods: BASS (lake blue), Maple (grey), BayesSpaceJoint (matcha), and IRIS_missing / IRIS_misClassify (red). Clustering accuracy of different methods in recapitulating the true tissue domains is measured by ARI, with a higher ARI indicating higher accuracy. IRIS again detects the spatial domains accurately in both challenging settings, more so than other methods.

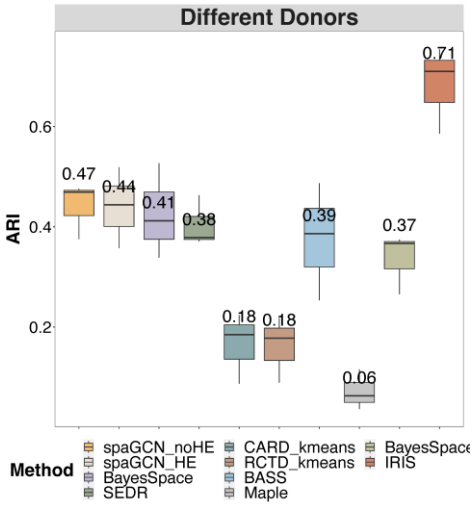


Figure S4.2 Evaluation on the spatial domain detection methods in the human dorsolateral prefrontal cortex (DLPFC) 10x Visium data for challenging settings.

The challenging settings include the setting where the tissue slices are from different donors (denoted as diffStr). Compared spatial domain detection methods (x-axis) include (1) single slice method: spaGCN without HE image (yellow), spaGCN with HE image (beige), BayesSpace (purple), and SEDR (green); (2) deconvolution-based method: CARD_kmeans (blue), RCTD_kmeans (tape); (3) multiple-slice methods: BASS (lake blue), Maple (grey), BayesSpaceJoint (matcha), and IRIS (red). Clustering accuracy of different methods in recapitulating the true tissue domains is measured by ARI, with a higher ARI indicating higher accuracy. IRIS again detects the spatial domains accurately, more so than other methods.

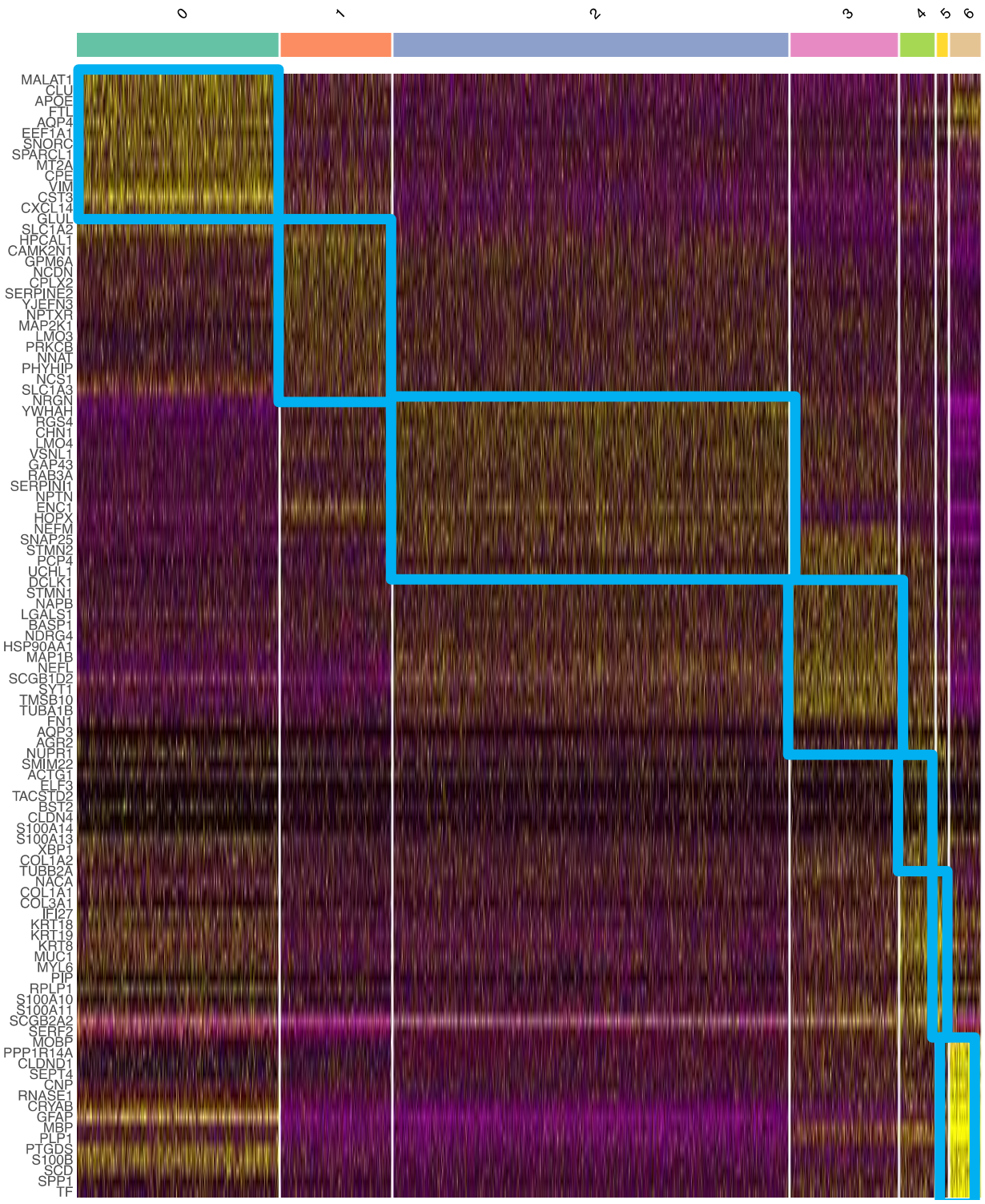


Figure S4.3 Heatmap of expression pattern of the domain specific genes.

Due to the limited space, we only display the top 15 selected domain specific genes. Yellow color represents a higher expression while purple color represents a lower expression. Here, the tissue slice is slice 151509 in the baseline same donor analysis.

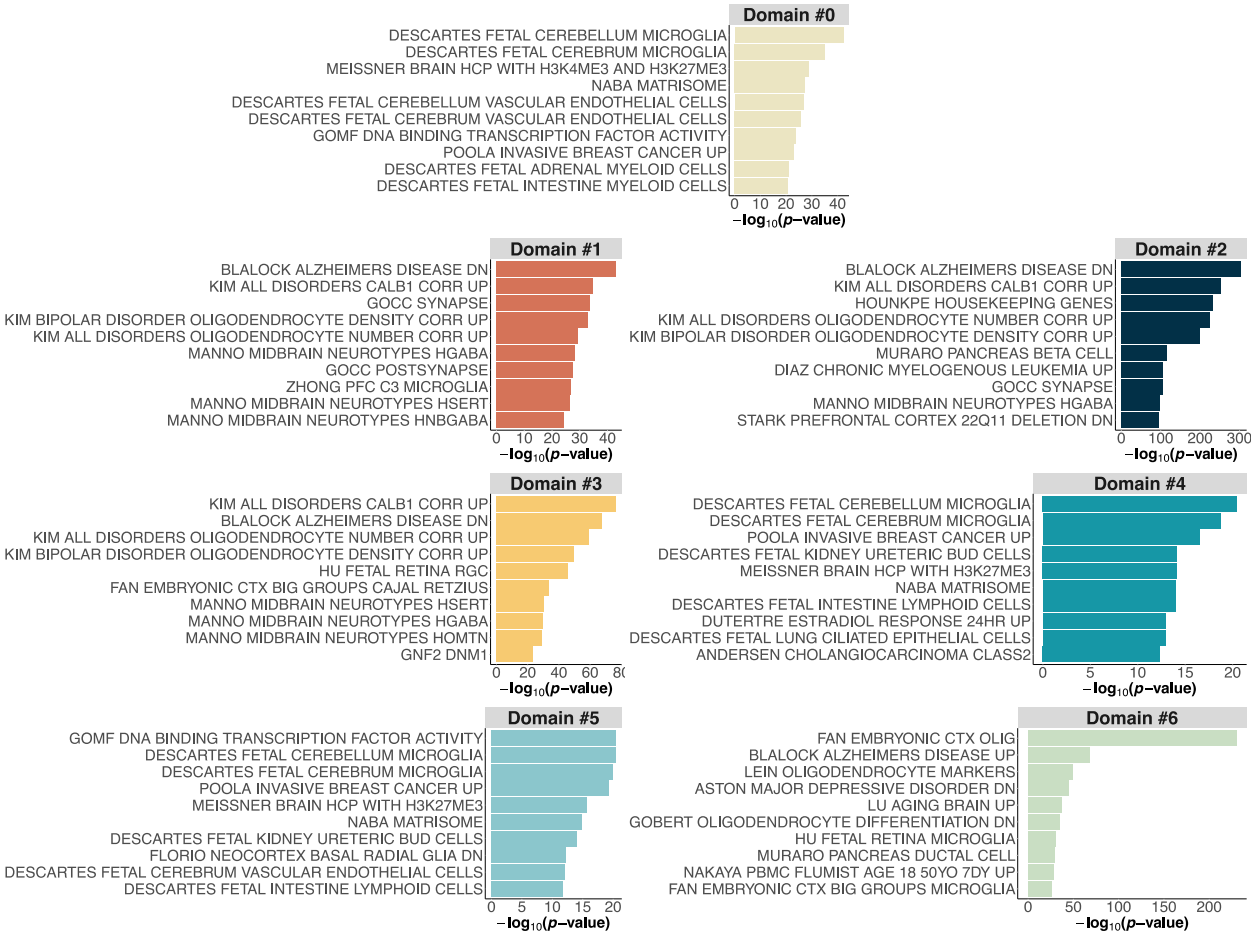


Figure S4.4 Gene set enrichment analysis on the domain-specific DE genes in the human DLPFC data.

The top 10 enriched gene sets are shown for each of the seven detected spatial domains. Here, the tissue slice is slice 151509 in the baseline same donor analysis.

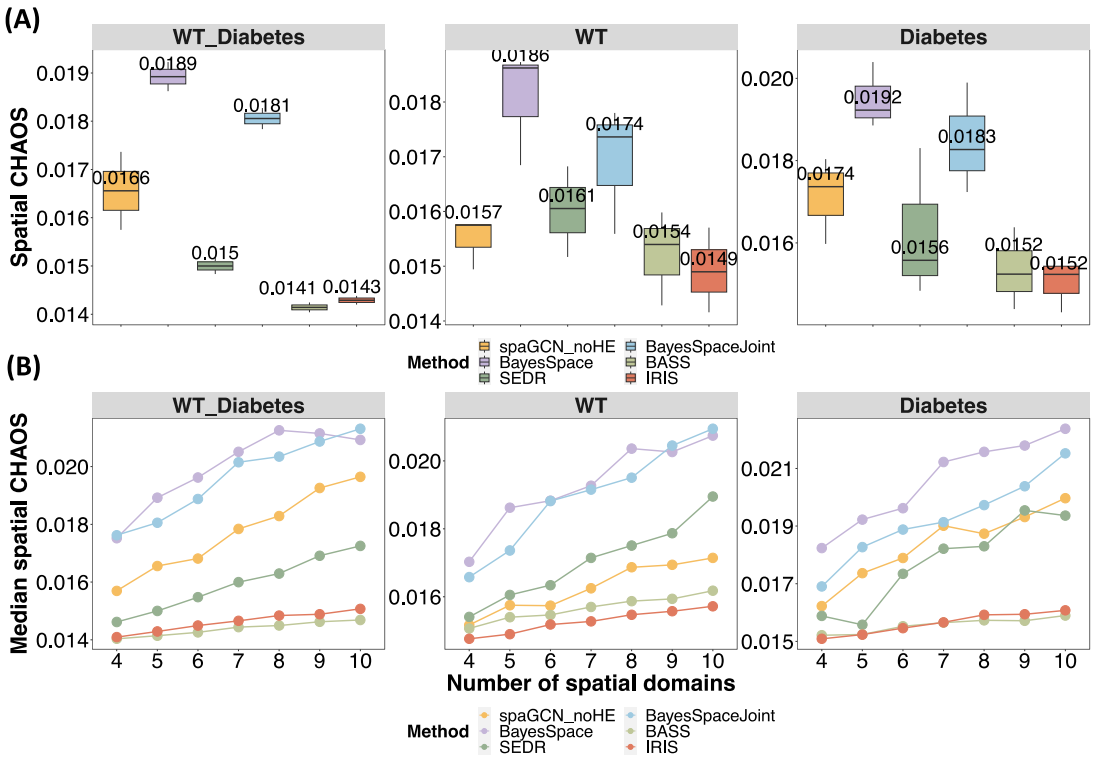


Figure S4.5 Evaluation on the robustness of different methods in different analysis settings.

Specifically, we performed the analysis on the slices that are from WT & Diabetes both, only from WT mice, and only from diabetic mice. **(A)** Boxplots display CHAOS values for different methods, which measure the spatial continuity and compactness of the detected spatial domains from different methods, when the number of spatial domains is set to be 5. **(B)** Line plots display the CHAOS values when varying the number of spatial domains. The median CHAOSs across all slices in each analysis setting were used. In general, the relative performance of these methods remains the same across other settings. IRIS consistently outperforms other methods.

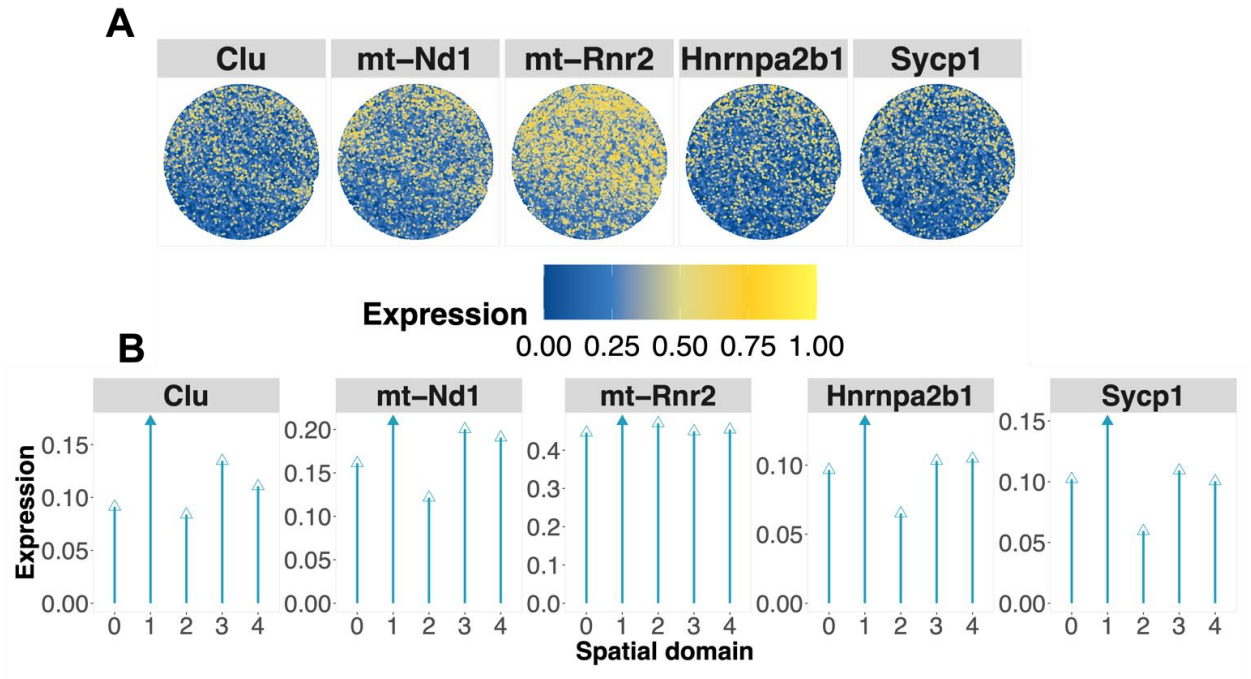


Figure S4.6 The spatial pattern of the domain #1 enriched mitochondrial genes and Sertoli cell marker genes in the WT mouse (WT3_Puck7) of the main analysis.

(A) Scatter plots (top) display the spatial distribution of important spermatogenesis related marker genes. **(B)** Lollipop plots (bottom) display the mean expression of important marker genes (y-axis) in each spatial domain identified by IRIS (x-axis). Solid arrow indicates the highest enrichment value. Specifically, several mitochondrial genes and Sertoli cells marker gene (*e.g.*, *Clu* (Zhang et al. 2015)) are enriched in spatial domain 1, suggesting it is the germinal epithelium structure.

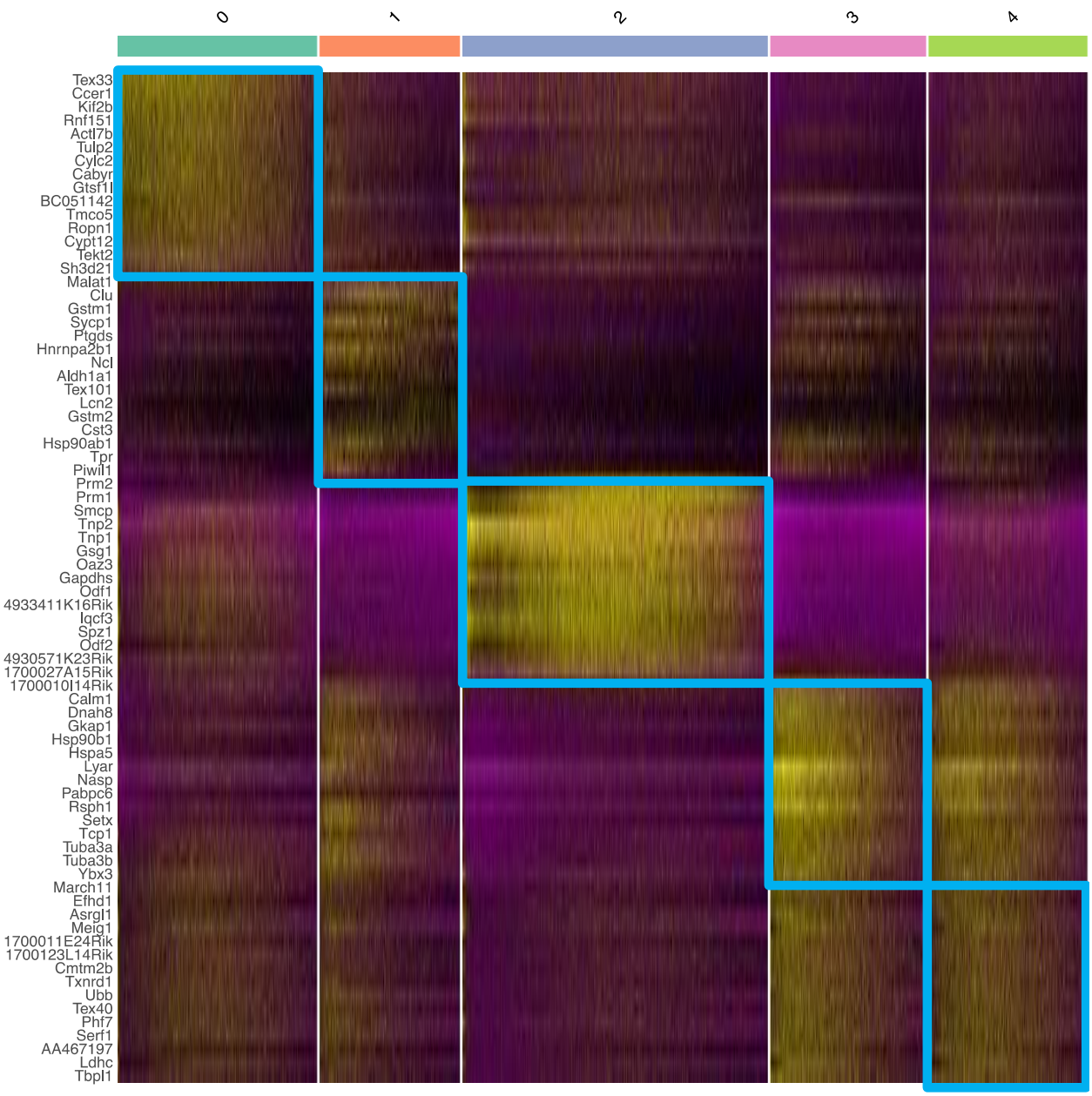


Figure S4.7 Heatmap of expression pattern of the domain specific genes.

Due to the limited space, we only display the top 15 selected domain specific genes. Yellow color represents a higher expression while purple color represents a lower expression. Here, the tissue slice is the slice WT3_Puck7 in the main analysis.

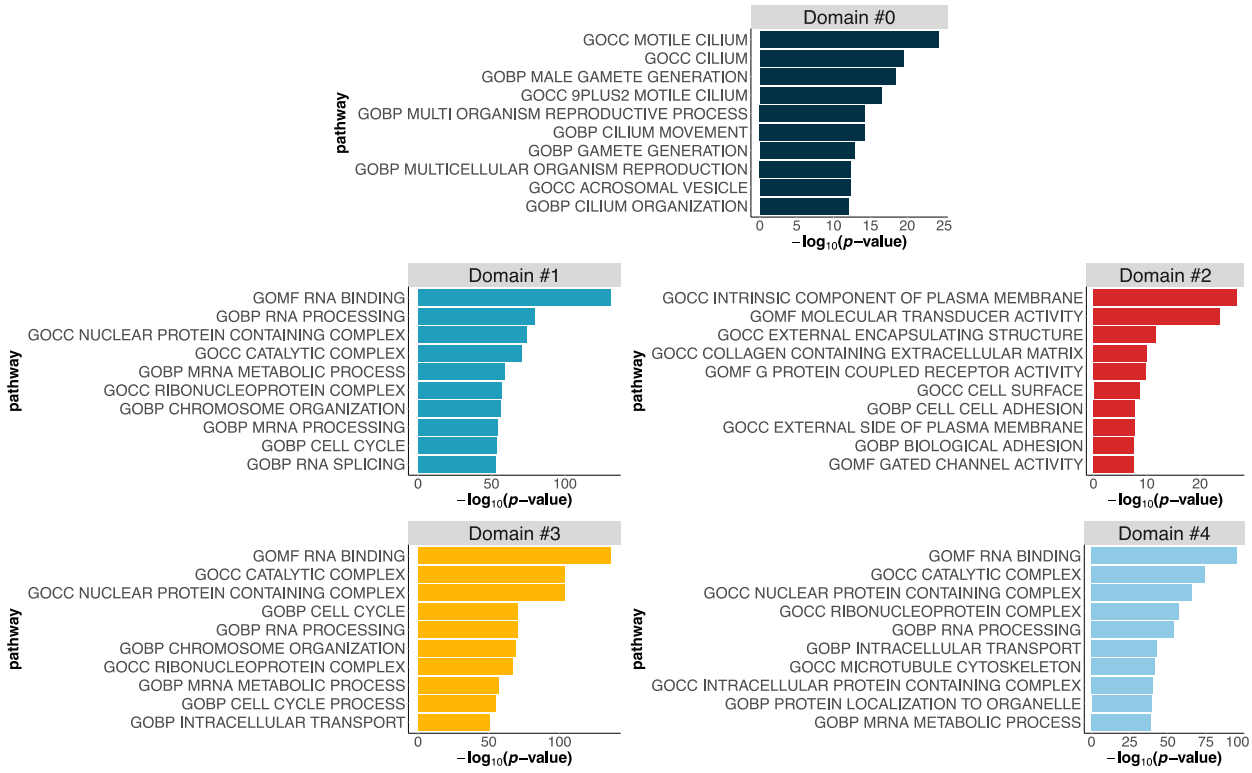


Figure S4.8 GSEA results on each spatial domain detected by IRIS in the mouse WT slice (WT3_Puck7) in the main analysis of the Slide-seq data.

Here Top enriched gene sets in all spatial domains were displayed.

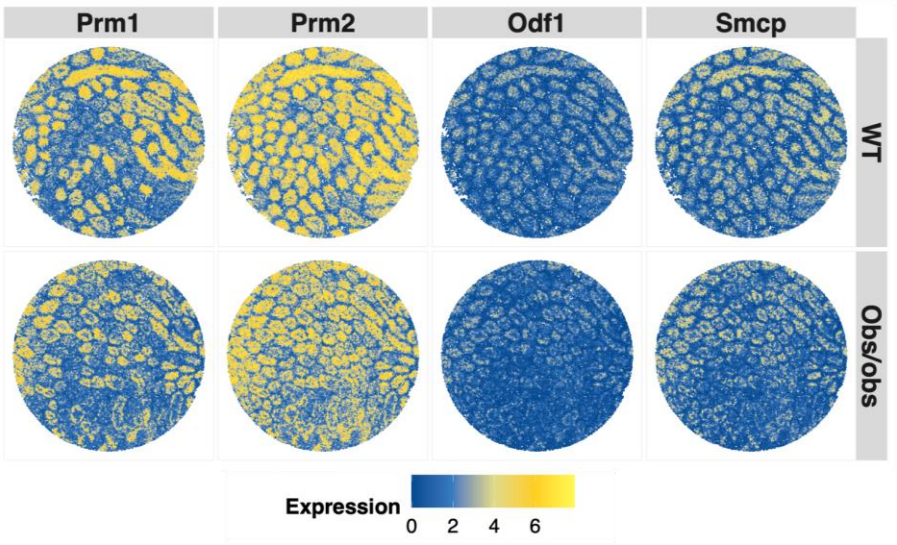


Figure S4.9 Comparison of the spatial pattern of ES marker genes in the WT mice (WT3_Puck7) and ob/ob (diabetic, Diabetes2_Puck10) mouse.

Scatter plot displays the spatial distribution of the marker genes. Specifically, the spatial arrangement of the spermatids was disrupted under the diabetic conditions.

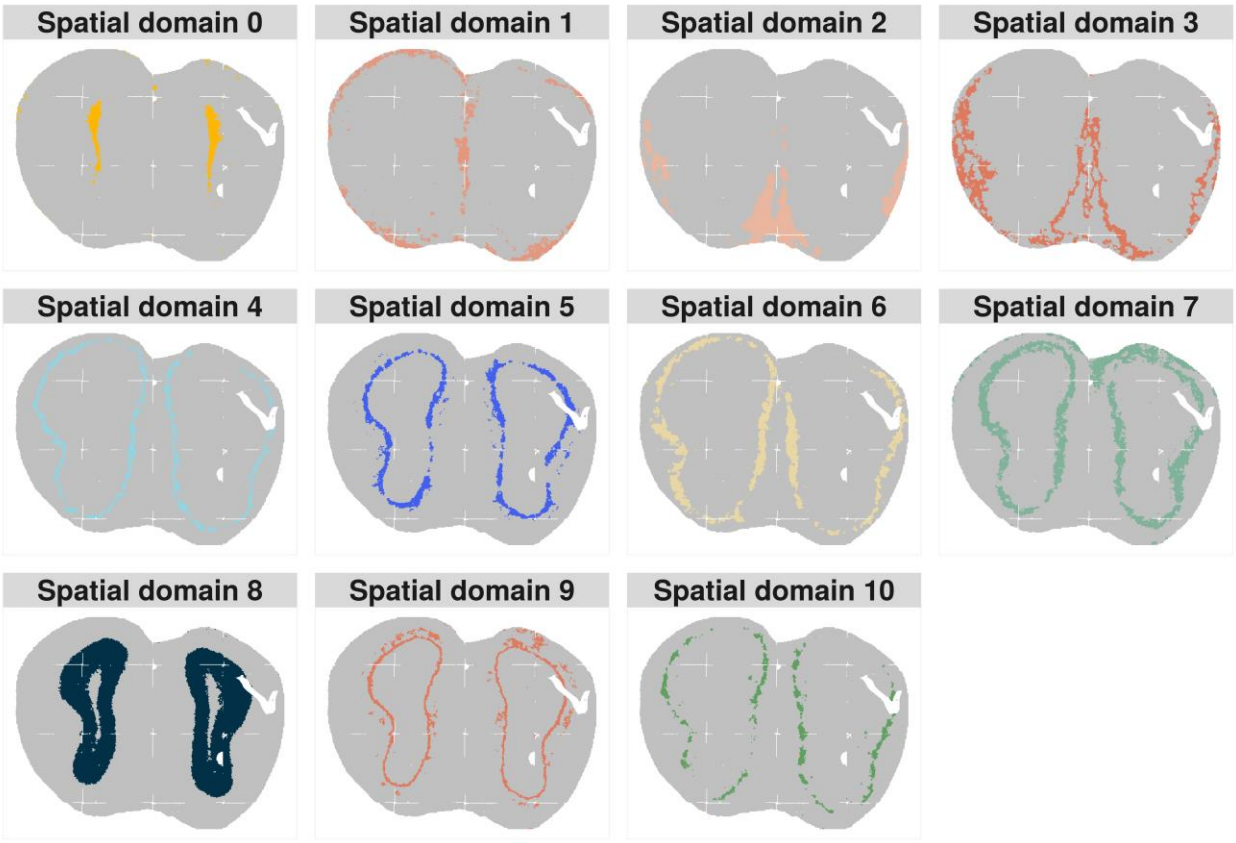


Figure S4.10 Spatial domains identified by IRIS in the slice S1 in the main analysis of the mouse olfactory bulb stereo-seq data.

This scatterplot displays the spatial domains separately. IRIS is capable of accurately and incisively depicting the layered structure of the mouse olfactory bulb.

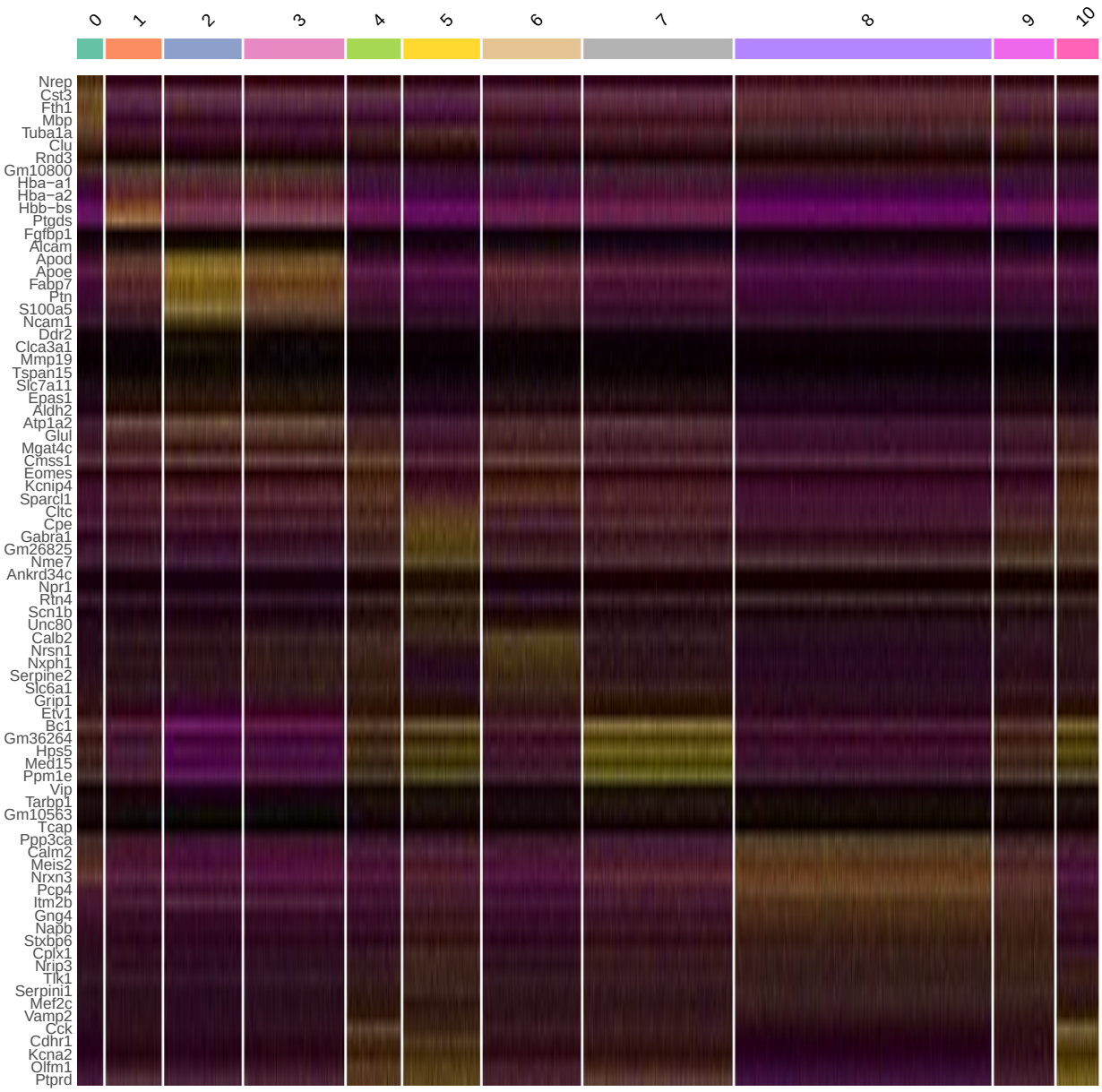


Figure S4.11 Heatmap of expression pattern of the domain specific genes.

Due to the limited space, we only display the top 5 selected domain specific genes. Yellow color represents a higher expression while purple color represents a lower expression. Here, the tissue slice is slice S1 in the main analysis.

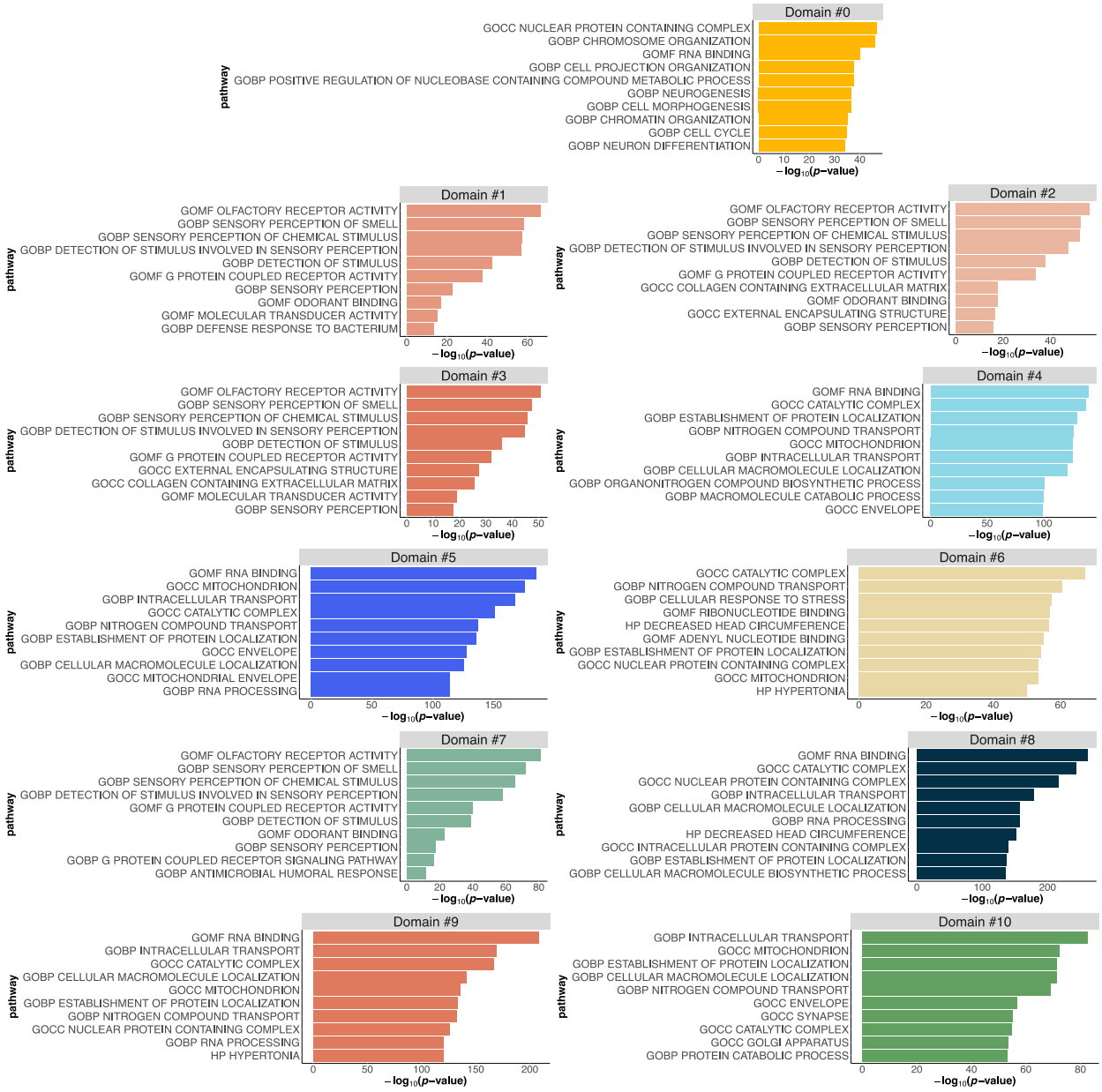


Figure S4.12 GSEA analysis results of the slice S1 in the main analysis of the mouse olfactory bulb data.

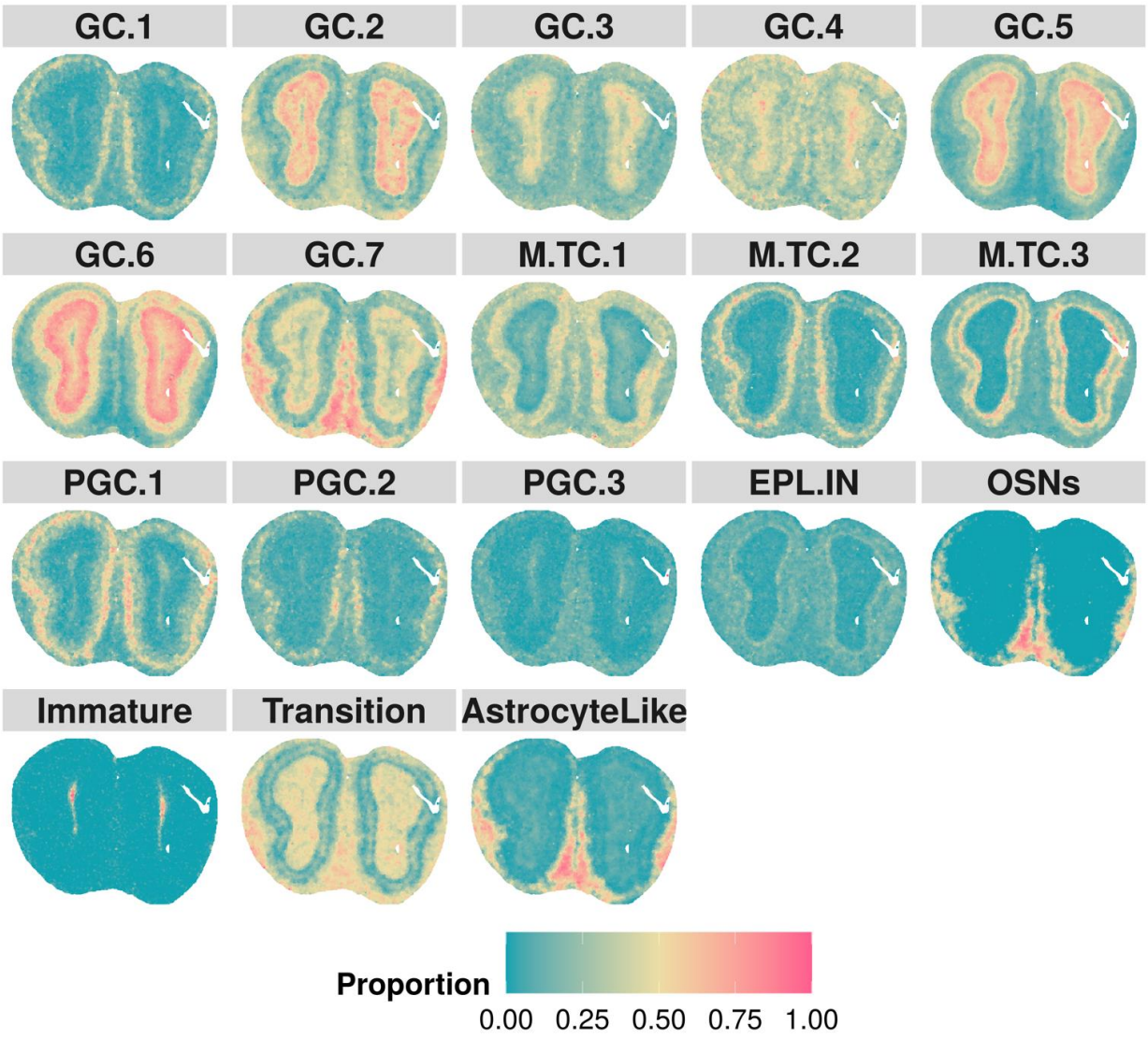


Figure S4.13 Spatial scatter plot displays the spatial distribution of IRIS estimated cell type proportion across spatial locations in the slice S1 in the main analysis of the mouse olfactory bulb stereo-seq data.

Specifically, cell type proportions estimated by IRIS can accurately depict the layered structure of mouse olfactory bulb. Here, for each cell type, the cell type proportion was scaled to 0-1 range. Color was shown to represent the 0-1 range of cell type proportions correspondingly.

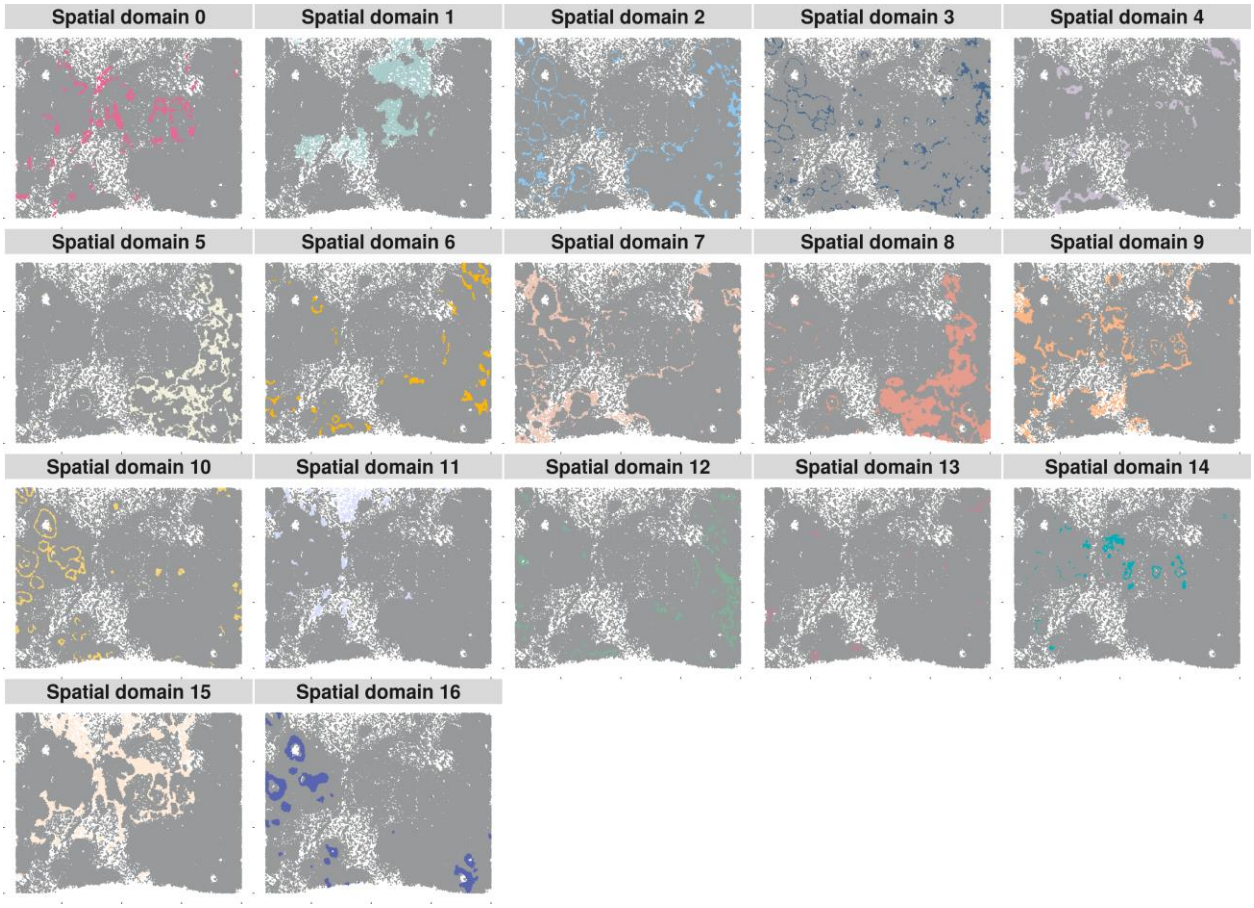


Figure S4.14 Spatial domains identified by IRIS in the slice Rep1 in the main analysis of the human breast cancer 10x Xenium data.

This scatterplot displays the spatial domains separately. IRIS is capable of accurately and incisively pinpointing the spatial locations of four distinct tumor domains (domain #5, #8, #10, and #16).

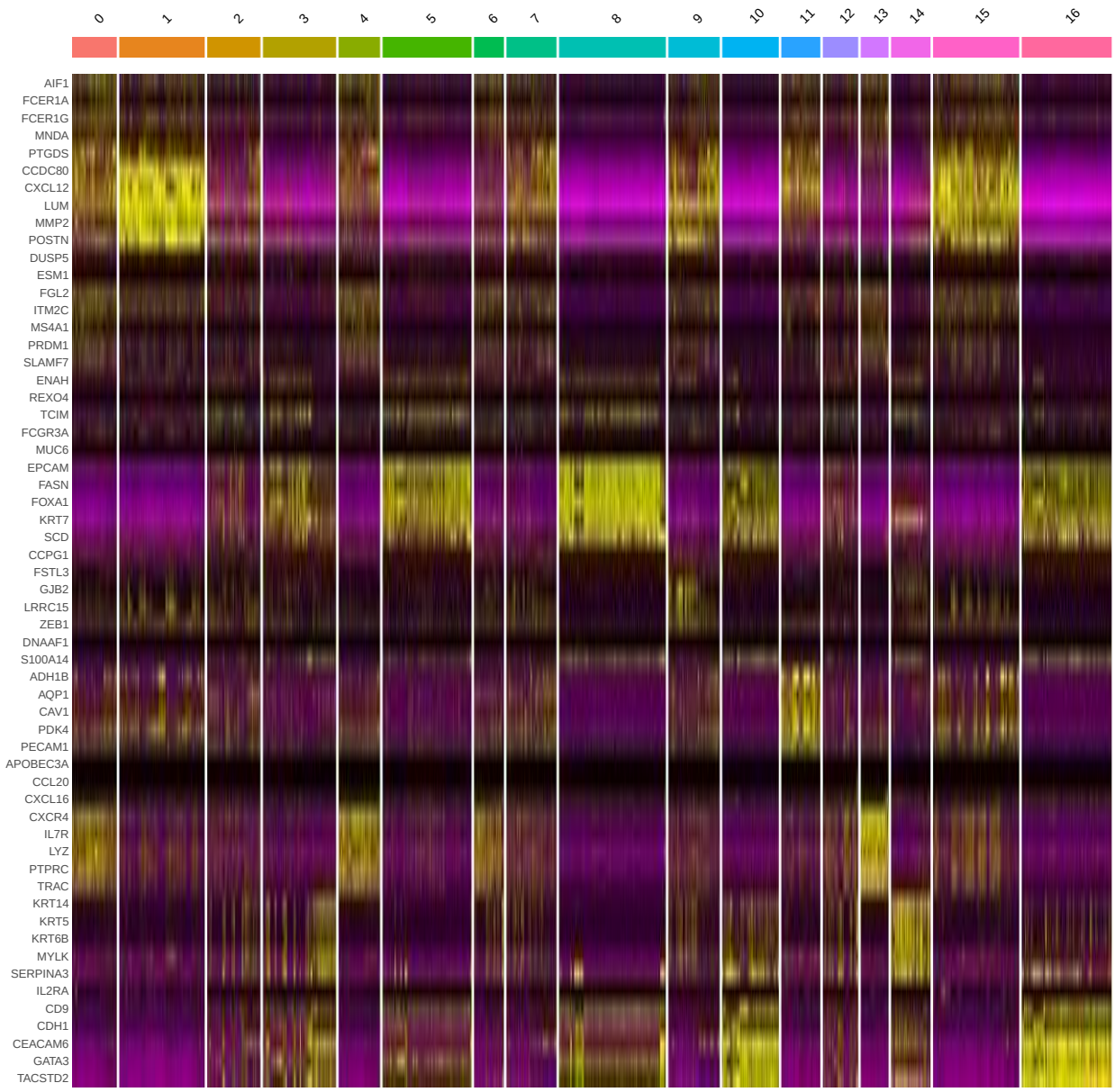


Figure S4.15 Heatmap of expression pattern of the domain specific genes in the human breast cancer 10x Xenium data.

Due to the limited space, we only display the top 5 selected domain specific genes. Yellow color represents a higher expression while purple color represents a lower expression. Here, the tissue slice is slice Rep1 in the main analysis.

Figure S4.16 GSEA analysis results of the slice Rep1 in the human breast cancer 10x Xenium data.

GSEA further revealed that the tumor related domain specific DE genes (domain #5, #8, #10, and #16) detected by IRIS are highly enriched in luminal-like breast cancer cell lines, and HER2/ERBB2 breast cancer cell line while immune-related domain specific DE genes (domain #0, #4, and #13) are highly enriched in immune cells pathways, lymphocyte activation pathway, monocyte macrophage cells, NK cell, T cells and immune response related pathways.

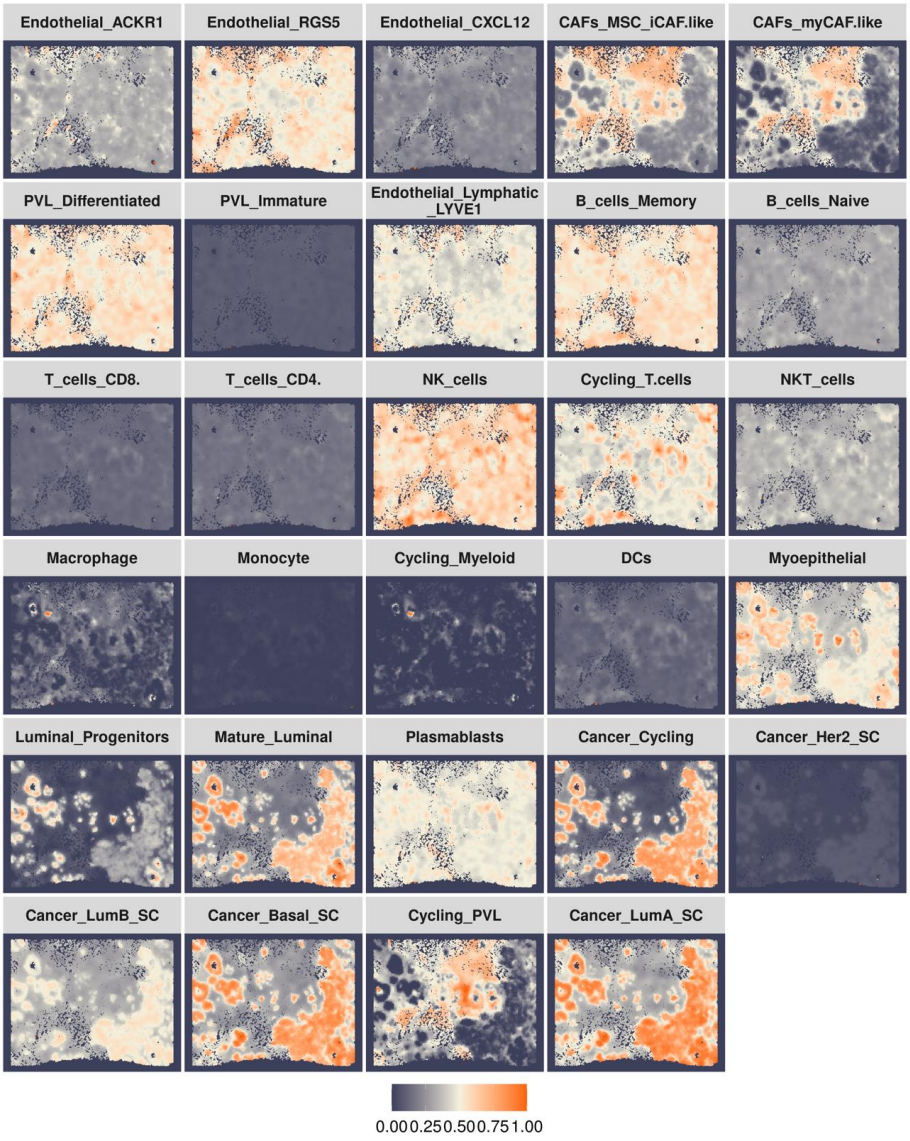


Figure S4.17 Spatial scatter plot displays the spatial distribution of IRIS estimated cell type proportion across spatial locations in the slice Rep1 of the human breast cancer 10x Xenium data.

Specifically, IRIS correctly locates cancer related cell type proportions (e.g., cancer cycling cells, cancer Her2 cells, cancer luminal B cells, cancer basal cells, and cancer luminal A cells) in cancer domains while locating immune cells (e.g., B cells, NK cells, T cells) into immune related domains. Here, for each cell type, the cell type proportion was scaled to 0-1 range. Color was shown to represent the 0-1 range of cell type proportions correspondingly.

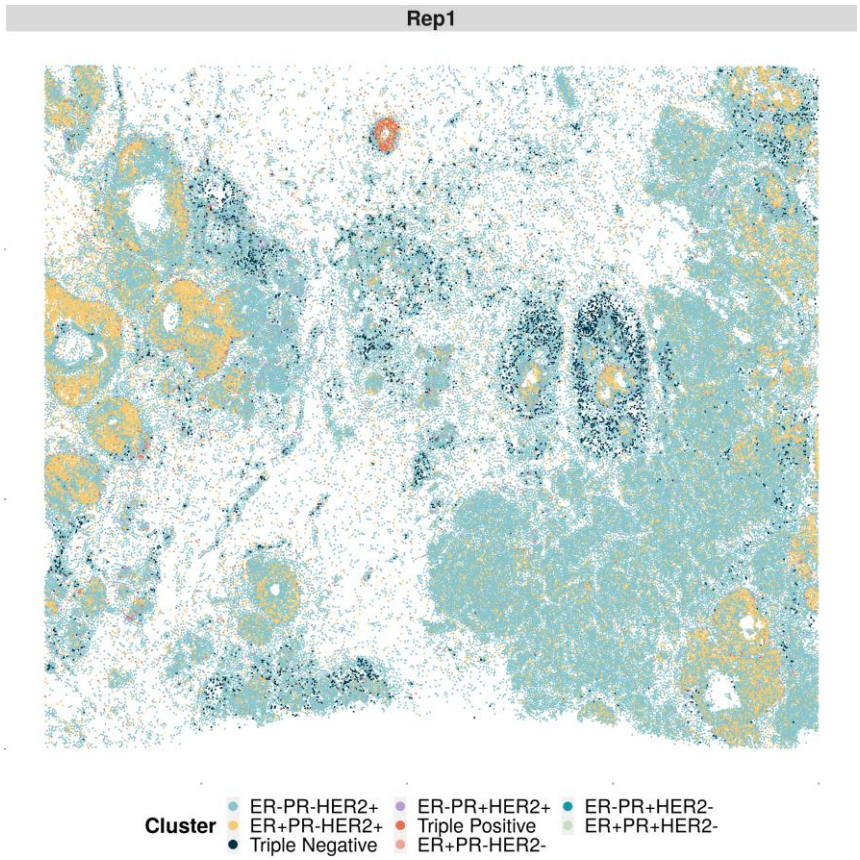


Figure S4.18 Breast cancer subtypes classified by the expression of hormonal receptors in each spatial location.

Specifically, in the slice Rep1 of the human breast cancer 10x Xenium dataset, each spatial location can be classified based on the expression of three hormone receptors: estrogen receptor- 1(ER), progesterone receptor (PR) and human epidermal growth factor receptor-2 (HER2), breast cancer can be divided into eight different subtypes.

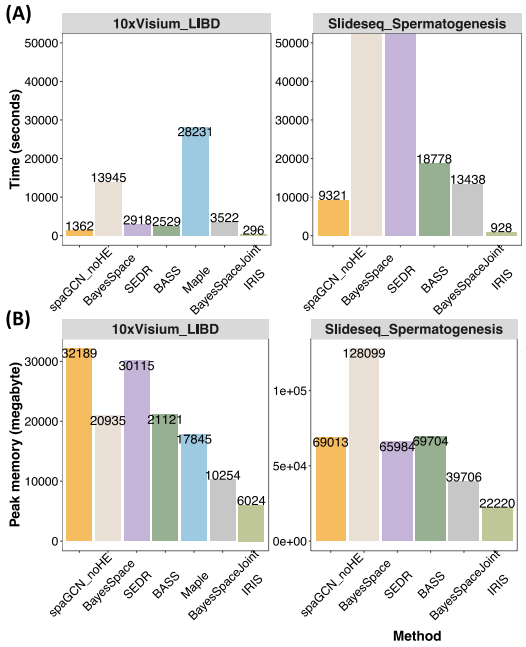


Figure S4.19 Computation time (minutes top panel) and (B) peak memory usage (MB, bottom panel) for spatial domain detection method on moderate-sized real spatial transcriptomics datasets.

Computation of IRIS, spaGCN, BayesSpace, BayesSpace_Joint, and SEDR was performed on a single core of an Intel(R) Xeon(R) CPU E5-2683 v3 2.00GHz processor. IRIS is computationally fast and efficient compared with other methods in the moderate-sized 10x Visium human DLPFC, Slide-seq mouse spermatogenesis. IRIS is also the only method that is scalable to the large mouse MOB stereo-seq and human breast cancer 10x Xenium datasets and can finish analysis there in 26 ~ 28 minutes with 15.5-26.1 GB physical memory.

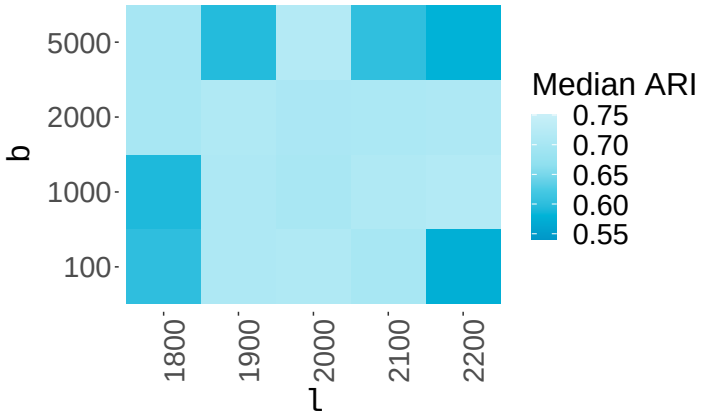


Figure S4.20 Selection of the penalty parameters according to the performance of IRIS in baseline analysis of the human DLPFC dataset.

ARI results when we varied the values of $\boldsymbol{\lambda}$ from and the values of $\boldsymbol{\beta}$ from 10, 100, 500, 1000 to 10000. Specifically, the median ARI across slices is robust to a wide range of the values of $\boldsymbol{\beta}$ and a reasonable range of the values of $\boldsymbol{\lambda}$.

Chapter 5 Conclusion

The dissertation has focused on developing efficient statistical methods and computational methods to address various analytical challenges in genomics and genetics. These challenges typically arise with the high-dimensional data generated by rapidly evolving sequencing techniques, e.g., single-cell RNA-seq (scRNA-seq), and spatially resolved transcriptomics (SRT). I presented iDEA for integrative differential expression (DE) and gene set enrichment (GSE) analysis for scRNA-seq data in chapter 2, CARD for spatially informed cell type deconvolution for SRT data in chapter 3, and IRIS for integrative and reference-informed spatial domain detection for multiple-slice SRT data in chapter 4. I have demonstrated these methods by showing results from extensive simulations and comprehensive real data applications. Together, the applications of these methods have advanced our understanding in cellular heterogeneity, tissue cell type composition and organization, and the underlying mechanisms of disease etiology. Below, I review these projects, discuss their limitations, and envision possible future directions.

In Chapter 2, I have focused on performing joint differential expression and gene set enrichment analysis through a hierarchical Bayesian framework using DE summary statistics as input. By integrating these two important analyses, our method iDEA can improve the power and consistency of DE analysis and accuracy of GSE analysis. Specifically, we assume that the true DE effect size follows a mixture of two distributions depending on whether the gene is a DE gene or not. However, the distribution of effect sizes between up-regulated and down-regulated genes might be different while the enrichment effect of a gene set associated with the proportion of being up-regulated or down-regulated genes might also be different. For example, previous studies of

malignant cell transformation show that the majority of DE genes are down-regulated and relate to a diverse set of functions such as extracellular matrix production, cell adhesion, while a minority of the differentially expressed genes are up-regulated and highly enriched for cellular proliferation control related gene sets (Danielsson et al. 2013). As another example, a recent study has modeled the log fold change of genes using a mixture of three Gaussian distributions to capture the non-DE, positive, and negative DE genes respectively and then perform the likelihood ratio test to test each gene set (Makrooni et al. 2022). Previous results have shown that analyzing up- and down-regulated genes separately can further improve the power than analyzing all the genes together (Makrooni et al. 2022, Hong et al. 2014). Therefore, extending iDEA's modeling assumption to three component Gaussian mixtures may help improve the power of detecting both DE genes and enriched gene sets. This will also facilitate the interpretation of gene sets enriched in up- and down-regulated DE genes. While we have primarily focused on analyses comparing two different cell types in scRNA-seq data, an increasing number of scRNA-seq studies focus on replicated multi-condition experiments to study population-specific changes in expression between conditions. Besides multi-condition scRNA-seq data, spatially resolved transcriptomics studies have identified several spatially expressed genes on tissue sections, which advances our understanding in the tissue organization. Therefore, exploring the utility of iDEA in multi-condition and spatial contexts might further provide novel biological insights in gene set enrichment patterns between different experimental conditions as well as spatial organization and structure.

In Chapter 3, I propose a spatially informed cell type deconvolution method CARD for SRT studies. CARD incorporates the conditional autoregressive (CAR) modeling assumption into the non-negative matrix factorization framework to accurately estimate the cell type compositions and reconstruct a high-resolution map for each spatial location. The results of extensive

simulations and real data applications are a demonstration of both the advantages of integrating cell type specific expression from scRNA-seq data and spatial correlation structure into the deconvolution framework. Specifically, CARD makes use of a fast optimization algorithm for the inference of cell type composition matrix and only outputs the point estimate. However, this estimation procedure ignores the uncertainty in cell type composition across spatial locations. Although many deconvolution methods have been proposed in both bulk RNA-seq and spatial transcriptomics fields, only a few methods have been proposed for quantifying the uncertainty associated with cell type proportions in bulk RNA-seq data (Cai et al. 2022, Vellame et al. 2023) and no method has been proposed to address this challenge in spatial transcriptomics deconvolution analysis (Sun, Ma and Zou 2023). Lack of consideration of these uncertainties can lead to missed or false findings in downstream analysis, i.e., cell type specific differential expression analysis. Indeed, previous studies in bulk RNA-seq deconvolution have shown that considering uncertainty in cell type proportions improves the accuracy in detecting the cell type specific DE genes between the patients of Alzheimer's disease and healthy controls (Cai et al. 2022). Therefore, incorporating the uncertainty measure in SRT data deconvolution might further benefit the down-stream analysis, i.e., spatial domain specific DE analysis, high-resolution enhancement. Additionally, in Chapter 3, we have only investigated two versions of CARD: reference-based (the main version) and reference-free. In particular, the reference-based version uses the cell type specific gene expression profile predefined by scRNA-seq data while the reference-free version only requires the marker gene list for the deconvolution of SRT data. However, both technical and biological batch variation exist in gene expression between SRT scRNA-seq reference data. It could be beneficial to model the prior distribution of reference basis matrix to further improve the deconvolution accuracy.

In Chapter 4, I develop an integrative reference-informed spatial domain detection method IRIS for SRT studies. I have primarily focused on modeling the primary feature of cell type compositional heterogeneity across spatial locations while accounting for the spatial relationship within each tissue slice and cell type compositional similarity in the same domain across slices. However, joint alignment of multiple tissue slices to reconstruct a three-dimensional (3D) tissue structures remains a challenge. Effective reconstruction of 3D tissue (Wang et al. 2023) organization is necessary for identifying 3D spatial tissue domains, estimating 3D cell type proportion distributions, and further improve the interpretability of results as well as revealing underlying biological mechanisms. Besides 3D reconstruction, batch effects correction has become an important problem with the increasing scale of SRT datasets from multiple slices. Extension of IRIS into a spatial factor analysis model to learn the shared embeddings of expression across slices can provide the first step towards identifying DE genes between multiple conditions (Liu et al. 2023). In addition to integrative analysis of multiple transcriptomics data, effectively integrating other data modalities such as histological images accompanied with SRT data will help with the integrative spatial domain detection analysis. For example, we can learn a consensus between spatial domains inferred by transcriptomics and those inferred by cell morphological features extracted from the image. Future methodological development for efficiently integrating image information into IRIS modeling framework may further improve the domain detection accuracy of IRIS. Finally, SRT datasets represent only a fraction of the multi-dimensional information encoded within cells. There has been a growing interest in spatial multi-omics studies, including transcriptomics, proteomics, and epigenomics to obtain a comprehensive understanding of biological systems. Further extension of IRIS into a multi-omics integration will enhance our understanding of cellular heterogeneity, tissue function, and further unravel disease mechanisms.

In summary, throughout the dissertation, I have developed three statistical and computational methods to address different challenges including: DE genes and enriched gene set detection, cell type composition estimation and biologically interpretable spatial domain detection. These methods have allowed us to reveal cellular heterogeneity, to identify coordinated gene expression patterns, and to better understand the role of cell type composition as well as tissue structure in biological processes. By harnessing the power of these methods, we are not only advancing our fundamental knowledge of biology but also paving the way for precision medicine and personalized therapeutics. We believe that these developed methods will serve as valuable tools for researchers to continue shaping and driving future investigations in this exciting and rapidly evolving field.

Appendices

Appendix A. Chapter 2 (iDEA) Supplementary Text.

A.1 EM-MCMC Inference Algorithm.

The iDEA model is described in detail in the **Methods**. Here, we describe the detailed algorithm for inference. As explained in the main text, our goal is to infer the posterior probability of $\gamma_j = 1$ as evidence for *j*-th gene being DE and test the null hypothesis H_0: $\tau_1 = 0$ that DE genes are not enriched in the gene set. To achieve both goals, we develop an efficient expectation maximization (EM)-Markov chain Monte Carlo (MCMC) algorithm. To simplify notation**,** we denote $\boldsymbol{\beta}$ as the *p*-vector of the underlying true effect sizes, or $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^T$. We denote $\boldsymbol{\gamma}$ as the *p*-vector of the indicator variables, or $\boldsymbol{\gamma} = \left(\gamma_1, \gamma_2, \cdots, \gamma_p\right)^T$. We denote $\sigma_{e_j}^2$ as the variance of the marginal DE effect size estimate for *j*-th gene, or $\sigma_{e_j}^2 = \mathrm{se}^2(\hat{\beta}_j)$. We treat both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as missing data and write out the complete likelihood as

$$\begin{aligned}
\log\Pr\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \tau_0, \tau_1, \sigma_\beta^2\right) &= \log\left\{\Pr\left(\widehat{\boldsymbol{\beta}} \middle| \boldsymbol{\beta}, \boldsymbol{\gamma}\right) \Pr\left(\boldsymbol{\beta} \middle| \boldsymbol{\gamma}, \sigma_\beta^2\right) \Pr(\boldsymbol{\gamma} | \tau_0, \tau_1) \Pr\left(\sigma_\beta^2 \middle| a_\beta, b_\beta\right)\right\} \\
&= -\tfrac{1}{2}\textstyle\sum_{j=1}^{p} \gamma_j \left(\log\left(\sigma_{e_j}^2\right) + \frac{\left(\hat{\beta}_j - \beta_j\right)^2}{\sigma_{e_j}^2}\right) \\
&\quad -\tfrac{1}{2}\textstyle\sum_{j=1}^{p} \gamma_j \left(\log\left(\sigma_{e_j}^2 \sigma_\beta^2\right) + \frac{\beta_j^2}{\sigma_{e_j}^2 \sigma_\beta^2}\right) \\
&\quad + \textstyle\sum_{j=1}^{p} \gamma_j \log\left(\pi_j\right) + \left(1 - \gamma_j\right) \log\left(1 - \pi_j\right) \\
&\quad - \left(a_\beta + 1\right) \log\left(\sigma_\beta^2\right) - b_\beta \sigma_\beta^{-2}\,, \qquad \text{(A1)}
\end{aligned}$$

where we have also ignored the constant terms in the above equation and $\pi_j = \frac{\exp(\tau_0 + a_j\tau_1)}{1+\exp(\tau_0 + a_j\tau_1)}$. With the above complete likelihood, we can derive the expectation step (E-Step) and maximization step (M-Step) as follows.

A.1.1 Expectation step (E-step)

In the E-Step, we obtain the expectation of equation (A1)

$$Q = E\left[\log \Pr\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}, \boldsymbol{\Gamma} | \boldsymbol{\tau}, \sigma_\beta^2\right)\right], \tag{A2}$$

which involves evaluating the expectations $E(\gamma_j), E(\gamma_j\beta_j)$ and $E(\gamma_j\beta_j^2)$. These expectations are obtained under the conditional distributions $\mathrm{P}(\beta_j, \gamma_j | \widehat{\boldsymbol{\beta}}, \tau_0^{(t)}, \tau_1^{(t)}, \left(\sigma^{(t)}\right)_\beta^2)$, with $\tau_0^{(t)}, \tau_1^{(t)}$ and $\left(\sigma^{(t)}\right)_\beta^2$ being the estimates from the previous iteration *t*. These conditional distributions are unfortunately not available in analytic forms. Therefore, we use Markov Chain Monte Carlo (MCMC) to obtain these expectations. Specifically, we develop a Gibbs sampling to sample the posterior distributions for β_j and γ_j in an alternate fashion. Afterwards, we use these posterior samples to evaluate the above expectations. To do so, we first integrate out β_j from the complete likelihood and obtain the conditional distribution for γ_j as

$$\Pr\left(\gamma_j = 1 \middle| \widehat{\boldsymbol{\beta}}, \tau_0^{(t)}, \tau_1^{(t)}, \left(\sigma^{(t)}\right)_\beta^2\right) \propto \exp\left\{\frac{m_j^2}{2s_j^2} + \log(s_j) - \log\left(\sigma_\beta^{(t)}\right) + \log\left(\pi_j^{(t)}\right)\right\}, \tag{A3}$$

$$\Pr\left(\gamma_j = 0 \middle| \widehat{\boldsymbol{\beta}}, \tau_0^{(t)}, \tau_1^{(t)}\right) \propto 1 - \pi_j^{(t)}. \tag{A4}$$

Then posterior distribution of γ_j is,

$$\gamma_j \sim \text{Bernoulli}\left(\frac{\Pr\left(\gamma_j = 1 \middle| \widehat{\boldsymbol{\beta}}, \tau_0^{(t)}, \tau_1^{(t)}, \left(\sigma^{(t)}\right)_\beta^2\right)}{\Pr\left(\gamma_j = 1 \middle| \widehat{\boldsymbol{\beta}}, \tau_0^{(t)}, \tau_1^{(t)}, \left(\sigma^{(t)}\right)_\beta^2\right) + \Pr\left(\gamma_j = 0 \middle| \widehat{\boldsymbol{\beta}}, \tau_0^{(t)}, \tau_1^{(t)}, \left(\sigma^{(t)}\right)_\beta^2\right)}\right) \tag{A5}$$

where $m_j = \frac{\widehat{\beta}_j}{1+\left(\sigma_\beta^{(t)}\right)^{-2}}$ and $s_j^2 = \frac{\sigma_{\widehat{e}_j}^2}{1+\left(\sigma_\beta^{(t)}\right)^{-2}}$. Next, we recognize from the complete likelihood that the conditional distribution of β_j given $\gamma_j = 1$ is normal:

$$\beta_j|\gamma_j = 1 \sim N\left(m_j, s_j^2\right) \tag{A6}$$

Certainly, $\beta_j = 0$ if $\gamma_j = 0$.

A.1.2 Maximization step (M-step)

In the M-Step, we obtain the parameter estimates for τ_0, τ_1 and σ_β^2 that maximize the Q function obtained in the E-Step. For τ_0 and τ_1, we obtain the first derivatives of the Q function with respect to each parameter as

$$\frac{\partial Q}{\partial \tau_0} = \sum\nolimits_{j=1}^{p} \left(E(\gamma_j) - \pi_j\right),$$

$$\frac{\partial Q}{\partial \tau_1} = \sum_{j=1}^{p} a_j \left(E(\gamma_j) - \pi_j\right). \tag{A7}$$

We also obtain the second derivatives as

$$\frac{\partial^2 Q}{\partial \tau_0^2} = \sum\nolimits_{j=1}^{p} \pi_j\left(1 - \pi_j\right),$$

$$\frac{\partial^2 Q}{\partial \tau_0 \partial \tau_1} = \sum\nolimits_{j=1}^{p} a_j \pi_j\left(1 - \pi_j\right),$$

$$\frac{\partial^2 Q}{\partial \tau_1^2} = \sum_{j=1}^{p} a_j^2 \pi_j\left(1 - \pi_j\right). \tag{A8}$$

where π_j is calculated as the expectation of the indicator variable γ_j $\mathrm{E}\left(\gamma_j \middle| \widehat{\boldsymbol{\beta}}, \tau_0^{(t)}, \tau_1^{(t)}, \left(\sigma^{(t)}\right)_\beta^2\right)$. And π_j is used in the following Newton-Raphson algorithm to obtain the parameter estimate of

the intercept τ_0 and gene set coefficient τ_1. Afterwards, we use the Newton-Raphson algorithm for optimization and obtain estimates of $\tau_0^{(t+1)}$ and $\tau_1^{(t+1)}$.

For σ_β^2, we obtain the first derivatives of the Q function with respect to σ_β^2 as

$$\frac{\partial Q}{\partial \sigma_\beta^2} = \sigma_\beta^{-4}\left(\sum_{j=1}^{p}\frac{E(\gamma_j\beta_j^2)}{2\sigma_{e_j}^2} + b_\beta\right) - \sigma_\beta^{-2}\left(\frac{\sum_{j=1}^{p}E(\gamma_j)}{2} + a_\beta + 1\right),$$

which leads to an analytical update for σ_β^2 as

$$\left(\sigma_\beta^{(t+1)}\right)^2 = \frac{\sum_{j=1}^{p}\frac{E(\gamma_j\beta_j^2)}{2\sigma_{e_j}^2} + b_\beta}{\frac{\sum_{j=1}^{p}E(\gamma_j)}{2} + a_\beta + 1}. \quad \text{(A9)}$$

The EM-MCMC algorithm thus iterates between the E-step and the M-step until converge. The EM-MCMC algorithm allows us to directly obtain the parameter estimate $E(\gamma_j)$, which is the posterior probability of *j*-th gene being a DE gene. This posterior probability is also commonly referred to as the posterior inclusion probability (PIP) in other settings. We use these posterior probabilities to serve as DE evidence. In addition, the EM-MCMC algorithm also provides an estimate for τ_1, which, when paired with its standard error computed in the following section, allows us to construct a Wald test to test the null hypothesis of no gene set enrichment $H_0: \tau_1 = 0$.

A.2 Louis Method for p-value Computation.

Here, we describe the details of the Louis method for computing the standard error of $\hat{\tau}_1$. In the EM-MCMC algorithm described in the previous section, we can obtain the information matrix for (τ_0, τ_1) based on the log complete likelihood $\log Pr\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \tau_0, \tau_1, \sigma_\beta^2\right)$ as described in equation (1). For completeness, we re-write the information matrix in the complete likelihood as a 2 by 2 matrix

$$I_c = \sum_{j=1}^{p} \hat{\pi}_j(1-\hat{\pi}_j)\begin{pmatrix} 1 & a_j \\ a_j & a_j^2 \end{pmatrix}, \quad \text{(A10)}$$

where $\hat{\pi}_j = \frac{\exp(A_j\hat{\boldsymbol{\tau}})}{1+\exp(A_j\hat{\boldsymbol{\tau}})}$ is computed based on the $\hat{\boldsymbol{\tau}}$ estimates from the last EM step and a_j is the annotation for j-th gene. Our goal, however, is to obtain the information matrix for (τ_0, τ_1) based on the marginal log likelihood $\log Pr(\hat{\boldsymbol{\beta}}|\tau_0, \tau_1, \sigma_\beta^2)$, also known as the observed likelihood. Such marginal information matrix can be obtained based on the complete information matrix through an adjustment using the Louis method (Louis 1982, Oakes 1999). Specifically, with the posterior inclusion probably PIP_j for *j*-th gene obtained from EM steps, we compute the information matrix in the incomplete likelihood as a 2 by 2 matrix

$$I_{ic} = \sum_{j=1}^{p} \text{PIP}_j(1-\text{PIP}_j)\begin{pmatrix} 1 & a_j \\ a_j & a_j^2 \end{pmatrix}. \quad \text{(A11)}$$

Finally, the observed information matrix I_o is adjusted by

$$I_o = I_c - I_{ic}$$

Once we compute the marginal information matrix, we can obtain the standard error $\text{se}^2(\hat{\tau}_1)$ as the corresponding element in the inverse of the information matrix I_o.

A.3 Application to an oral carcinoma bulk RNAseq dataset.

To illustrate the flexibility of the modeling framework in iDEA, we applied iDEA to analyze a publicly available bulk RNASeq dataset from Tuch et al (Tuch et al. 2010). The bulk RNAseq dataset consists of gene expression measurements for 10,540 genes on tumors and matched normal tissue from three patients with oral squamous cell carcinomas. We carried out both GSE analyses and DE analyses on comparing the matched tumor and normal pairs.

We first applied iDEA and other GSE methods to detect significantly enriched gene sets

across our compiled database of 12,033 human gene sets. The *p*-values of the enriched gene sets from iDEA are shown in **Figure S2.23A**. We also constructed an empirical null *p*-value distribution by permuting the gene labels for each gene set 10 times. Consistent with both simulations and scRNA-seq data applications, we found that the *p*-values in the permuted data from iDEA ($\lambda_{\text{gc}} = 1.07$), fGSEA ($\lambda_{\text{gc}} = 0.99$), PAGE ($\lambda_{\text{gc}} = 0.98$), and GSEA ($\lambda_{\text{gc}} = 0.99$) are well behaved, while that from CAMERA show severe deflation ($\lambda_{\text{gc}} = 0.12$)(**Figure S2.23B**). For each method, we relied on the empirical null distribution of *p*-values to compute power in detecting enriched gene sets based on a fixed empirical FDR. Consistent with both simulations and scRNA-seq data applications, iDEA displays higher power compared to the other GSE methods (**Figure S2.23C**). For example, at an empirical FDR of 5%, iDEA identified 2075 significantly enriched gene sets, which is 17%, 80%, 20% higher than fGSEA (1777), CAMERA (1154), and GSEA (1733) respectively. While PAGE (2079) also displays higher power in number of detecting the significant gene sets, the top gene sets identified by iDEA are most closely related to oral squamous cell carcinomas or tumor related pathways. For example, among the top 10 gene sets identified by iDEA, 7 are related to tumor pathways. As a comparison, 5 among the top 10 gene sets identified by PAGE are related to tumor pathways. Specifically, enriched gene sets identified by iDEA include the SMID_BREAST_CANCER_NORMAL_LIKE_UP (Shah and Mehta 2009), SWEET_LUNG_CANER_KRAS_DN (Daly et al. 2011) and relevant GO items such as GO:0031012 (extracellular matrix (Pickup, Mouw and Weaver 2014)), GO:0043292 (contractile fiber (Mazzoccoli et al. 2017)). In order to quantify the biological significance of gene sets identified by different GSE methods, we quantified the relevance between gene sets and oral squamous cell carcinomas in an unbiased way by searching the related literatures in PubMed (details in **Methods**). Indeed, in the top 50 enriched gene sets identified by different methods,

iDEA identified more gene sets relevant to oral squamous cell carcinomas (30) than fGSEA (23), CAMERA (30), PAGE (23), and GSEA (25). The higher number of detected enriched gene sets relevant to oral squamous cell carcinomas and cancer growth by iDEA provides convergent support for the higher power of iDEA for GSE analysis.

Next, we applied iDEA for DE analysis to identify DE genes. Consistent with both simulations and scRNA-seq data applications, iDEA identified more DE genes than zingeR. For example, at an empirical FDR of 1%, iDEA identified 409 DE genes, while DESeq2 only identified 1 (**Figure S2.23D**). The 50 selected important DE genes identified by iDEA clearly distinguishes the normal tissue and cancer tissue (**Figure S2.23F**). Importantly, using the key markers provided by the original study (Usoskin et al. 2015), iDEA identified 262 genes directly related to oral squamous cell carcinomas or important genes involved in common tumors; while zingeR only identified 1. The higher number of DE genes relevant to oral squamous cell carcinomas or common tumors detected by iDEA provides convergent support for its higher power for DE analysis. Important DE genes involved in Oral squamous cell carcinoma development that are detected by iDEA but missed by zingeR include *CRNN* (Salahshourifar et al. 2015), *WNT10A* (Uraguchi et al. 2004), *PTHLH* (Lv et al. 2014), *KRT6* (Harris et al. 2015), *IGF1* (Zhi et al. 2014), *PTGFR* (Akiyama et al. 2013), *TGFBR3* (Cheng et al. 2016). Among them, *CRNN* has been studied to be the potential prognostic marker of OSCC due to its downregulation in oral squamous cell carcinoma samples, *WNT10A* plays an important role in accelerating of the progression of carcinomas via activating EMTs and local invasiveness (Uraguchi et al. 2004), *PTHLH* is indispensable for the pathogenesis of oral squamous cell carcinoma by affecting cell proliferation and cell cycle (Lv et al. 2014). *TGFBR3* is an important activator of *GDF10*, which is

downregulated during oral carcinogenesis and involved in the suppression of cell survival (Cheng et al. 2016).

Further, we also evaluated the GC content and gene length effect. For all the genes in the dataset, we first calculated the GC content and gene length and then we create two gene sets corresponding to the levels of GC content and gene length. Specifically, for the GC content, we use the continuous value of GC content for each gene as the gene set. For the gene length, we created a binary gene set if gene length is higher than the average of the gene length, than this gene is in this gene set and annotated as 1 otherwise 0. Finally, by adding these two gene sets into iDEA, we calculated the *p*-values for these two gene set to represent the significance of GC content effect and gene length effect correspondingly. In the analyses, we did not observe obvious GC content effect (*p*-value = 0.31) and gene length effect (*p*-value = 0.08) in this dataset.

A.4 Bayesian model averaging (BMA) approach.

Besides performing DE analysis in iDEA in the real data based on a pre-selected gene set, we also developed a new strategy to aggregate DE evidence on a particular gene across all gene sets through Bayesian model averaging (BMA). Specifically, for the given gene, we denote its posterior inclusion probability (PIP) obtained using the gene set k as PIP_k. The corresponding Bayes factor quantifying its DE evidence based on the gene set k is $\mathrm{BF}_k = \mathrm{PIP}_k/(1-\mathrm{PIP}_k)$. With equal prior weights on different gene sets, the average Bayes factor quantifying its DE evidence based on all K gene sets is thus $\mathrm{ABF} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{BF}_k$, which can be converted back to a posterior inclusion probability as $\mathrm{PIP} = \mathrm{ABF}/(1+\mathrm{ABF})$. We found that PIPs computed this way is highly correlated with the PIPs computed based on the pre-selected gene set (**Figure S2.24**). We now provide both options for computing PIPs for quantifying DE evidence: biologists can choose to use pre-selected gene sets that are known to be relevant to the particular experiments, as is the case

for all the real data applications; alternatively, biologists also have the option of using the Bayesian model averaging when such prior knowledge is not available.

A.5 Gene set overlap.

Previously we followed most existing GSE approaches and accounted for test non-independence due to gene set overlap through permutation. In addition, we also performed new analysis to further examine the issue of gene set overlap in the real data applications. We adopted the method proposed by Jiang and Gentleman (Jiang and Gentleman 2007) to examines pairs of gene sets one at a time. For each pair of gene set, Jiang's method divides genes into three categories: one category of genes that are only in the first gene set, one category of genes that are only in the second gene set, and one category of genes that are common in both gene sets. Afterwards, Jiang's method calculates three *p*-values, one for each category of genes. By computing *p*-values in each set, we can explicitly deconvolute the results in the presence of gene set overlap. Here, we mainly applied Jiang's method to analyze the top 50 gene sets identified by iDEA in human embryonic data and mouse neuron cell data in order to further dissect particular set of genes that drive the enrichment signal (Note that we did not apply to all significant gene sets due to the heavy computational burden of Jiang's method and the gene set overlap is moderate compare to the gene set size). Specifically, there are 1,225 pairwise combinations among top 50 gene sets. For each real data we checked, we first construct the pairwise combinations of gene sets among top 50 significant gene sets identified by iDEA and for each pair, and then we filtered out gene set pairs which has less than 20 genes overlap (due to computational stability). For each pair which has larger than 20 genes in overlap, we calculated above mentioned three categories of *p*-values. Then we checked the *p*-values of the category of genes that are common in both gene sets and the *p*-values of the category of genes that are unique in gene sets respectively. For example,

in the human embryonic data, 692 of the 1,225 gene set pairs have higher than 20 genes in overlap. For each of these 692 gene set pairs in turn, we calculated the three *p*-values as mentioned in the previous paragraph. Among the total 2,076 adjusted *p*-values (Bonferroni correction) we calculated, 1,397 of them are less than 0.05. We first look at the intersection part, 35 out of 692 intersection sets have adjusted *p*-value is less than 0.05. For the disjoint parts, 1,362 out of 1,384 are significant. This observation suggests that among the top 50 significant gene sets we identified, gene set specific genes are significantly enriched, suggesting that it is not the overlapped genes that drive the enrichment signal, and that gene set overlap does not appear to introduce excessive false signals. We further looked at the combination of the top first gene set GO:0001944 (vasculature development) (**Table 2.10**). From the table, we observed that the significance of this gene set is induced by both the overlapping parts and non-overlapping parts. Following the same procedure, we also applied Jiang's method to analyze the top 50 gene sets identified by iDEA in the mouse sensory neuron scRNA-seq data. 1,025 out of 1,225 gene set pairs have higher than 20 genes in overlap. For each of these 1,025 gene set pairs in turn, we calculated the three *p*-values as mentioned in the previous paragraph. Among the total 3,075 adjusted *p*-values (Bonferroni correction) we calculated, 2,603 of them are less than 0.05. We first look at the intersection part, 889 out of 1,025 intersection sets have adjusted *p*-value is less than 0.05. For the disjoint parts, 1,714 out of 2,050 are significant. We further looked at the combination of the top first gene set GO:0044425 (obsolete membrane part) (**Table 2.11**). From the table, we observed that the significance of this gene set is induced by both the overlapping parts and non-overlapping parts.

A.6 Cell type identification in the three scRNA-seq datasets.

For all the real datasets we analyzed, one of our real data contains cell types that are known *a priori* and not inferred from the whole expression matrix, while the other two data contain cell

types that are extensively validated through approaches other than inferring based on the whole expression matrix. Specifically, for the human embryonic stem cell scRNAseq dataset, the cell types are obtained from fluorescence-activated cell sorting (FACS) analysis before mixing for scRNA-seq. FACS relies on known cell type markers and represents a somewhat unbiased strategy for cell type clustering (Baron et al. 2019). For the mouse neuronal scRNAseq dataset, the cell types are initially inferred through an iterative PCA-based procedure and are further validated by comparing the hierarchical relationship of the neuronal types with the known developmental origin of sensory neuron types, as well as by comparing neurons with distinct and characteristic soma sizes in their identified neuronal class. In addition, the inferred neuronal cell types are further confirmed by double and triple immunohistochemical staining (e.g., NP1 cell type by staining of *PLXNC1*). For the 10x Genomics PBMC scRNASeq dataset, the identity of cell types was inferred by aligning cluster-specific genes to known markers of distinct PBMC populations as well as comparing against the transcriptomes of the purified populations in PBMC subsets. Their approach has been found to be largely consistent with conventional marker-based methods and the major cell types reach to the expected ratios in PBMCs. We have also displayed t-SNE plot in **Figure S2.16**, which clearly shows distinct cell clusters. Because the cell types in these data are validated through various approaches, the DE analysis results are less likely influenced by the cell type inference step as compared to other data that are fully relying on the whole gene expression matrix for cell type inference.

Appendix B. Chapter 3 (CARD) Supplementary Text

B.1 Description on compared deconvolution methods

We compared CARD with six recently developed deconvolution methods such as RCTD (Cable et al. 2021), stereoscope (Andersson et al. 2020), SPOTlight (Elosua-Bayes et al. 2021), cell2location (Kleshchevnikov et al. 2022), and spatialDWLS (Dong and Yuan 2021). Specifically, RCTD directly models expression count data from spatial transcriptomics based on a Poisson factor analysis model, which extends the linear factor analysis models commonly used for bulk RNA-seq deconvolution. In the factor analysis model, RCTD introduces additional variance parameters to account for the different platform effects between scRNA-seq and spatial transcriptomics. Stereoscope uses a similar approach as RCTD but with a negative binomial model for modeling the observed count data. Because of direct count modeling, RCTD and stereoscope are particularly suited for high resolution spatial transcriptomics that measures a couple of cells on each tissue location with relatively low sequencing depth per location. SPOTlight takes low-dimensional components from both scRNA-seq and spatial transcriptomics as input. The low-dimensional components of scRNA-seq are referred to as cell-type specific topic profiles in SPOTlight and are obtained through non-negative matrix factorization and aggregated across cells within each cell type. With the input of the low-dimensional components, SPOTlight relies on a nonnegative least squares estimation procedure commonly used in bulk RNA-seq deconvolution for spatial transcriptomics deconvolution21,22. cell2location models the expression count data with a negative binomial model and accounts for cell type composition variation across distinct tissue segments through incorporating the tissue segmentation information as a latent factor.

spatialDWLS extends the bulk RNA-seq deconvolution method DWLS to first identify the cell types existed on each spatial location through enrichment analysis and further estimate the cell type proportions using the inferred signature mean gene expression in scRNA-seq though a weighted least squares framework.

B.2 Simulation Design

B.2.1 Model-free simulation design

We performed realistic simulations to evaluate the performance of CARD and compare it with other deconvolution approaches. To do so, we obtained two published datasets: a scRNA-seq data collected on the mouse nervous system (Zeisel et al. 2018)and a spatial transcriptomics data collected on the mouse olfactory bulb (Ståhl et al. 2016). We used the scRNA-seq data to construct the expression levels for 18,215 genes on 260 spatial locations that were measured in the spatial transcriptomics data. Specifically, in the scRNA-seq data, we obtained expression measurements for a total of 20,515 cells from six common cell types. These cell types include neurons (n=11,702), astrocytes (n=5,039), oligodendrocytes (n=1397), vascular cells (n=1162), immmune cells (1078), and ependymal cells (n = 15). Following (Andersson et al. 2020), we split the scRNA-seq data into two sets: one set (50% cells, denoted as split1) was used to simulate the spatial transcriptomics count while the set (50% cells, denoted as split2) was used to evaluate the performance of deconvolution methods. In the spatial transcriptomics data, we obtained location information for 260 spatial locations and followed (Svensson et al. 2018) to categorize these 260 locations into three main anatomic regions. The three anatomic regions include the granule cell layer (75 locations), the mitral cell layer (140 locations), and the nerve layer (45 locations). In the simulations, we assumed that each anatomic region contains a dominant cell type. In particular, oligodendrocyte is the dominant cell type in the granule cell layer; astrocyte is the dominant cell

type in the nerve cell layer; and neuron is the dominant cell type in the mitral cell layer. For each anatomic region, we determined the number of cell types colocalized with the dominant cell type based on a uniform distribution U (0, 5). We randomly draw the proportions for each cell type from a Dirichlet distribution with the concentration parameter set to be 1.0 for all cell types in the region. We assigned the largest proportion to be the proportion of the dominant cell type, and randomly assigned the remaining proportions for the other cell types. In order to mimic the number of cells at each location observed by nuclear segmentation of the mouse brain histology images, we followed (Andersson et al. 2020) and fixed the total number of cells on each location to be 10. We then set the number of cells for each cell type on each location as 10 times the sampled cell type proportions (further rounded to the nearest integer). Afterwards, we randomly sampled the corresponding number of cells from the six cell types in the split1 scRNA-seq data without replacement to serve as the cells residing on the spatial location. We then summed the expression levels for each gene across the sampled cells as the expression level of the corresponding gene on the location.

In the above procedure, we also added additional noise by setting a percentage of the spatial locations in each anatomic region to be noisy locations. We denote the percentage of noisy locations as p_n. On the noisy locations, rather than setting the dominant cell type for each layer, we randomly draw the cell type proportions from a Dirichlet distribution with the concentration parameter set to be 1.0 for all cell types without assigning a dominant cell type. We followed the same procedure described above to sample the spatial count data. Note that our simulation strategy does not match the CARD model and thus allows us to examine the robustness of CARD. In the simulations, we varied the percentage of noisy locations, p_n, to be either 0, 0.2, 0.4, or 0.6. These choices of p_n cover a wide range of measurement noise that can be encountered in spatial

transcriptomics. We focused on the p_n up to 0.6 because the spatial correlation pattern is almost completely gone at 0.6 (**Figure S3.1**), while the cell type composition in the simulated spatial data using $p_n = 0.8$ or 1.0 is almost indistinguishable from the simulated data using P_n= 0.6 (Ma and Zhou 2022). Consequently, the deconvolution accuracy of different methods is also highly consistent when P_n ranges from 0.6, 0.8 to 1.0 (Ma and Zhou 2022). Therefore, we examined a total of 4 simulation settings, each of which consisted of 5 simulation replicates. Note that we used multiple simulation replicates, instead of using one replicate as in previous deconvolution studies (Cable et al. 2021, Elosua-Bayes et al. 2021), in order to capture data variation and examine method robustness. Because some simulation replicates are easier to perform cell type deconvolution on while others are harder, the absolute value of RMSE for any method can vary substantially across replicates, even though the performance rank of different methods remains consistent across replicates.

B.2.2 Simulation Analysis Scenarios

We applied CARD along with other deconvolution methods (details in **Methods**) to analyze the simulated data. In the analysis, we examined five analysis scenarios:

(1) **Analysis scenario I**: (the correct scRNA-seq reference): we applied the scRNA-seq reference data (split2) that contains all cell types to deconvolute the simulated spatial transcriptomics data.

(2) **Analysis scenario II**: (missing one cell type in the reference): we applied the scRNA-seq reference data (split2) to deconvolute the simulated spatial transcriptomics data. Different from scenario I, however, we removed one cell type in the reference during deconvolution to examine the robustness of deconvolution.

(3) **Analysis scenario III**: (one extra cell type in the reference): we applied the scRNA-seq reference data (split2) to deconvolute the simulated spatial transcriptomics data. Different from scenario I, however, we added a new cell type in the reference during deconvolution. Specifically, we examined adding one new cell type, the "blood cells", which contains n=70 cells.

(4) **Analysis scenario IV**: (miss classified cell type in the reference): we applied the scRNA-seq reference data (split2) to deconvolute the simulated spatial transcriptomics data. Different from scenario I, however, we randomly merged two cell types into one cell type as the merge one in the reference during deconvolution. Because we have six cell types in the original simulations, we created 15 cell type misclassification settings, each consisting of five cell types: one merged cell type based on two out of the six cell types, along with the four remaining cell types.

(5) **Analysis scenario V**: (a similar scRNA-seq reference from a different platform): we applied a scRNA-seq reference data from a different platform for deconvolution. Specifically, we obtained another scRNA-seq reference data (Mizrak et al. 2019) that was sequenced on a different platform microwell-seq + Drop-seq on the mouse brain (ventricular-subventricular zone). The new scRNA-seq data contains a similar set of cell types with similar expression patterns as the cell types in the old scRNA-seq data (Ma and Zhou 2022). For the new scRNA-seq data, we extracted six matched cell types from GSE109447 dataset for deconvolution: astrocytes (n=13765), neurons (n=3110), oligodendrocytes (n=7513), endothelial cells (n=1774), immune cells (specifically microglia cells, n = 5525), and ependymal cells (n = 997).

We also applied CARD and the other deconvolution methods in the simulation scenarios where the scRNA-seq reference is provided at different cell type resolutions. To do so, we first performed hierarchical clustering in the scRNA-seq data on each of the original six cell types that were used to simulate the spatial transcriptomics data. The hierarchical clustering separated each of the six cell types (except ependymal cells due to its low sample size, n = 7 in split1) into either 2, 4, 6, or 8 sub cell types, resulting in a total of 10 to 40 sub cell types in the reference scRNA-seq data. We performed simulations using the original six major cell types but used the scRNA-seq data with different number of sub cell types to serve as the reference for deconvolution, thus creating the scenario of deconvolution with different/higher cell type resolutions (Ma and Zhou 2022).

B.3 Evaluations on Simulations

In all simulation's scenarios, we evaluated the deconvolution accuracy by comparing the RMSE between the truth and estimated proportions. Specifically, for the scenarios I-III, we calculate the RMSE between $\widehat{\boldsymbol{V}}$ and $\boldsymbol{V}$ using the above equation based on the cell types existing in the truth. For the scenario IV, we calculate the RMSE between $\widehat{\boldsymbol{V}}$ and $\boldsymbol{V}$, where the true proportion of the merged cell type on each spatial location is the summation of the proportions for the two underlying cell types that were merged. For the scenario V, we calculate the RMSE between $\widehat{\boldsymbol{V}}$ and $\boldsymbol{V}$ based on the matched cell types. For each scenario in scenarios II, III, IV, and V, we also compared the results of CARD with the deconvolution of CARD when using the six cell types in the original scRNAseq data to evaluate the accuracy loss of CARD. Specifically, for both the simulation scenario II and III, the RMSE of the deconvolution of CARD when using the six cell types in the original scRNAseq data is calculated based on the cell types existing in the truth. For the scenario IV, the RMSE of the deconvolution of CARD when using the six cell types in the original scRNAseq data is that we used the original six cell types for deconvolution and obtained

the estimated proportion for the merged cell type as the summation of the estimated proportions for the two underling cell types, and the true proportion of the merged cell type on each spatial location is also the summation of the proportions for the two underlying cell types that were merged. We then compared the estimated cell type proportion with the underlying truth. For scenario V, the RMSE of the deconvolution of CARD when using the six cell types in the original scRNAseq data is calculated based on matched cell types (equal to the RMSE in scenario I). Then for the scenarios II-V, we calculated the percentage loss of the deconvolution accuracy of CARD by $\frac{RMSE(scenario\ S) - RMSE(using\ the\ six\ cell\ types\ in\ original\ scRNAseq\ )}{RMSE(oracle\ resultsa\ in\ scenario\ S)} * 100$, with the RMSE ($scenario\ S$) represents the RMSE of CARD in a specific scenario S (S = II, III, IV or V) and the RMSE ($using\ the\ six\ cell\ types\ in\ original\ scRNAseq$) is calculated based on the results of CARD when using the six cell types in the original scRNA-seq data. Thus, in this way, we can evaluate the percentage of accuracy loss due to the missing (scenario II), additional (scenario III), misclassified cell types (scenario IV) in the scRNAseq reference or the use of scRNAseq reference from a different platform (scenario V).

For the calculation of the percentage of accuracy improvement over other methods, for example, when we compare CARD with a specific method (e.g., RCTD) in a specific setting, we calculate the accuracy improvement as by - $\frac{RMSE(CARD) - RMSE(RCTD)}{RMSE(RCTD)} * 100$. For the evaluation on the deconvolution performance on scRNA-seq references at different resolution, we treated the cell type composition for the six major cell types in the simulated spatial transcriptomics data as the underlying truth. After deconvolution with sub cell types, we summed the estimated proportions of the sub cell types for each major cell type to serve as the estimated cell type proportion. We then compared the estimated cell type proportion with the underlying truth for these six main cell types.

B.4 Details on Preprocessing Spatial Transcriptomics and scRNA-seq Datasets

We performed four sets of deconvolution analyses on five published spatial transcriptomics datasets (**Table S3.2**). For each spatial transcriptomics data, we applied one or more scRNA-seq data to serve as the reference for deconvolution.

B.4.1 Mouse Olfactory Spatial Transcriptomics Data

We downloaded the mouse olfactory bulb spatial transcriptomics data (Ståhl et al. 2016) from the spatial transcriptomics research website (https://www.spatialresearch.org/). This data consists of gene expression measurements in the form of read counts that are collected on several spatial locations known as spots. We followed (Ståhl et al. 2016, Edsgärd, Johnsson and Sandberg 2018) to focus on the MOB section #12, which contains 16,034 genes and 282 spatial locations. For deconvolution, we obtained the Tepe et al (Tepe et al. 2018) scRNA-seq data from Gene Expression Omnibus (GEO; accession number GSE121891) to serve as the reference. This scRNA-seq data was collected from the mouse olfactory bulb and contains 18,560 genes and 12,801 cells. The cells have already been clustered into the following main cell types: granule cells (GC, n = 8,614), olfactory sensory neurons (OSNs, n = 1,200), periglomerular cells (PGC, n = 1,693), mitral and tufted cells (M-TC, n = 1,133), and external plexiform layer interneurons (EPL-IN, n = 161). In the data, we filtered out genes that have zero counts on all cells and filtered out cells that have zero counts on all genes. These filtering criteria led to a final set of 17,812 genes and 12,801 cells for analysis.

B.4.2 Human Pancreatic Ductal Adenocarcinoma (PDAC) Data

We downloaded the human pancreatic ductal adenocarcinoma (PDAC) data from GEO website (accession number GSE111672) (Moncada et al. 2020). This dataset consists of both spatial transcriptomics data and scRNA-seq data collected on the same tissue obtained using the

inDrop technology. Following the original paper, we focused on the PDAC-A spatial transcriptomics data for the patient ID GSM3036911. In the analysis, we filtered out genes that have zero counts on all spatial locations and filtered out locations that have less than 100 total read counts. These filtering criteria led to a final set of 22,269 genes and 428 locations.

Two scRNA-seq data are available in the same study: PDAC-A-inDrop, which is a matched data collected from the same patient GSM3036911; and PDAC-B-inDrop, which is an unmatched data collected from a different patient. PDAC-A-inDrop consists of 19,736 genes and 1,926 cells that were clustered into 20 cell types in the original study. PDAC-B-inDrop consists of 19,736 genes and 1,733 cells that were clustered into 13 cell types in the original study. For both scRNA-seq data, we filtered out genes that have zero counts on all cells and filtered out cells that have zero counts on all genes. These filtering criteria led to a final set of 16,381 genes and 1,926 cells for PDAC-B-inDrop and a final set of 15,919 genes and 1,733 cells for PDAC-B-inDrop. We performed deconvolution using either scRNA-seq data to serve as the reference.

In addition to using the scRNA-seq data from the same study, we also obtained one external scRNA-seq data (10x Chromium) from Peng et al (Junya et al. 2019) (Genome Sequence Archive under project PRJCA001063). This data was collected from 11 control pancreas and 24 PDAC tumors and was sequenced using the 10x Chromium platform. The data contains 24,005 genes and 57,530 cells from 10 cell types. We split the data into three subsets: the PengNormal data that consists of the 11 control samples; the PengTumor data that consists of the 24 PDAC tumors; and the Peng data that consists of all 35 samples. We used the same filtering criteria for quality control in each subset, leading to a final set of 23,886 genes and 57,530 cells for Peng, 21,151 genes and 15,544 cells for Peng_normal, 23,527 genes and 41986 cells for Peng_tumor. We treated each of the three subsets as the reference to examine the robustness of deconvolution.

B.4.3 Mouse Hippocampus Slide-seqV2 Data and Mouse Brain 10x Visium Data

We obtained the mouse hippocampus Slide-seqV2 dataset (Stickels et al. 2021) from the Broad Institute's Single Cell Portal. This dataset consists of gene expression measurements in the form of read counts for 23,265 genes and 53,208 spatial locations. In the analysis, we filtered out genes that have zero counts on all locations and filtered out locations that have less than 100 total read counts. We focused on the remaining set of 23,238 genes and 41,768 spatial locations for analysis.

We obtained the mouse brain (coronal section) 10x Visium dataset from the 10x genomics website (https://www.10xgenomics.com/resources/datasets/). This dataset consists of gene expression measurements in the form of read counts on 21,143 genes and 2,698 spatial locations. We filtered out genes that have zero counts on all spatial locations and filtered out spatial locations that have less than 100 total read counts. We focused on the remaining set of 20,984 genes and 2,698 spatial locations for analysis.

For deconvoluting the above two spatial transcriptomics datasets, we obtained the DropViZ scRNA-seq dataset from the Broad Institute's Single Cell Portal to serve as the reference. This dataset was collected from the mouse hippocampus using the Drop-seq technology and was used for cell type deconvolution in the RCTD paper. The dataset contains 27,953 genes and 15,095 cells from 17 cell types. We filtered out genes that have zero counts on all cells and filtered out cells that have zero counts on all genes. These filtering criteria led to a final set of 23,282 genes and 15,095 cells.

Finally, we created a low-resolution version of the Slide-seqV2 data to examine the performance of refined spatial map construction. Specifically, we created a binned Slide-seqV2 data (10 μm) that match the feature size of 10x Visium data (55 μm) by following the github code

provided by the original publication (Stickels et al. 2021) (https://github.com/rstickels/Slide_seqv2/blob/master/SlideseqV2_visium_comparison_submission.ipynb). To do so, we first divided the area of the original Slide-seqV2 hippocampus data into equal-size bins based on the parameters provided in the github code. Afterwards, we aggregated the spatial locations that are in the same bin into new locations as our final binned spatial data, leading to a resolution that matches that of the 10x Visium dataset.

B.5 Methodological Details of CARD

B.5.1 The Gaussian kernel function

Following the previous literatures (Svensson et al. 2018, Sun et al. 2020a), we used the Gaussian kernel function (a.k.a. squared exponential kernel function or radial basis kernel function) as our spatial kernel in the CAR model. The Gaussian kernel is in the form of

$$K_G(\boldsymbol{s}_i, \boldsymbol{s}_j) = exp(-\frac{\left|\left|\boldsymbol{s}_i - \boldsymbol{s}_j\right|\right|^2}{2\sigma^2}) \quad \text{(B1)}$$

where $\left|\left|\boldsymbol{s}_i - \boldsymbol{s}_j\right|\right|^2$ is the Euclidean distance; and σ the length/scale parameter that effectively characterizes the size of the focal expression patterns. The Gaussian kernel function is infinitely differentiable. Because any reasonable values of σ gives out almost identical results, we simply fixed σ to be 0.1 throughout the study.

B.5.2 Reference basis matrix and its relationship to spatial transcriptomics

Here, we provide details on how we construct the reference basis matrix from the scRNA-seq data, how the reference basis matrix is related to the spatial transcriptomics data, and how the equation (1) in the main text is derived. Our approach follows closely the previous bulk RNA-seq

deconvolution work in (Wang et al. 2019, Dong et al. 2020). To facilitate description, we first introduce the necessary notations as follows.

Notations:

g: gene index

i: location index

k: cell type index

c: cell index

C_{ik}: set of cells on location *i* that belong to cell type *k*

C_i: set of all cells on location *i*

n_{ik}: number of cells on location *i* that belong to cell type *k*

n_i: number of total cells on location *i*

x_{ig}: read counts for gene *g* on location *i*

x_i: total number of read counts on location *i*; $x_i = \sum_g x_{ig}$

x_{igc}: read counts for gene *g* on location *i* contributed from cell *c*

With the above notations, we can obtain the total read counts for gene *g* on location *i* (x_{ig}) as

$$x_{ig} = \sum_{c \in C_i} x_{igc}$$

$$= \sum_k \sum_{c \in C_{ik}} x_{igc}$$

$$= \sum_k n_{ik} \frac{\sum_{c \in C_{ik}} x_{igc}}{n_{ik}}$$

$$= \sum_k n_{ik} \boldsymbol{B}_{igk}. \quad \text{(B2)}$$

We denote $\boldsymbol{B}_{igk} = \frac{\sum_{c \in C_{ik}} x_{igc}}{n_{ik}}$ as the mean expression counts for gene g on location i that are contributed by cell type k. We denote y_{ig} as the relative abundance of gene g on location i in the spatial transcriptomics data, defined as the ratio of reads mapped to gene g on location i out of the total read depth on the location. Thus, we have

$$y_{ig} = \frac{x_{ig}}{\sum_g x_{ig}}$$

$$= \frac{\sum_k n_{ik} \boldsymbol{B}_{igk}}{\sum_g x_{ig}}$$

$$= \frac{\sum_k n_{ik} \boldsymbol{B}_{igk}}{\sum_g x_{ig} n_i} * n_i. \quad \text{(B3)}$$

We denote $\boldsymbol{P}_{ik} = \frac{n_{ik}}{n_i}$ as the proportion of read counts on location i that are contributed by cell type k. Consequently, the relative abundance y_{ig} can be expressed as

$$y_{ig} = \sum_k \boldsymbol{B}_{igk} \boldsymbol{P}_{ik} * \frac{n_i}{\sum_g x_{ig}}$$

$$= \sum_k \boldsymbol{B}_{igk} \boldsymbol{P}_{ik} * \frac{n_i}{x_i}$$

$$= \sum_k \boldsymbol{B}_{igk} \boldsymbol{P}_{ik} * l_i^{-1}$$

$$= \sum_k \boldsymbol{B}_{igk} \boldsymbol{V}_{ik}. \quad \text{(B4)}$$

Above, $l_i = \frac{x_i}{n_i}$ represents the average read counts per cell on location i. We denote $\boldsymbol{V}$ as the cell type composition matrix, with each element $\boldsymbol{V}_{ik} = \boldsymbol{P}_{ik} \boldsymbol{l}_i^{-1}$. For the reference basis matrix $\boldsymbol{B}$, we followed (Dong et al. 2020, Wang et al. 2019) and set

$$\boldsymbol{B}_{igk} = \frac{\sum_{c \in C_{ik}} x_{igc}}{n_{ik}}$$

$$= \frac{\sum_{c \in C_{ik}} x_{igc}}{\sum_{c \in C_{ik}} \sum_{g'=1}^{G} x_{ig'c}} \frac{\sum_{c \in C_{ik}} \sum_{g'=1}^{G} x_{ig'c}}{n_{ik}}$$

$$= \theta_{igk} S_{ik}, \quad \text{(B5)}$$

where $\theta_{igk} = \frac{\sum_{c \in C_{ik}} x_{igc}}{\sum_{c \in C_{ik}} \sum_{g'=1}^{G} x_{ig'c}}$ is the relative abundance of gene *g* on location *i* for cell type *k*; and $S_{ik} = \frac{\sum_{c \in C_{ik}} \sum_{g'=1}^{G} x_{ig'c}}{n_{ik}}$ represents the average number of total read counts for cells of cell type *k* on location *i*. We assume that across all spatial locations, the relative abundance of gene *g* on location *i* for cell type *k* has the same mean θ_{gk} as the scRNA-seq data. We also assume that the average number of total counts for cells of cell type *k* on location *i* has the same mean as the scRNA-seq data. Under these assumptions, we can use the available scRNA-seq dataset, which contains either one or multiple samples, to estimate $B_{gk} = \theta_{gk} S_k$. Here, θ_{gk} is estimated from the scRNA-seq data as the mean relative abundance of gene *g* in cell type *k* and S_k is estimated from the scRNA-seq data as the mean number of total read counts for cells of cell type *k*. Specifically, with *J* samples in the scRNA-seq reference data and with $\tilde{x}_{jgc}$ denoting the observe read count for gene *g* in cell *c* in sample *j*, we have $\theta_{gk} = \frac{\sum_{j=1}^{J} \frac{\sum_{c \in C_{jk}} \tilde{x}_{jgc}}{\sum_{c \in C_{jk}} \sum_{g'=1}^{G} \tilde{x}_{jg'c}}}{J}$, $S_k = \frac{\sum_{j=1}^{J} \frac{\sum_{c \in C_{jk}} \sum_{g'=1}^{G} \tilde{x}_{jg'c}}{n_{jk}}}{J}$. With these estimates, we finalize the equation in (4) as:

$$y_{ig} = \sum_k \boldsymbol{B}_{gk} \boldsymbol{V}_{ik} + \epsilon_{gi}.$$

In matrix format, this is our factor model introduced in the first equation in the main text:

$$\boldsymbol{Y} = \boldsymbol{BV} + \boldsymbol{E}. \tag{B6}$$

B.5.3 Statistical inference for CARD

The CARD model is defined by equations (1) (2) in the main text, with details described in the **Methods**. We follow Brook's Lemma to get the joint distribution for the N-size column vector $\boldsymbol{V}_k$ as in equation (3) in the main text. Specifically, the Brook's Lemma states that:

Define the sample space Ω to be the set of all possible realizations $\mathbf{x} = (x_1, x_2, \ldots x_N)$ that $\Omega = \{\mathbf{x}: \boldsymbol{P}(\mathbf{x}) > 0\}$. Then for any two given realizations $\mathbf{x}$ and $\mathbf{y} \in \Omega$,

$$\frac{P(\mathbf{x})}{P(\mathbf{y})} = \prod_{i=1}^{N} \frac{\boldsymbol{P}(x_i|x_1, \ldots, x_{i-1}, y_{i+1}, \ldots y_N)}{\boldsymbol{P}(y_i|x_1, \ldots, x_{i-1}, y_{i+1}, \ldots y_N)}$$

Let $\mathbf{x} = \mathbf{V}_\mathrm{k} = (\mathbf{V}_{1\mathrm{k}}, \mathbf{V}_{2\mathrm{k}}, \ldots \mathbf{V}_{\mathrm{Nk}})$ and $\mathbf{y} = (\boldsymbol{b}_k, \ldots, \boldsymbol{b}_k)$ with each element equals to $\boldsymbol{b}_k$, then we have:

$$\boldsymbol{P}(\mathbf{V}_\mathrm{k}) = \prod_{i=1}^{N} \frac{\exp\left(-\frac{1}{2\sigma_{ik}^2}\left(\boldsymbol{V}_{ik} - \boldsymbol{b}_k - \phi\sum_{j<i}\boldsymbol{W}_{ij}\left(\boldsymbol{V}_{jk} - \boldsymbol{b}_k\right) - \phi\sum_{j>i}\boldsymbol{W}_{ij}(\boldsymbol{b}_k - \boldsymbol{b}_k)\right)^2\right)}{\exp\left(-\frac{1}{2\sigma_{ik}^2}\left(\boldsymbol{b}_k - \boldsymbol{b}_k - \phi\sum_{j<i}\boldsymbol{W}_{ij}\left(\boldsymbol{V}_{jk} - \boldsymbol{b}_k\right) - \phi\sum_{j>i}\boldsymbol{W}_{ij}(\boldsymbol{b}_k - \boldsymbol{b}_k)\right)^2\right)}\boldsymbol{P}(\mathrm{y})$$

$$\propto \prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma_{ik}^2}(\boldsymbol{V}_{ik} - \boldsymbol{b}_k)^2 + \frac{1}{\sigma_{ik}^2}(\boldsymbol{V}_{ik} - \boldsymbol{b}_k)\phi\sum_{j<i}\boldsymbol{W}_{ij}\left(\boldsymbol{V}_{jk} - \boldsymbol{b}_k\right)\right)$$

$$\propto \exp\left(-\sum_{i=1}^{N}\frac{1}{2\sigma_{ik}^2}(\boldsymbol{V}_{ik} - \boldsymbol{b}_k)^2 + \phi\sum_{i=1}^{N}\sum_{j<i}\frac{1}{2\sigma_{ik}^2}2(\boldsymbol{V}_{ik} - \boldsymbol{b}_k)\boldsymbol{W}_{ij}\left(\boldsymbol{V}_{jk} - \boldsymbol{b}_k\right)\right)$$

If $\frac{\boldsymbol{W}_{ij}}{\sigma_{ik}^2} = \frac{\boldsymbol{W}_{ji}}{\sigma_{jk}^2}$ and $\boldsymbol{W}_{ii} = 0$, then we have

$$\boldsymbol{P}(\mathbf{V_k}) \propto \exp\left(-\sum_{i=1}^{N}\frac{\mathbf{1}}{2\sigma_{ik}^2}(\boldsymbol{V}_{ik} - \boldsymbol{b}_k)^2 + \phi\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mathbf{1}}{2\sigma_{ik}^2}(\boldsymbol{V}_{ik} - \boldsymbol{b}_k)\boldsymbol{W}_{ij}\left(\boldsymbol{V}_{jk} - \boldsymbol{b}_k\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{V}_k - \boldsymbol{b}_k\mathbf{1}_N)^T\boldsymbol{M}_k^{-1}(\boldsymbol{V}_k - \boldsymbol{b}_k\mathbf{1}_N) + \frac{1}{2}(\boldsymbol{V}_k - \boldsymbol{b}_k\mathbf{1}_N)^T M_k^{-1}\phi\boldsymbol{W}(\boldsymbol{V}_k - \boldsymbol{b}_k\mathbf{1}_N)\right.$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{V}_k - \boldsymbol{b}_k\mathbf{1}_N)^T(\boldsymbol{M}_k^{-1}(\boldsymbol{I}_N - \phi\boldsymbol{W}))\,(\boldsymbol{V}_k - \boldsymbol{b}_k\mathbf{1}_N)\right), \quad \text{(B7)}$$

where $\mathbf{1}_N$ is a *N*-vector of 1's and $\boldsymbol{M}_k = diag\left(\sigma_{1\mathrm{k}}^2, \ldots \sigma_{Nk}^2\right)$

We have defined the row-standardized weight matrix $\widetilde{\boldsymbol{W}}_{ij} = \boldsymbol{W}_{ij}/\boldsymbol{W}_{i+}$ that satisfy the symmetric condition $\widetilde{\boldsymbol{W}}_{ij}\sigma_{jk}^2 = \widetilde{\boldsymbol{W}}_{ji}\sigma_{ik}^2$ and $\sigma_{ik}^2 = \boldsymbol{\lambda}_{\mathrm{k}}\,/\,\boldsymbol{W}_{i+}$ in the main text. So, we replace the $\boldsymbol{W}$ matrix with the $\widetilde{\boldsymbol{W}}$ and we have $\boldsymbol{P}(\mathbf{y}) = \mathbf{1}$. We finally obtain the joint distribution of $\boldsymbol{V}_k$, as what we described in the equation (3) in the main text.

$$\boldsymbol{V}_k \sim MVN(b_k\mathbf{1}_N, \boldsymbol{\Sigma_k}),$$

where, $\boldsymbol{\Sigma} = \left(\boldsymbol{I}_N - \phi\widetilde{\boldsymbol{W}}\right)^{-1}M_k$ is a positive definite covariance matrix.

The covariance matrix $\boldsymbol{\Sigma}$ can be further reparametrized by:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{k}} &= \left(\boldsymbol{I_n} - \phi\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{M_k} \\ &= \left(\boldsymbol{I_n} - \phi\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{\lambda}_{\mathrm{k}}D^{-1} \\ &= \boldsymbol{\lambda}_{\mathrm{k}}\left(\boldsymbol{D} - \phi\boldsymbol{D}\widetilde{\boldsymbol{W}}\right)^{-1} \\ &= \boldsymbol{\lambda}_{\mathrm{k}}(\boldsymbol{D} - \phi\boldsymbol{W})^{-1}, \quad \text{(B8)}\end{aligned}$$

where, $\boldsymbol{D} = \boldsymbol{diag}(\boldsymbol{W}_{1+}, \boldsymbol{W}_{2+}, \ldots, \boldsymbol{W}_{N+})$.

To simplify notation, we denote $\boldsymbol{L} = \boldsymbol{D} - \phi\boldsymbol{W}$. The CARD model defined in equations (B1) and (B3) in the main text contains several hyper-parameters including $\boldsymbol{b}_k, \boldsymbol{\lambda}_{\mathrm{k}}$, ϕ and σ_e^2. We specify priors on each of them and infer them based on the data at hand.

For ϕ, this is the spatial autocorrelation parameter in the CAR model and represents the property parameter that ensures the $\boldsymbol{L}$ matrix to be positive definite[54]. We followed previous

research[55] and specified a discrete uniform distribution on ϕ by placing it equally on seven grid values as (0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99). These values lead to a wide range of spatial correlation structures modeled by the CAR model. We selected the ϕ value that achieves the highest log-likelihood as the final estimate. For the cell type specific scaling factor $\boldsymbol{\lambda}_{\mathrm{k}}$, we specified a conjugate inverse-gamma distribution $\boldsymbol{\lambda}_{\mathrm{k}} \sim InvG(\alpha, \beta)$. In the inverse-gamma distribution, we set the shape parameter α = 1.0 and we followed [56–58] to set the scale parameter $\beta = \frac{\#spots}{2.0}$ to ensure its adaptivity in balancing the information from data and prior assumption, through which appropriate optimization for $\boldsymbol{\lambda}_k$ can be achieved. For the hyperparameters $\boldsymbol{b}_{\mathrm{k}}$ and σ_e^2, we assigned the non-informative priors that are proportional to one.

Our goal is to infer the cell type composition matrix $\boldsymbol{V}$. To do so, we perform optimization based on the log likelihood in the following form with the grided value ϕ:

$$logPr(\boldsymbol{V}, \boldsymbol{\lambda}_{\mathrm{k}}, \sigma_e^2, \boldsymbol{b}_k | \boldsymbol{B}, \boldsymbol{X}, \phi, \alpha, \beta)$$

$$= -\frac{G*n}{2} log\sigma_e^2 - \frac{1}{2\sigma_e^2}\sum_i (\boldsymbol{X}_i - \boldsymbol{B}\boldsymbol{V}_i^T)^T (\boldsymbol{X}_i - \boldsymbol{B}\boldsymbol{V}_i^T) - \frac{\sum_k log\left(det\left(\boldsymbol{\lambda}_{\mathrm{k}}\boldsymbol{L}^{-1}\right)\right)}{2}$$

$$- \sum_k \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}} (\boldsymbol{V}_k - \boldsymbol{b}_k 1_{\boldsymbol{n}})^{\boldsymbol{T}} \boldsymbol{L} (\boldsymbol{V}_{\boldsymbol{k}} - \boldsymbol{b}_k 1_{\boldsymbol{n}}) - \sum_k ((\alpha + 1) log\boldsymbol{\lambda}_{\mathrm{k}} + \beta \boldsymbol{\lambda}_{\mathrm{k}}^{-\mathbf{1}})$$

$$= -\frac{G*n}{2} log\sigma_e^2 - \frac{1}{2\sigma_e^2} \left| |\boldsymbol{X} - \boldsymbol{B}\boldsymbol{V}^{\boldsymbol{T}}| \right|^2 + \frac{n}{2}\sum_k log\frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}} + \frac{k}{2} log\left(det(\boldsymbol{L})\right)$$

$$- \sum_k \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}} (\boldsymbol{V}_k - \boldsymbol{b}_k \mathbf{1}_n)^{\boldsymbol{T}} \boldsymbol{L} (\boldsymbol{V}_k - \boldsymbol{b}_k \mathbf{1}_n) - \sum_k ((\alpha + 1) log\boldsymbol{\lambda}_{\mathrm{k}} + \beta \boldsymbol{\lambda}_{\mathrm{k}}^{-1}). \quad \text{(B9)}$$

We developed an iterative algorithm to perform constrained optimization with non-negativity constraints on each element of $\boldsymbol{V}$. The iterative algorithm iterates through $\boldsymbol{V}$ and the three hyper-parameters ($\boldsymbol{b}_k, \boldsymbol{\lambda}_{\mathrm{k}}$, and σ_e^2) to minimize the negative log-likelihood in each iteration.

With the estimated cell type composition matrix $\boldsymbol{V}$, we normalized it further to ensure that each row of $\boldsymbol{V}$ has a summation of one. We describe the detailed optimization algorithm for each set of parameters in the following subsections.

B.5.3.1 Optimization for V

Maximizing the above log-likelihood function $logPr(\boldsymbol{V}, \boldsymbol{\lambda}_{\mathrm{k}}, \sigma_e^2, \boldsymbol{b}_k|\boldsymbol{B}, \boldsymbol{X}, \phi, \alpha, \beta)$ is equivalent to minimizing the negative log-likelihood Q = $-logPr(\boldsymbol{V}, \boldsymbol{\lambda}_{\mathrm{k}}, \sigma_e^2, \boldsymbol{b}_k|\boldsymbol{B}, \boldsymbol{X}, \phi, \alpha, \beta)$. To perform constrained optimization, we follow the common NMF framework (Lee and Seung 2000, Burred 2014, Cai et al. 2010) to derive a multiplicative learning role to ensure the non-negativity of $\boldsymbol{V}$. Specifically, the negative log-likelihood Q is:

$$Q = \frac{G*n}{2} log\sigma_e^2 + \frac{1}{2\sigma_e^2}\left||\boldsymbol{X} - \boldsymbol{B}\boldsymbol{V}^{\boldsymbol{T}}|\right|^2 - \frac{n}{2}\sum_k log\frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}} - \frac{k}{2}log\big(det(\boldsymbol{L})\big)$$

$$+ \sum_k \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{V}_k - \boldsymbol{b}_k 1_{\boldsymbol{n}})^{\boldsymbol{T}}\boldsymbol{L}(\boldsymbol{V}_{\boldsymbol{k}} - \boldsymbol{b}_k 1_{\boldsymbol{n}}) + \sum_k ((\alpha+1)log\boldsymbol{\lambda}_{\mathrm{k}} + \beta\boldsymbol{\lambda}_{\mathrm{k}}^{-\mathbf{1}})$$

$$= \frac{G*n}{2} log\sigma_e^2 + \frac{\mathbf{1}}{2\sigma_e^2}\left(tr\left(\boldsymbol{X} - \sum_k \boldsymbol{B}_k\boldsymbol{V}_k^T\right)\left(\boldsymbol{X} - \sum_k \boldsymbol{B}_k\boldsymbol{V}_k^T\right)^T\right)$$

$$- \frac{n}{2}\sum_k log\frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}} - \frac{K}{2}log\big(det(\boldsymbol{L})\big) + \sum_k \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{V}_k - \boldsymbol{b}_k 1_{\boldsymbol{n}})^{\boldsymbol{T}}\boldsymbol{L}(\boldsymbol{V}_k - \boldsymbol{b}_k 1_{\boldsymbol{n}})$$

$$+ \sum_k ((\alpha+1)log\boldsymbol{\lambda}_{\mathrm{k}} + \beta\boldsymbol{\lambda}_{\mathrm{k}}^{-\mathbf{1}})$$

$$= \frac{G*n}{2} log\sigma_e^2 + \frac{1}{2\sigma_e^2}(tr(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}) - 2tr(\textstyle\sum_k \boldsymbol{B}_k\boldsymbol{V}_k^T\boldsymbol{X}^T) + tr((\textstyle\sum_k \boldsymbol{B}_k\boldsymbol{V}_k^T)(\textstyle\sum_{\boldsymbol{k}} \boldsymbol{V}_k\boldsymbol{B}_k^T))$$

$$- \frac{n}{2}\sum_k log\frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}} - \frac{K}{2}log\big(det(\boldsymbol{L})\big) + \sum_k \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{V}_k^T\boldsymbol{L}\boldsymbol{V}_k - 2\boldsymbol{b}_k\boldsymbol{V}_k^T\boldsymbol{L}\mathbf{1}_n + \boldsymbol{b}_k^2\mathbf{1}_n^T\boldsymbol{L}\mathbf{1}_n) + \sum_k((\alpha+1)log\boldsymbol{\lambda}_{\mathrm{k}} + \beta\boldsymbol{\lambda}_{\mathrm{k}}^{-\mathbf{1}}) \quad \text{(B10)}$$

When we ignore the constants in the above equation, we can obtain the Q with respect to the column vector V_k as

$$Q_k \propto \frac{\mathbf{1}}{\mathbf{2}\sigma_e^2}(-2tr(\sum_k \boldsymbol{B}_k \boldsymbol{V}_k^T \boldsymbol{X}^T) + tr(\left(\sum_k \boldsymbol{B}_k \boldsymbol{V}_k^T)(\sum_{\boldsymbol{k}} \boldsymbol{V}_k \boldsymbol{B}_k^T)\right) + \sum_k \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{V}_k^T \boldsymbol{L} \boldsymbol{V}_k - 2\boldsymbol{b}_k \boldsymbol{V}_k^T \boldsymbol{L}\mathbf{1}_n)$$

$$\propto \frac{\mathbf{1}}{\mathbf{2}\sigma_e^2}\left(-2tr(\boldsymbol{X}^T \boldsymbol{B}_k \boldsymbol{V}_k^T) + tr\left(\boldsymbol{B}_k \boldsymbol{V}_k^T \boldsymbol{V}_k \boldsymbol{B}_k^T + 2\boldsymbol{B}_k \boldsymbol{V}_k^T \sum_{j\neq k} \boldsymbol{V}_j \boldsymbol{B}_j^T\right)\right) + \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{V}_k^T \boldsymbol{L} \boldsymbol{V}_k - 2\boldsymbol{b}_k \boldsymbol{V}_k^T \boldsymbol{L}\mathbf{1}_n) \quad \text{(B11)}$$

We can obtain the partial derivative of Q with respect to the column vector $\boldsymbol{V_k}$ as

$$\nabla_{v_k} Q_k = \frac{\partial Q_k}{\partial \boldsymbol{V}_k} = \frac{1}{2\sigma_e^2}\left(-2\boldsymbol{X}^T \boldsymbol{B}_k + 2\boldsymbol{V}_k \boldsymbol{B}_k^T \boldsymbol{B}_k + 2\sum_{j\neq k} \boldsymbol{V}_j \boldsymbol{B}_j^T \boldsymbol{B}_k\right) + \frac{1}{2\boldsymbol{\lambda}_{\mathrm{k}}} 2\boldsymbol{L}\boldsymbol{V}_k - \frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}} \boldsymbol{b}_k \boldsymbol{L}\mathbf{1}_n \quad \text{(B12)}$$

We substitute $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{\phi W}$ and simplify the above equation as

$$\nabla_{v_k} Q_k = -\frac{1}{\sigma_e^2}\boldsymbol{X}^T \boldsymbol{B}_k + \frac{1}{\sigma_e^2}(\boldsymbol{V}_k \boldsymbol{B}_k^T \boldsymbol{B}_k + \sum_{j\neq k} \boldsymbol{V}_j \boldsymbol{B}_j^T \boldsymbol{B}_k) + \frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}}\{(\boldsymbol{D} - \boldsymbol{\phi W})\boldsymbol{V}_k - \boldsymbol{b}_k(\boldsymbol{D} - \boldsymbol{\phi W})\mathbf{1}_n\}$$

$$= -\frac{1}{\sigma_e^2}\boldsymbol{X}^T \boldsymbol{B}_k - \frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{\phi W V}_k + \boldsymbol{b}_k \boldsymbol{D}\mathbf{1}_n) + \frac{1}{\sigma_e^2}(\boldsymbol{V}_k \boldsymbol{B}_k^T \boldsymbol{B}_k + \sum_{j\neq k} \boldsymbol{V}_j \boldsymbol{B}_j^T \boldsymbol{B}_k) + \frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{D}\boldsymbol{V}_k + \boldsymbol{\phi} \boldsymbol{b}_k \boldsymbol{W}\mathbf{1}_n)$$

$$= \nabla_{v_k}^+ Q_k - \nabla_{v_k}^- Q_k$$

where $\nabla_{v_k}^+ Q_k = \frac{1}{\sigma_e^2}(\boldsymbol{V}_k \boldsymbol{B}_k^T \boldsymbol{B}_k + \sum_{j\neq k} \boldsymbol{V}_j \boldsymbol{B}_j^T \boldsymbol{B}_k) + \frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{D}\boldsymbol{V}_k + \boldsymbol{\phi} \boldsymbol{b}_k \boldsymbol{W}\mathbf{1}_{\mathrm{n}})$ represents the positive terms in the gradient and $\nabla_{v_k}^- Q_k = \frac{1}{\sigma_e^2}\boldsymbol{X}^T \boldsymbol{B}_k + \frac{1}{\boldsymbol{\lambda}_{\mathrm{k}}}(\boldsymbol{\phi W V}_k + \boldsymbol{b}_k \boldsymbol{D}\mathbf{1}_{\mathrm{n}})$ represents the negative terms in the gradient. Following (Lee and Seung 2000, Burred 2014, Lee, Seung and Saul 2002), we have the following updating rules where

$$\boldsymbol{V}_k \leftarrow \boldsymbol{V}_k \circ \frac{\nabla_{v_k}^{-} Q_k}{\nabla_{v_k}^{+} \boldsymbol{Q}_k} \tag{B13}$$

Hence, element wisely, the equation (13) leads to the following updating rule:

$$\boldsymbol{V}_{jk} \leftarrow \boldsymbol{V}_{jk} \frac{(\boldsymbol{\lambda}_k \boldsymbol{X}^T \boldsymbol{B}_k + \sigma_e^2(\phi \boldsymbol{W} \boldsymbol{V}_k + \boldsymbol{b}_k \boldsymbol{D} \boldsymbol{1}_n))_j}{(\boldsymbol{\lambda}_k (\boldsymbol{V}_k \boldsymbol{B}_k^T \boldsymbol{B}_k + \sum_{j \neq k} \boldsymbol{V}_j \boldsymbol{B}_j^T \boldsymbol{B}_k) + \sigma_e^2(\boldsymbol{D} \boldsymbol{V}_k + \phi \boldsymbol{b}_k \boldsymbol{W} \boldsymbol{1}_n))_j} \tag{B14}$$

Thus, for each cell type k, we use the equation (14) to update the cell type composition across spatial locations.

B.5.3.2 Optimization for b_k

We take the partial derivative of complete loglikelihood with respect to b_k and set it to be zero. We can obtain a closed form update for $\boldsymbol{b_k}$ as

$$\frac{\partial log P(\boldsymbol{b}_k)}{\partial \boldsymbol{b}_k} = \frac{1}{\sigma_{Lk}^2} \boldsymbol{V}_k^T \boldsymbol{L} \boldsymbol{1}_n - \frac{1}{\sigma_{Lk}^2} b_k \boldsymbol{1}_n^T \boldsymbol{L} \boldsymbol{1}_n = 0$$

$$\rightarrow b_k = \boldsymbol{V}_k^T \boldsymbol{L} \boldsymbol{1}_n (\boldsymbol{1}_n^T \boldsymbol{L} \boldsymbol{1}_n)^{-\boldsymbol{1}} \tag{B15}$$

B.5.3.3 Optimization for λ_k

We take the partial derivative of complete loglikelihood with respect to $\boldsymbol{\lambda}_k$ and set it to be zero. We can obtain a closed form update for $\boldsymbol{\lambda}_k$ as

$$log P(\boldsymbol{\lambda}_k) \propto \frac{n}{2} log \frac{1}{\boldsymbol{\lambda}_k} - \frac{1}{2\boldsymbol{\lambda}_k} (\boldsymbol{V}_k - \boldsymbol{b}_k 1_{\boldsymbol{n}})^T \boldsymbol{L} (\boldsymbol{V}_k - \boldsymbol{b}_k 1_{\boldsymbol{n}}) - (\alpha + 1) log \boldsymbol{\lambda}_k - \beta \boldsymbol{\lambda}_k^{-\boldsymbol{1}}$$

$$\propto -\left(\frac{n}{2} + \alpha + 1\right) log \boldsymbol{\lambda}_k - \left(\frac{(\boldsymbol{V}_k - \boldsymbol{b}_k \boldsymbol{1}_n)^T \boldsymbol{L} (\boldsymbol{V}_k - \boldsymbol{b}_k \boldsymbol{1}_n)}{2} + \beta\right) \boldsymbol{\lambda}_k^{-\boldsymbol{1}}$$

$$\frac{\partial log P(\boldsymbol{\lambda}_k)}{\partial \boldsymbol{\lambda}_k} = -\frac{\left(\frac{n}{2} + \alpha + 1\right)}{\boldsymbol{\lambda}_k} + \frac{1}{\boldsymbol{\lambda}_k^2} \left(\frac{(\boldsymbol{V}_k - \boldsymbol{b}_k \boldsymbol{1}_n)^{\boldsymbol{T}} \boldsymbol{L} (\boldsymbol{V}_k - \boldsymbol{b}_k \boldsymbol{1}_n)}{2} + \beta\right)$$

$$\rightarrow \widehat{\boldsymbol{\lambda}_k} = \frac{\frac{(\boldsymbol{v}_k - \boldsymbol{b}_k \mathbf{1}_n)^{\boldsymbol{T}} \boldsymbol{L} (\boldsymbol{v}_k - \boldsymbol{b}_k \mathbf{1}_n)}{2} + \beta}{\left(\frac{n}{2} + \alpha + 1\right)} \tag{B16}$$

B.5.3.4 Optimization for σ_e^2

We take the partial derivative of complete loglikelihood with respect to σ_e^2 and set it to be zero. We can obtain a closed form update for σ_e^2 as

$$\frac{\partial log P(\sigma_e^2)}{\partial \sigma_e^2} = -\frac{G * n}{2\sigma_e^2} + \frac{\left| |\boldsymbol{X} - \boldsymbol{B}\boldsymbol{V}^{\boldsymbol{T}}| \right|^2}{2\sigma_e^4} = 0$$

$$\rightarrow \widehat{\sigma_e^2} = \frac{\left| |\boldsymbol{X} - \boldsymbol{B}\boldsymbol{V}^{\boldsymbol{T}}| \right|^2}{G * n} \tag{B17}$$

Appendix C. Chapter 4 (IRIS) Supplementary Text

C.1 Evaluations in Real Data Sets

In the human DLPFC dataset, because we have the ground truth label, we directly compared identified spatial domains by each method with the ground truth via adjusted rand index (ARI) by using the compare function in the igraph R package (v1.0.0). Specifically, the details of calculating ARI is provided in ref (Ma and Zhou 2022)

In the remaining datasets, we evaluated the performance of different methods by calculating the spatial CHAOS score as there was no ground truth available for these data. Specifically, to calculate the spatial chaos in each slice t, we first constructed a one-nearest-neighbor (1NN) graph for each spatial locations in each domain r. We then specified the edge weight between spot I and j by the following:

$$\boldsymbol{E}_{t_{rij}} = \begin{cases} e_{ij}, & if\ spatial\ location\ i\ and\ j\ are\ one-nearest\ neighbor \\ 0, & otherwise \end{cases}$$

where, e_{ij} is the Euclidean distance and the spatial CHAOS was calculated as the mean of the edge weight across spatial domains:

$$CHAOS = \frac{\sum_{r=1}^{R} \sum_{i,j} \boldsymbol{E}_{t_{rij}}}{N_t}.$$

where N_t is the total number of spatial locations in slice t; R is the total number of spatial domains; i, j belongs to all the i-th, and j-th spatial locations belonging to spatial domain r in slice t.

C.2 Robustness and sensitivity analysis

C.2.1 Selection of the penalty parameters β and λ

Following previous studies, we fixed the penalty parameters $\beta = 1000$ and $\lambda = 2000$ for all datasets. Such choice is robust when varying the number of spatial domains in all real datasets. In addition, we evaluated different choices of β and λ in the human DLPFC dataset where the ground truth spatial domains are known. We found that IRIS is generally robust for a reasonable range of β and λ in the baseline setting analysis (**Figure S4.20**).

C.2.2 Algorithmic innovations to ensure scalability of IRIS

IRIS iteratively updates the spatial domain labels and cell type compositions for all tissue slices to ensure optimal clustering performance. IRIS also relies on several algorithmic innovations to make it highly computationally efficient. First, the modeling framework of IRIS is in essence based on a non-negative matrix factorization (NMF) model, expressing the mean gene expression profile in the spatial transcriptomics as a linear function of that from scRNA-seq. The NMF modeling framework streamlines the inference procedure and facilitates scalable computation. Second, IRIS detects spatial domains based on the concatenated cell type composition matrix, which represents a low-dimensional sub-space with enriched signals for the noisy high-dimensional gene expression data. Because of this, the preprocessing steps of IRIS are relatively simple and do not contain a dimension reduction step. Third, IRIS makes use of the fast multiplicative updating rules (Lee and Seung 2000, Janecek and Tan 2011) for updating the nonnegative cell type composition matrix in a supervised fashion. The multiplicative updating rules allow for algorithmic optimization without explicit inverse of the graph Laplacian matrix, which incurs heavy computation burden. Fourth, IRIS relies on the computationally efficient K-means clustering algorithms for updating the spatial domain labels in each optimization iteration. Fifth, IRIS takes advantage of the sparse matrix computation properties when constructing the graph Laplacian matrix, which induces a local geometric structure to further reduce the

computational cost. Finally, while IRIS is implemented in R, its core deconvolution algorithm is implemented with an efficient C++ code that is linked back to the main functions of IRIS through Rcpp, ensuring scalable computation.

C.3 Spatial Transcriptomics and Single-cell RNA-seq Datasets.

We applied IRIS to analyze four published spatial transcriptomics datasets collected by different techniques, with distinct spatial resolutions, and from multiple species and tissues. For each spatial transcriptomics data, we obtained an external scRNA-seq collected on the same type of tissue but with a different sequencing technology to serve as the reference.

Human prefrontal cortex (DLPFC) data by 10x Visium

We downloaded human prefrontal cortex (DLPFC) data (Maynard et al. 2021) generated by 10x Visium from the spatialLIBD website (http://spatial.libd.org/spatialLIBD/). This study sequenced 12 brain tissue slices measured on 33,538 genes and 3,460 ~ 4,789 spatial locations from three donors. Each tissue slice contains seven spatial domains including six cortical layers and white matter. We obtained the domain annotations for each measured spatial location from the original study and used them as the ground truth to evaluate the accuracy of different methods on spatial domain detection. For quality control, for each tissue slice in turn, we retained genes with non-zero expression on at least five spots. We filtered spots that was labeled “discarded” from the original study and retained spots that have a minimum of 100 UMIs following (Cable et al. 2022, Ma and Zhou 2022). These filtering criteria led to a final set of 17,151 genes and 3,454 ~ 4,730 spatial locations across the 12 tissue slices for analysis. Besides the spatial transcriptomics, we also obtained a single nuclear RNA-seq (snRNA-seq) data sequenced by 10x Chromium technology on human post-mortem brain to serve as the reference (Mathys et al. 2019).

Because the DLPFC data contains known spatial domain annotations, we examined the data extensively and performed two types of analyses by either analyzing consecutive tissue slices from the same donor (e.g., on samples 151507-151510) or analyzing inconsecutive tissue slices from different donors (e.g., samples 151509, 151671, and 151675). The latter analyses represent a more challenging scenario than the former as the spatial domains on different slices can be of different shape.

High-resolution mouse spermatogenesis data by Slide-seq

We obtained the mouse spermatogenesis data by Slide-seq (Chen et al. 2021) from the link provided in the original paper (https://www.dropbox.com/s/ygzpj0d0oh67br0/Testis_Slideseq_Data.zip?dl=0&file_subpath=%2FData). This study consists of gene expression measurements on 23,515 ~ 24,450 genes and 27,194 ~ 42,776 spatial locations on testicular tissues harvested from six adult male mice with 3-10 months of age. One tissue slice is collected for each mouse. Among the six mice, three are leptin-deficient diabetic mice and three are matching wild-type (WT) mice. For quality control, for each tissue slice in turn, we retained genes with non-zero expression on at least five spots. We retained spots that have a minimum of 100 UMIs following (Cable et al. 2022, Ma and Zhou 2022). These filtering criteria led to a final set of 18,865 ~ 19,457 genes and 25,377 ~ 42,776 spatial locations across all six tissue slices for analysis. Besides the spatial transcriptomics, we also obtained a scRNA-seq data sequenced from Drop-seq on six batches of 7- to 9-week-old adult male mice to serve as the reference (Green et al. 2018). In the spermatogenesis data, we performed four analyses: we either analyzed two tissue slices of either diabetic mouse and WT mouse (dataset WT3_Puck7 and dataset Diabetes2_Puck10) following the original paper; or analyzed all three

tissue slices of the WT mice (dataset WT1_T3, WT2_Puck24, and WT3_Puck7) or that of the diabetic mice (dataset Diabetes1_T4, and Diabetes2_Puck10, Diabetes3_Puck11).

Sub-cellular resolution mouse olfactory bulb (MOB) data by Stereo-seq

We downloaded the sub-cellular mouse olfactory bulb from the MOSTA website (https://db.cngb.org/stomics/mosta/download.html). This study consists of gene expression measurements on 23,815 ~ 26,145 genes and 104,931 ~ 107,416 spatial locations from two adjacent mouse olfactory bulb sections. For quality control, for each tissue slice in turn, we retained genes with non-zero expression on at least five spots. We retained spots that have a minimum of 100 UMIs following (Cable et al. 2022, Ma and Zhou 2022). These filtering criteria led to a final set of 23,815 ~ 26,145 genes and 103,610 ~ 106,770 spatial locations across six tissue slices for analysis. Besides the spatial transcriptomics, we also obtained a scRNA-seq data sequenced by 10x Chromium on mouse olfactory bulb from six mice models to serve as the reference (Tepe et al. 2018). In the MOB data, we analyzed the two sections together (Mouse_olfa_S1 data and Mouse_olfa_S2)

High-resolution human breast cancer data by 10x Xenium

We downloaded the high-resolution human breast cancer data from 10x Genomics website (https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast). This study consists of gene expression measurements on 313 genes and 118,708 ~ 167,782 spatial locations from two adjacent human breast cancer slices. For quality control, for each tissue slice in turn, we retained genes with non-zero expression on at least five spots. We retained spots that have a minimum of 100 UMIs following (Cable et al. 2022, Ma and Zhou 2022). These filtering criteria led to a final set of 313 genes and 90,424 ~ 124,945 spatial locations across the two tissue slices for analysis. Besides the spatial transcriptomics, we also obtained a scRNA-seq data

sequenced by 10x Chromium from 26 breast cancer patients to serve as the reference (Wu et al. 2021). In the human breast cancer data, we analyzed the two tissue slices together (Rep1 and Rep2)

C.4 Challenging settings for the evaluation on the benchmark DLPFC dataset

In the DLPFC dataset by 10x Visium, we examined four settings in total: (1) baseline setting when the slices are from the same donor that share high similarity; (2) challenging setting when we use a scRNA-seq reference with one missing cell type information at a time; (3) challenging setting when we use a scRNA-seq reference with mis-classified cell type information by randomly merging two cell types; (4) challenging setting when the slices are from different donors that share low similarity in structures. For each setting, we applied IRIS with the input of multiple slices spatial transcriptomics and scRNA-seq reference data. Specifically, for the baseline setting (1), when the slices are from the same donors and challenging setting when the slices are from different donors, we use the scRNA-seq data from 10x Chromium on the post-mortem brain tissue with 44 cell types to serve as the reference (Mathys et al. 2019). For the challenging setting (2) when there is a missing cell type in the scRNA-seq data, we remove one cell type in the scRNA-seq reference at a time. Therefore, it consists of 44 scenarios in total as there are 44 cell types in the complete scRNA-seq reference. (3) When there is mis-classified cell type in the scRNA-seq data, we randomly merge two cell types among the 44 cell types to create an artificially mis-classified cell type. Therefore, it consists of 946 scenarios in total as there are 946 pair-wise combinations.

Bibliography

GENCODE V41.

10XGenomics. 10X Genomics: Visium Spatial Gene Expression

Abdelaal, T., S. Mourragui, A. Mahfouz & M. J. Reinders (2020) SpaGE: spatial gene enhancement using scRNA-seq. *Nucleic acids research,* 48**,** e107-e107.

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

Akiyama, K., N. Ohga, N. Maishi, Y. Hida, K. Kitayama, T. Kawamoto, T. Osawa, Y. Suzuki, N. Shinohara, K. Nonomura, M. Shindoh & K. Hida (2013) The F-prostaglandin receptor is a novel marker for tumor endothelial cells in renal cell carcinoma. *Pathol Int,* 63**,** 37-44.

Al-Kafaji, G., M. A. Sabry & M. Bakhiet (2016) Increased expression of mitochondrial DNA-encoded genes in human renal mesangial cells in response to high glucose-induced reactive oxygen species. *Molecular Medicine Reports,* 13**,** 1774-1780.

Allen, C., Y. Chang, Q. Ma & D. Chung (2022) MAPLE: a hybrid framework for multi-sample spatial transcriptomics data. *bioRxiv***,** 2022.02. 28.482296.

Alves, M. G., A. D. Martins, J. E. Cavaco, S. Socorro & P. F. Oliveira (2013) Diabetes, insulin-mediated glucose metabolism and Sertoli/blood-testis barrier function. *Tissue barriers,* 1**,** e23992.

Aly, H. A. (2021) Mitochondria-mediated apoptosis induced testicular dysfunction in diabetic rats: ameliorative effect of resveratrol. *Endocrinology,* 162**,** bqab018.

Andersson, A., J. Bergenstråhle, M. Asp, L. Bergenstråhle, A. Jurek, J. F. Navarro & J. Lundeberg (2020) Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications biology,* 3**,** 1-8.

Angerer, P., L. Simon, S. Tritschler, F. A. Wolf, D. Fischer & F. J. Theis (2017) Single cells make big data: New challenges and opportunities in transcriptomics. *Current opinion in systems biology,* 4**,** 85-91.

Antonetti, D. A., C. Reynet & C. R. Kahn (1995) Increased expression of mitochondrial-encoded genes in skeletal muscle of humans with diabetes mellitus. *The Journal of clinical investigation,* 95**,** 1383-1388.

Aran, D., Z. C. Hu & A. J. Butte (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology,* 18.

Argelaguet, R., A. S. Cuomo, O. Stegle & J. C. Marioni (2021) Computational principles and challenges in single-cell data integration. *Nature biotechnology,* 39**,** 1202-1215.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock & G. O. Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25**,** 25-29.

Asp, M., J. Bergenstråhle & J. Lundeberg (2020) Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. *BioEssays : news and reviews in molecular, cellular and developmental biology,* 42**,** 1900221-NA.

Baghban, R., L. Roshangar, R. Jahanban-Esfahlan, K. Seidi, A. Ebrahimi-Kalan, M. Jaymand, S. Kolahian, T. Javaheri & P. Zare (2020) Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Communication and Signaling,* 18**,** 1-19.

Ballester, J., M. C. Muñoz, J. Domínguez, T. Rigau, J. J. Guinovart & J. E. Rodríguez‐Gil (2004) Insulin‐dependent diabetes affects testicular function by FSH‐and LH‐linked mechanisms. *Journal of andrology,* 25**,** 706-719.

Banerjee, S., B. P. Carlin & A. E. Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data*.

Barnes, R. M., B. A. Firulli, S. J. Conway, J. W. Vincentz & A. B. Firulli (2010) Analysis of the Hand1 Cell Lineage Reveals Novel Contributions to Cardiovascular, Neural Crest, Extra-Embryonic, and Lateral Mesoderm Derivatives. *Developmental Dynamics,* 239**,** 3086-3097.

Baron, C. S., A. Barve, M. J. Muraro, R. van der Linden, G. Dharmadhikari, A. Lyubimova, E. J. P. de Koning & A. van Oudenaarden (2019) Cell Type Purification by Single-Cell Transcriptome-Trained Sorting. *Cell,* 179**,** 527-542 e19.

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

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological),* 36**,** 192-225.

Biancalani, T., G. Scalia, L. Buffoni, R. Avasthi, Z. Lu, A. Sanger, N. Tokcan, C. R. Vanderburg, Å. Segerstolpe & M. Zhang (2021) Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature methods,* 18**,** 1352-1362.

Bivand, R., L. Anselin, O. Berke, A. Bernat, M. Carvalho, Y. Chun, C. Dormann, S. Dray, R. Halbersma & N. Lewin-Koh (2011) spdep: Spatial dependence: weighting schemes,

statistics and models. *R package version 0.5-31, URL http://CRAN. R-project. org/package= spdep*.

Bolton, E. E., Y. L. Wang, P. A. Thiessen & S. H. Bryant (2010) PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports In Computational Chemistry, Vol 4,* 4**,** 217-241.

Bove, A., D. Gradeci, Y. Fujita, S. Banerjee, G. Charras & A. R. Lowe (2017) Local cellular neighborhood controls proliferation in cell competition. *Molecular biology of the cell,* 28**,** 3215-3228.

Brook, D. (1964) On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika,* 51**,** 481-483.

Buettner, F., N. Pratanwanich, D. J. McCarthy, J. C. Marioni & O. Stegle (2017) f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol,* 18**,** 212.

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

Burred, J. J. (2014) Detailed derivation of multiplicative update rules for NMF. *Paris, France*.

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

Cable, D. M., E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen & R. A. Irizarry (2021) Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology***,** 1-10.

--- (2022) Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology,* 40**,** 517-526.

Cai, B., J. Zhang, H. Li, C. Su & H. Zhao (2022) Statistical Inference of Cell-type Proportions Estimated from Bulk Expression Data. *arXiv preprint arXiv:2209.04038*.

Cai, D., X. He, J. Han & T. S. Huang (2010) Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence,* 33**,** 1548-1560.

Camp, J. G., F. Badsha, M. Florio, S. Kanton, T. Gerber, M. Wilsch-Brauninger, E. Lewitus, A. Sykes, W. Hevers, M. Lancaster, J. A. Knoblich, R. Lachmann, S. Paabo, W. B. Huttner & B. Treutlein (2015) Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings Of the National Academy Of Sciences Of the United States Of America,* 112**,** 15672-15677.

Cang, Z. & Q. Nie (2020) Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications,* 11**,** 1-13.

Chen, A., S. Liao, M. Cheng, K. Ma, L. Wu, Y. Lai, X. Qiu, J. Yang, J. Xu & S. Hao (2022) Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell,* 185**,** 1777-1792. e21.

Chen, G., J. Sima, M. Jin, K.-y. Wang, X.-j. Xue, W. Zheng, Y.-q. Ding & X.-b. Yuan (2008) Semaphorin-3A guides radial migration of cortical neurons during development. *Nature neuroscience,* 11**,** 36-44.

Chen, H., E. Murray, A. Sinha, A. Laumas, J. Li, D. Lesman, X. Nie, J. Hotaling, J. Guo & B. R. Cairns (2021) Dissecting mammalian spermatogenesis using spatial transcriptomics. *Cell reports,* 37**,** 109915.

Chen, K. H., A. N. Boettiger, J. R. Moffitt, S. Wang & X. Zhuang (2015) Spatially resolved, highly multiplexed RNA profiling in single cells. *Science,* 348**,** aaa6090.

Cheng, C. W., J. R. Hsiao, C. C. Fan, Y. K. Lo, C. Y. Tzen, L. W. Wu, W. Y. Fang, A. J. Cheng, C. H. Chen, I. S. Chang, S. S. Jiang, J. Y. Chang & A. Y. Lee (2016) Loss of GDF10/BMP3b as a prognostic marker collaborates with TGFBR3 to enhance chemotherapy resistance and epithelial-mesenchymal transition in oral squamous cell carcinoma. *Mol Carcinog,* 55**,** 499-513.

Cho, C.-S., J. Xi, Y. Si, S.-R. Park, J.-E. Hsu, M. Kim, G. Jun, H. M. Kang & J. H. Lee (2021) Microscopic examination of spatial transcriptome using Seq-Scope. *Cell,* 184**,** 3559-3572. e22.

Choe, K., U. Pak, Y. Pang, W. Hao & X. Yang (2023) Advances and Challenges in Spatial Transcriptomics for Developmental Biology. *Biomolecules,* 13**,** 156.

Chu, L. F., N. Leng, J. Zhang, Z. G. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart & J. A. Thomson (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology,* 17**,** 173.

Chung, W.-S., P. B. Verghese, C. Chakraborty, J. Joung, B. T. Hyman, J. D. Ulrich, D. M. Holtzman & B. A. Barres (2016) Novel allele-dependent role for APOE in controlling the rate of synapse pruning by astrocytes. *Proceedings of the National Academy of Sciences,* 113**,** 10186-10191.

Cobos, F. A., J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh & K. De Preter (2020) Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications,* 11**,** 1-14.

Comito, G., L. Ippolito, P. Chiarugi & P. Cirri (2020) Nutritional Exchanges Within Tumor Microenvironment: Impact for Cancer Aggressiveness. *Frontiers in oncology,* 10**,** 396-396.

Conway, J. R., A. Lex & N. Gehlenborg (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics,* 33**,** 2938-2940.

Cressie, N. (1992) STATISTICS FOR SPATIAL DATA. *Terra Nova,* 4**,** 613-617.

Daly, M. E., Q. T. Le, A. K. Jain, P. G. Maxim, A. Hsu, B. W. Loo, Jr., M. J. Kaplan, N. J. Fischbein, A. D. Colevas, H. Pinto & D. T. Chang (2011) Intensity-modulated radiotherapy for locally advanced cancers of the larynx and hypopharynx. *Head Neck,* 33**,** 103-11.

Danaher, P., Y. Kim, B. Nelson, M. Griswold, Z. Yang, E. Piazza & J. M. Beechem (2022) Advances in mixed cell deconvolution enable quantification of cell types in spatial transcriptomic data. *Nature communications,* 13**,** 385-NA.

Danielsson, F., M. Skogs, M. Huss, E. Rexhepaj, G. O'Hurley, D. Klevebring, F. Pontén, A. K. Gad, M. Uhlén & E. Lundberg (2013) Majority of differentially expressed genes are down-regulated during malignant transformation in a four-stage model. *Proceedings of the National Academy of Sciences,* 110**,** 6853-6858.

Das, S., A. Rai & S. N. Rai (2022) Differential Expression Analysis of Single-Cell RNA-Seq Data: Current Statistical Approaches and Outstanding Challenges. *Entropy,* 24**,** 995.

De Oliveira, V. (2010) Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics,* 64**,** 107-133.

Del Bigio, M. R. (2009) Ependymal cells: biology and pathology. *Acta neuropathologica,* 119**,** 55-73.

DeSisto, J., R. O'Rourke, H. E. Jones, B. Pawlikowski, A. D. Malek, S. Bonney, F. Guimiot, K. L. Jones & J. A. Siegenthaler (2020) Single-cell transcriptomic analyses of the developing meninges reveal meningeal fibroblast diversity and function. *Developmental cell,* 54**,** 43-59. e4.

Ding, G.-L., Y. Liu, M.-E. Liu, J.-X. Pan, M.-X. Guo, J.-Z. Sheng & H.-F. Huang (2015) The effects of diabetes on male fertility and epigenetic regulation during spermatogenesis. *Asian journal of andrology,* 17**,** 948.

Dong, M., A. Thennavan, E. Urrutia, Y. Li, C. M. Perou, F. Zou & Y. Jiang (2020) SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in bioinformatics,* 22**,** 416-427.

Dong, R. & G.-C. Yuan (2021) SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome biology,* 22**,** 1-10.

Dries, R., Q. Zhu, R. Dong, C.-H. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar & F. Bao (2021) Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology,* 22**,** 1-31.

Dudas, M., A. Wysocki, B. Gelpi & T.-L. Tuan (2008) Memory encoded throughout our bodies: molecular and cellular basis of tissue regeneration. *Pediatric research,* 63**,** 502-512.

Duò, A., M. D. Robinson & C. Soneson (2018) A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research,* 7**,** 1141-1141.

Eberwine, J., J.-Y. Sul, T. Bartfai & J. Kim (2014) The promise of single-cell sequencing. *Nature methods,* 11**,** 25-27.

Edsgärd, D., P. Johnsson & R. Sandberg (2018) Identification of spatial expression trends in single-cell gene expression data. *Nature methods,* 15**,** 339-342.

Efron, B. 2001. *Empirical Bayes analysis of a microarray experiment*. Stanford, Calif.: Division of Biostatistics, Stanford University.

Efron, B. & R. Tibshirani (2002) Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol,* 23**,** 70-86.

Elosua-Bayes, M., P. Nieto, E. Mereu, I. Gut & H. Heyn (2021) SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic acids research,* 49**,** e50-e50.

Eng, C.-H. L., M. J. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. J. Cronin, C. D. Karp, G.-C. Yuan & L. Cai (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature,* 568**,** 235-239.

Fan, J., N. Salathia, R. Liu, G. E. Kaeser, Y. C. Yung, J. L. Herman, F. Kaper, J. B. Fan, K. Zhang, J. Chun & P. V. Kharchenko (2016) Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods,* 13**,** 241-4.

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

Feng, Y., C. Guo, H. Wang, L. Zhao, W. Wang, T. Wang, Y. Feng, K. Yuan & G. Huang (2020) Fibrinogen-like protein 2 (FGL2) is a novel biomarker for clinical prediction of human breast cancer. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research,* 26**,** e923531-1.

Fiers, M. W., L. Minnoye, S. Aibar, C. Bravo González-Blas, Z. Kalender Atak & S. Aerts (2018) Mapping gene regulatory networks from single-cell omics data. *Briefings in functional genomics,* 17**,** 246-254.

Finak, G., A. McDavid, M. Yajima, J. Y. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley & R. Gottardo (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology,* 16**,** 278.

Fischl, A. M., P. M. Heron, A. J. Stromberg & T. S. McClintock (2014) Activity-Dependent Genes in Mouse Olfactory Sensory Neurons. *Chemical senses,* 39**,** 439-449.

Fu, F., X. Yang, M. Zheng, Q. Zhao, K. Zhang, Z. Li, H. Zhang & S. Zhang (2020) Role of Transmembrane 4 L Six Family 1 in the Development and Progression of Cancer. *Frontiers in molecular biosciences,* 7**,** 202-NA.

Fu, H., H. Xu, K. Chong, M. Li, K. S. Ang, H. K. Lee, J. Ling, A. Chen, L. Shao & L. Liu (2021) Unsupervised spatially embedded deep representation of spatial transcriptomics. *Biorxiv*.

Fujiwara, N. & J. W. Cave (2016) Partial conservation between mice and humans in olfactory bulb interneuron transcription factor codes. *Frontiers in neuroscience,* 10**,** 337.

Gadue, P., T. L. Huber, P. J. Paddison & G. M. Keller (2006) Wnt and TGF-beta signaling are required for the induction of an in vitro model of primitive streak formation using embryonic stem cells. *Proceedings Of the National Academy Of Sciences Of the United States Of America,* 103**,** 16806-16811.

Gao, L. L., J. Bien & D. Witten (2022) Selective inference for hierarchical clustering. *Journal of the American Statistical Association***,** 1-11.

Gayoso, A., R. Lopez, G. Xing, P. Boyeau, V. Valiollah Pour Amiri, J. Hong, K. Wu, M. Jayasuriya, E. Mehlman & M. Langevin (2022) A Python library for probabilistic analysis of single-cell omics data. *Nature biotechnology,* 40**,** 163-166.

Godec, J., Y. Tan, A. Liberzon, P. Tamayo, S. Bhattacharya, A. J. Butte, J. P. Mesirov & W. N. Haining (2016) Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity,* 44**,** 194-206.

Goeman, J. J., S. A. van de Geer, F. de Kort & H. C. van Houwelingen (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics,* 20**,** 93-99.

Green, C. D., Q. Ma, G. L. Manske, A. N. Shami, X. Zheng, S. Marini, L. Moritz, C. Sultan, S. J. Gurczynski & B. B. Moore (2018) A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-seq. *Developmental cell,* 46**,** 651-667. e10.

Greenough, T. C., J. R. Straubhaar, L. Kamga, E. R. Weiss, R. M. Brody, M. M. McManus, L. K. Lambrecht, M. Somasundaran & K. F. Luzuriaga (2015) A Gene Expression Signature That Correlates with CD8(+) T Cell Expansion in Acute EBV Infection. *Journal Of Immunology,* 195**,** 4185-4197.

Gudjohnsen, S. A., D. A. Atacho, F. Gesbert, G. Raposo, I. Hurbain, L. Larue, E. Steingrimsson & P. H. Petersen (2015) Meningeal melanocytes in the mouse: distribution and dependence on Mitf. *Frontiers in Neuroanatomy,* 9**,** 149.

Guo, J. B., Y. Zhu, B. L. Chen, G. Song, M. S. Peng, H. Y. Hu, Y. L. Zheng, C. C. Chen, J. Z. Yang, P. J. Chen & X. Q. Wang (2019) Network and pathway-based analysis of microRNA role in neuropathic pain in rat models. *J Cell Mol Med,* 23**,** 4534-4544.

Gwon, S.-Y., K.-J. Rhee & H. J. Sung (2018) Gene and protein expression profiles in a mouse model of collagen-induced arthritis. *International journal of medical sciences,* 15**,** 77.

Han, S., T. Kwak, K. Her, Y. Cho, C. Choi, H. J. Lee, S. Hong, Y. Park, Y. Kim & T. Kim (2008) CEACAM5 and CEACAM6 are major target genes for Smad3-mediated TGF-β signaling. *Oncogene,* 27**,** 675-683.

Harris, T. M., P. Du, N. Kawachi, T. J. Belbin, Y. Wang, N. F. Schlecht, T. J. Ow, C. E. Keller, G. J. Childs, R. V. Smith, R. H. Angeletti, M. B. Prystowsky & J. Lim (2015) Proteomic analysis of oral cavity squamous cell carcinoma specimens identifies patient outcome-associated proteins. *Arch Pathol Lab Med,* 139**,** 494-507.

Haslinger, A., T. J. Schwarz, M. Covic & D. Chichung Lie (2009) Expression of Sox11 in adult neurogenic niches suggests a stage‐specific role in adult neurogenesis. *European Journal of Neuroscience,* 29**,** 2103-2114.

Hawrylycz, M., L. Ng, D. Feng, S. Sunkin, A. Szafer & C. Dang (2014a) The allen brain atlas. *Springer Handbook of Bio-/Neuroinformatics***,** 1111-1126.

Hawrylycz, M., L. Ng, D. Feng, S. M. Sunkin, A. Szafer & C. Dang. 2014b. The Allen Brain Atlas. 1111-1126.

He, Y., X. Tang, J. Huang, J. Ren, H. Zhou, K. Chen, A. Liu, H. Shi, Z. Lin & Q. Li (2021) ClusterMap for multi-scale clustering analysis of spatial gene expression. *Nature communications,* 12**,** 1-13.

Hong, G., W. Zhang, H. Li, X. Shen & Z. Guo (2014) Separate enrichment analysis of pathways for up-and downregulated genes. *Journal of the Royal Society Interface,* 11**,** 20130950.

Hu, J., X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara & M. Li (2021) SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods,* 18**,** 1342-1351.

Huang, C. K., L. J. Zhan, Y. X. Ai & J. Jongstra (1997) LSP1 is the major substrate for mitogen-activated protein kinase-activated protein kinase 2 in human neutrophils. *Journal Of Biological Chemistry,* 272**,** 17-19.

Huang, D. W., B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane & R. A. Lempicki (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research,* 35**,** W169-W175.

Hubel, K. A. (1985) Intestinal nerves and ion transport: stimuli, reflexes, and responses. *Am J Physiol,* 248**,** G261-71.

Hwang, B., J. H. Lee & D. Bang (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine,* 50**,** 1-14.

Janecek, A. & Y. Tan (2011) ICNC - Iterative improvement of the Multiplicative Update NMF algorithm using nature-inspired optimization. *2011 Seventh International Conference on Natural Computation,* 3**,** 1668-1672.

Janesick, A., R. Shelansky, A. Gottscho, F. Wagner, M. Rouault, G. Beliakoff, M. F. de Oliveira, A. Kohlway, J. Abousoud & C. Morrison (2022) High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. *bioRxiv*.

Jew, B., M. Alvarez, E. Rahmani, Z. Miao, A. Ko, K. M. Garske, J. H. Sul, K. H. Pietiläinen, P. Pajukanta & E. Halperin (2020) Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications,* 11**,** 1971-1971.

Jian, L., J. Xie, S. Guo, H. Yu, R. Chen, K. Tao, C. Yang, K. Li & S. Liu (2020) AGR3 promotes estrogen receptor-positive breast cancer cell proliferation in an estrogen-dependent manner. *Oncology Letters,* 20**,** 1441-1451.

Jiang, Z. & R. Gentleman (2007) Extensions to gene set enrichment. *Bioinformatics,* 23**,** 306-13.

Johnson, J., R. T. Fremeau, J. L. Duncan, R. C. Rentería, H. Yang, Z. Hua, X. Liu, M. M. LaVail, R. H. Edwards & D. R. Copenhagen (2007) Vesicular glutamate transporter 1 is required for photoreceptor synaptic signaling but not for intrinsic visual functions. *Journal of Neuroscience,* 27**,** 7245-7255.

Joshi-Tope, G., M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney & L. Stein (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research,* 33**,** D428-D432.

Junya, P., B.-F. Sun, C. Chen, J.-Y. Zhou, Y.-S. Chen, C. Hao, L. Lulu, H. Dan, J. Jiang, G. Cui, Y. Yang, W. Wang, D. Guo, M. Dai, J. Guo, T. Zhang, Q. Liao, Y. Liu, Y.-L. Zhao, D. Han, Y. Zhao, Y.-G. Yang & W. Wu (2019) Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell research,* 29**,** 725-738.

Kanehisa, M. & S. Goto (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research,* 28**,** 27-30.

Kharchenko, P. V., L. Silberstein & D. T. Scadden (2014) Bayesian approach to single-cell differential expression analysis. *Nature Methods,* 11**,** 740-742.

Khatri, P., M. Sirota & A. J. Butte (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *Plos Computational Biology,* 8.

Kim, S. H. & W. Whitt (2015) The Power of Alternative Kolmogorov-Smirnov Tests Based on Transformations of the Data. *Acm Transactions on Modeling And Computer Simulation,* 25.

Kim, S. Y. & D. J. Volsky (2005) PAGE: Parametric analysis of gene set enrichment. *Bmc Bioinformatics,* 6**,** 144.

Kleshchevnikov, V., A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian & A. Gayoso (2022) Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology***,** 1-11.

Korsunsky, I., A. Nathan, N. Millard & S. Raychaudhuri (2019) Presto scales Wilcoxon and auROC analyses to millions of observations. *BioRxiv***,** 653253.

Lähnemann, D., J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel & A. Mahfouz (2020) Eleven grand challenges in single-cell data science. *Genome biology,* 21**,** 1-35.

Lambrechts, D., E. Wauters, B. Boeckx, S. Aibar, D. Nittner, O. T. Burton, A. Bassez, H. Decaluwé, A. Pircher, K. Van den Eynde, B. Weynand, E. Verbeken, P. De Leyn, A. Liston, J. Vansteenkiste, P. Carmeliet, S. Aerts & B. Thienpont (2018) Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature medicine,* 24**,** 1277-1289.

Lee, D. (2011) A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology,* 2**,** 79-89.

Lee, D. & H. S. Seung (2000) Algorithms for non-negative matrix factorization. *Advances in neural information processing systems,* 13.

Lee, D. D., H. S. Seung & L. K. Saul (2002) Multiplicative updates for unsupervised and contrastive learning in vision. *Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies,* 1**,** 91.

Lee, J. H., E. R. Daugharthy, J. Scheiman, R. Kalhor, T. C. Ferrante, R. Terry, B. M. Turczyk, J. L. Yang, H. S. Lee & J. Aach (2015a) Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature protocols,* 10**,** 442-458.

Lee, J. Y., C. N. Skon, Y. J. Lee, S. Oh, J. J. Taylor, D. Malhotra, M. K. Jenkins, M. G. Rosenfeld, K. A. Hogquist & S. C. Jameson (2015b) The Transcription Factor KLF2 Restrains CD4(+) T Follicular Helper Cell Differentiation. *Immunity,* 42**,** 252-264.

Lehtiniemi, T. & N. Kotaja. 2017. The genetics of postmeiotic male germ cell differentiation from round spermatids to mature sperm. In *Genetics of Human Infertility*, 101-115. Karger Publishers.

Li, B., W. Zhang, C. Guo, H. Xu, L. Li, M. Fang, Y. Hu, X. Zhang, X. Yao & M. Tang (2022) Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature Methods***,** 1-9.

Li, C., F. Su, L. Zhang, F. Liu, W. Fan, Z. Li & J. Ma (2021) Identifying potential diagnostic genes for diabetic nephropathy based on hypoxia and immune status. *Journal of Inflammation Research,* 14**,** 6871.

Li, H., C. A. Calder & N. A. Cressie (2007) Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. *Geographical Analysis,* 39**,** 357-375.

Li, H., J. Zhou, Z. Li, S. Chen, X. Liao, B. Zhang, R. Zhang, Y. Wang, S. Sun & X. Gao (2023) A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics. *Nature Communications,* 14**,** 1548.

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

Li, Z. & X. Zhou (2022) BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome biology,* 23**,** 1-35.

Liao, J., X. Lu, X. Shao, L. Zhu & X. Fan (2020) Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics. *Trends in biotechnology,* 39**,** 43-58.

Lin, S., L. Fang, G. E. Liu & C.-J. Li. 2019. Epigenetics and heritable phenotypic variations in livestock. In *Transgenerational Epigenetics,* 283-313. Elsevier.

Linn, E., L. Ghanem, H. Bhakta, C. Greer & M. Avella (2021) Genes regulating spermatogenesis and sperm function associated with rare disorders. *Frontiers in Cell and Developmental Biology,* 9**,** 634536.

Liu, W., X. Liao, Z. Luo, Y. Yang, M. C. Lau, Y. Jiao, X. Shi, W. Zhai, H. Ji & J. Yeong (2023) Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with PRECAST. *Nature Communications,* 14**,** 296.

Liu, X. & S. Liu (2018) Cholecystokinin selectively activates short axon cells to enhance inhibition of olfactory bulb output neurons. *The Journal of Physiology,* 596**,** 2185-2207.

Lopez, R., B. Li, H. Keren-Shaul, P. Boyeau, M. Kedmi, D. Pilzer, A. Jelinski, E. David, A. Wagner & Y. Addad (2021) Multi-resolution deconvolution of spatial transcriptomics data reveals continuous patterns of inflammation. *BioRxiv*.

Lopez, R., B. Li, H. Keren-Shaul, P. Boyeau, M. Kedmi, D. Pilzer, A. Jelinski, I. Yofe, E. David & A. Wagner (2022) DestVI identifies continuums of cell types in spatial transcriptomics data. *Nature biotechnology,* 40**,** 1360-1369.

Lopez, R., A. Nazaret, M. Langevin, J. Samaran, J. Regier, M. I. Jordan & N. Yosef (2019) A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv preprint arXiv:1905.02269*.

Louis, T. A. (1982) Finding the Observed Information Matrix When Using the Em Algorithm. *Journal Of the Royal Statistical Society Series B-Methodological,* 44**,** 226-233.

Love, M. I., W. Huber & S. Anders (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15**,** 550.

Lowe, R., N. Shirley, M. Bleackley, S. Dolan & T. Shafee (2017) Transcriptomics technologies. *PLoS computational biology,* 13**,** e1005457.

Lubeck, E., A. F. Coskun, T. Zhiyentayev, M. Ahmad & L. Cai (2014) Single-cell in situ RNA profiling by sequential hybridization. *Nature methods,* 11**,** 360-361.

Lundholm, M., S. Mayans, V. Motta, A. Lofgren-Burstrom, J. Danska & D. Holmberg (2010) Variation in the Cd3 zeta (Cd247) Gene Correlates with Altered T Cell Activation and Is Associated with Autoimmune Diabetes. *Journal Of Immunology,* 184**,** 5537-5544.

Lv, Y., D. Lv, X. Lv, P. Xing, J. Zhang & Y. Zhang (2021) Immune cell infiltration-based characterization of triple-negative breast cancer predicts prognosis and chemotherapy response markers. *Frontiers in Genetics,* 12**,** 616469.

Lv, Z., X. Wu, W. Cao, Z. Shen, L. Wang, F. Xie, J. Zhang, T. Ji, M. Yan & W. Chen (2014) Parathyroid hormone-related protein serves as a prognostic indicator in oral squamous cell carcinoma. *J Exp Clin Cancer Res,* 33**,** 100.

Ma, Y., S. Sun, X. Shang, E. T. Keller, M. Chen & X. Zhou (2020) Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. *Nature Communications,* 11**,** 1585.

Ma, Y. & X. Zhou (2022) Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology*.

Mages, S., N. Moriel, I. Avraham-Davidi, E. Murray, J. Watter, F. Chen, O. Rozenblatt-Rosen, J. Klughammer, A. Regev & M. Nitzan (2023) TACCO unifies annotation transfer and decomposition of cell identities for single-cell and spatial omics. *Nature Biotechnology***,** 1-9.

Makrooni, M. A., D. O'Shea, P. Geeleher & C. Seoighe (2022) Random-effects meta-analysis of effect sizes as a unified framework for gene set analysis. *PLOS Computational Biology,* 18**,** e1010278.

Malcher, A., N. Rozwadowska, T. Stokowy, T. Kolanowski, P. Jedrzejczak, W. Zietkowiak & M. Kurpisz (2013) Potential biomarkers of nonobstructive azoospermia identified in microarray gene expression analysis. *Fertility and sterility,* 100**,** 1686-1694. e7.

Maleki, F., K. Ovens, D. J. Hogan & A. J. Kusalik (2020) Gene set analysis: challenges, opportunities, and future research. *Frontiers in genetics,* 11**,** 654.

Malvaut, S., S. Gribaudo, D. Hardy, L. S. David, L. Daroles, S. Labrecque, M.-A. Lebel-Cormier, Z. Chaker, D. Coté & P. De Koninck (2017) CaMKIIα expression defines two functionally distinct populations of granule cells involved in different types of odor behavior. *Current Biology,* 27**,** 3315-3329. e6.

Mathys, H., J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, A. J. Martorell, R. M. Ransohoff, B. P. Hafler, D. A. Bennett, M. Kellis & L.-H. Tsai (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature,* 570**,** 332-337.

Maynard, K. R., L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, J. Chew, Y. Yin, J. E. Kleinman, T. M. Hyde, N. Rao, S. C. Hicks, K. Martinowich & A. E. Jaffe (2021) Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience,* 24**,** 425-436.

Mazzoccoli, G., S. Castellana, M. Carella, O. Palumbo, C. Tiberio, C. Fusilli, D. Capocefalo, T. Biagini, T. Mazza & L. Lo Muzio (2017) A primary tumor gene expression signature identifies a crucial role played by tumor stroma myofibroblasts in lymph node involvement in oral squamous cell carcinoma. *Oncotarget,* 8**,** 104913-104927.

Mentlein, R. & J. Held‐Feindt (2002) Pleiotrophin, an angiogenic and mitogenic growth factor, is expressed in human gliomas. *Journal of neurochemistry,* 83**,** 747-753.

Meyer, G. (2010) Building a human cortex: the evolutionary differentiation of Cajal‐Retzius cells and the cortical hem. *Journal of anatomy,* 217**,** 334-343.

Missarova, A., L. U. Rosen, E. Dann, R. Satija & J. Marioni (2023) Sensitive cluster-free differential expression testing. *bioRxiv***,** 2023.03. 08.531744.

Mizrak, D., H. M. Levitin, A. C. Delgado, V. Crotet, J. Yuan, Z. Chaker, V. Silva-Vargas, P. A. Sims & F. Doetsch (2019) Single-cell analysis of regional differences in adult V-SVZ neural stem cell lineages. *Cell reports,* 26**,** 394-406. e5.

Moen, E., D. Bannon, T. Kudo, W. Graf, M. W. Covert & D. Van Valen (2019) Deep learning for cellular image analysis. *Nature methods,* 16**,** 1233-1246.

Moffitt, J. R., D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac & X. Zhuang (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science (New York, N.Y.),* 362**,** NA-NA.

Moncada, R., D. Barkley, F. Wagner, M. Chiodin, J. C. Devlin, M. Baron, C. H. Hajdu, D. M. Simeone & I. Yanai (2020) Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature biotechnology,* 38**,** 333-342.

Moriel, N., E. Senel, N. Friedman, N. Rajewsky, N. Karaiskos & M. Nitzan (2021) NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols,* 16**,** 4177-4200.

Morvaridi, S., D. Dhall, M. I. Greene, S. J. Pandol & Q. Wang (2015) Role of YAP and TAZ in pancreatic ductal adenocarcinoma and in stellate cells associated with cancer and chronic pancreatitis. *Scientific reports,* 5**,** 16759-16759.

Moses, L. & L. Pachter (2022) Museum of spatial transcriptomics. *Nature Methods*.

Mou, T., W. Deng, F. Gu, Y. Pawitan & T. N. Vu (2020) Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. *Frontiers in genetics,* 10**,** 1331.

Mukherjee, D. & J. Zhao (2013) The role of chemokine receptor CXCR4 in breast cancer metastasis. *American journal of cancer research,* 3**,** 46.

Mullen, A. C. & J. L. Wrana (2017) TGF-beta Family Signaling in Embryonic and Somatic Stem-Cell Renewal and Differentiation. *Cold Spring Harbor Perspectives In Biology,* 9.

Nabi, S., M. Askari, M. Rezaei-Gazik, N. Salehi, N. Almadani, Y. Tahamtani & M. Totonchi (2022) A rare frameshift mutation in SYCP1 is associated with human male infertility. *Molecular Human Reproduction,* 28**,** gaac009.

Nagayama, S., R. Homma & F. Imamura (2014) Neuronal organization of olfactory bulb circuits. *Frontiers in neural circuits,* 8**,** 98.

Narisetty, N. N. & X. He (2014) Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics,* 42**,** 789-817.

Neufeld, A., L. L. Gao, J. Popp, A. Battle & D. Witten (2022) Inference after latent variable estimation for single-cell RNA sequencing data. *arXiv preprint arXiv:2207.00554*.

Nguyen, A. L. & K. Schindler (2017) Specialize and divide (twice): functions of three aurora kinase homologs in mammalian oocyte meiotic maturation. *Trends in Genetics,* 33**,** 349-363.

Nielsen, M. F. B., M. B. Mortensen & S. Detlefsen (2016) Key players in pancreatic cancer-stroma interaction: Cancer-associated fibroblasts, endothelial and inflammatory cells. *World journal of gastroenterology,* 22**,** 2678-2700.

Nishimura, D. (2001) BioCarta. Biotech Software & Internet Report. *Biotech Software & Internet Report,* 2**,** 117-120.

Nitzan, M., N. Karaiskos, N. Friedman & N. Rajewsky (2019) Gene expression cartography. *Nature,* 576**,** 132-137.

Noureen, N., Z. Ye, Y. Chen, X. Wang & S. Zheng (2022) Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *Elife,* 11**,** e71994.

O'Brien, S. L., A. Fagan, E. J. Fox, R. C. Millikan, A. C. Culhane, D. J. Brennan, A. H. McCann, S. Hegarty, S. Moyna & M. J. Duffy (2007) CENP-F expression is associated with poor prognosis and chromosomal instability in patients with primary breast cancer. *International journal of cancer,* 120**,** 1434-1443.

Oakes, D. (1999) Direct calculation of the information matrix via the EM algorithm. *Journal Of the Royal Statistical Society Series B-Statistical Methodology,* 61**,** 479-482.

Ocklenburg, S., C. Anderson, W. M. Gerding, C. Fraenz, C. Schlüter, P. Friedrich, M. Raane, B. Mädler, L. Schlaffke & L. Arning (2019) Myelin water fraction imaging reveals hemispheric asymmetries in human white matter that are associated with genetic variation in PLP1. *Molecular neurobiology,* 56**,** 3999-4012.

Oikonomopoulou, K., E. P. Diamandis & M. D. Hollenberg (2010) Kallikrein-related peptidases: proteolysis and signaling in cancer, the new frontier. *Biological Chemistry,* 391**,** 299-310.

Oron, A. P., Z. Jiang & R. Gentleman (2008) Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics,* 24**,** 2586-2591.

Palla, G., H. Spitzer, M. Klein, D. Fischer, A. C. Schaar, L. B. Kuemmerle, S. Rybakov, I. L. Ibarra, O. Holmberg & I. Virshup (2022) Squidpy: a scalable framework for spatial omics analysis. *Nature methods,* 19**,** 171-178.

Palmer, C., M. Diehn, A. A. Alizadeh & P. O. Brown (2006) Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *Bmc Genomics,* 7**,** 115.

Papalexi, E. & R. Satija (2018) Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology,* 18**,** 35-45.

Pardo, B., A. Spangler, L. M. Weber, S. C. Page, S. C. Hicks, A. E. Jaffe, K. Martinowich, K. R. Maynard & L. Collado-Torres (2022) spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. *BMC genomics,* 23**,** 434.

Park, J.-S. & S.-J. Oh (2012) A new concave hull algorithm and concaveness measure for n-dimensional datasets. *Journal of Information science and engineering,* 28**,** 587-600.

Pauklin, S., P. Madrigal, A. Bertero & L. Vallier (2016) Initiation of stem cell differentiation involves cell cycle-dependent regulation of developmental genes by Cyclin D. *Genes & Development,* 30**,** 421-433.

Pham, D., X. Tan, J. Xu, L. F. Grice, P. Y. Lam, A. Raghubar, J. Vukovic, M. J. Ruitenberg & Q. Nguyen (2020) stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv*.

Phillips, D. J., M. Matusiak, B. R. Gutierrez, S. S. Bhate, G. L. Barlow, S. Jiang, J. Demeter, K. S. Smythe, R. H. Pierce, S. P. Fling, N. Ramchurren, M. A. Cheever, Y. Goltsev, R. West, M. S. Khodadoust, Y. H. Kim, C. M. Schürch & G. P. Nolan (2021) Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. *Nature communications,* 12**,** 6726-NA.

Pickup, M. W., J. K. Mouw & V. M. Weaver (2014) The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep,* 15**,** 1243-53.

Pont, F., M. Tosolini & J. J. Fournié (2019) Single-Cell Signature Explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. *Nucleic acids research,* 47**,** e133-e133.

Prinz, M., J. Priller, S. S. Sisodia & R. M. Ransohoff (2011) Heterogeneity of CNS myeloid cells and their roles in neurodegeneration. *Nature neuroscience,* 14**,** 1227-1235.

ProteinAtlas.

Qiu, C., J. Cao, B. K. Martin, T. Li, I. C. Welsh, S. Srivatsan, X. Huang, D. Calderon, W. S. Noble & C. M. Disteche (2022) Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nature genetics,* 54**,** 328-341.

Radeloff, V. C., T. Miller, H. S. He & D. J. Mladenoff (2000) Periodicity in spatial data and geostatistical models: autocorrelation between patches. *Ecography,* 23**,** 81-91.

Ralston, A. & K. Shaw (2008) Gene expression regulates cell differentiation. *Nat Educ,* 1**,** 127-131.

Ramachandran, P., K. P. Matchett, R. Dobie, J. R. Wilson-Kanamori & N. C. Henderson (2020) Single-cell technologies in hepatology: new insights into liver biology and disease pathogenesis. *Nature reviews Gastroenterology & hepatology,* 17**,** 457-472.

Ramachandran, V. S. 2002. Encyclopedia of the human brain. Academic Press.

Rao, A., D. Barkley, G. S. França & I. Yanai (2021) Exploring tissue architecture using spatial transcriptomics. *Nature,* 596**,** 211-220.

Ren, S., X. Chen, X. Tian, D. Yang, Y. Dong, F. Chen & X. Fang (2021) The expression, function, and utilization of Protamine1: a literature review. *Translational Cancer Research,* 10**,** 4947.

Robinson, M. D., D. J. McCarthy & G. K. Smyth (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-140.

Rodriques, S. G., R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen & E. Z. Macosko (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science,* 363**,** 1463-1467.

Rőszer, T. (2015) Understanding the mysterious M2 macrophage through activation markers and effector mechanisms. *Mediators of inflammation,* 2015.

Rousset, F. & J.-B. Ferdy (2014) Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography,* 37**,** 781-790.

Rue, H. & L. Held. 2005. *Gaussian Markov Random Fields*.

Salahshourifar, I., V. K. Vincent-Chong, H. Y. Chang, H. L. Ser, A. Ramanathan, T. G. Kallarakkal, Z. A. Rahman, S. M. Ismail, N. Prepageran, W. M. Mustafa, M. T. Abraham, K. K. Tay & R. B. Zain (2015) Downregulation of CRNN gene and genomic instability at 1q21.3 in oral squamous cell carcinoma. *Clin Oral Investig,* 19**,** 2273-83.

Saliba, A.-E., A. J. Westermann, S. A. Gorski & J. Vogel (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic acids research,* 42**,** 8845-8860.

Sampaio-Baptista, C., A. Vallès, A. A. Khrapitchev, G. Akkermans, A. M. Winkler, S. Foxley, N. R. Sibson, M. Roberts, K. Miller & M. E. Diamond (2020) White matter structure and myelin-related gene expression alterations with experience in adult rats. *Progress in neurobiology,* 187**,** 101770.

Saunders, A., E. Z. Macosko, A. Wysoker, M. Goldman, F. M. Krienen, H. de Rivera, E. Bien, M. A. Baum, L. Bortolin, S. Wang, A. Goeva, J. Nemesh, N. Kamitaki, S. A. Brumbaugh, D. Kulp & S. A. McCarroll (2018) Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell,* 174**,** 1015-1030.

Schaefer, C. F., K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay & K. H. Buetow (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research,* 37**,** D674-D679.

Schürch, C. M., S. S. Bhate, G. L. Barlow, D. J. Phillips, L. Noti, I. Zlobec, P. Chu, S. Black, J. Demeter, D. R. McIlwain, S. Kinoshita, N. Samusik, Y. Goltsev & G. P. Nolan (2020) Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell,* 182**,** 1341-1359.e19.

Sergushichev, A. (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv***,** 060012.

Shah, M. Y. & A. R. Mehta (2009) Metastasis from breast cancer presenting as an epulis in the upper gingiva. *J Oral Maxillofac Pathol,* 13**,** 38-40.

Shang, L. & X. Zhou (2022) Spatially aware dimension reduction for spatial transcriptomics. *Nature Communications,* 13**,** 7203.

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

Soldatov, R. A., M. Kaucka, M. E. Kastriti, J. Petersen, T. Chontorotzea, L. Englmaier, N. Akkuratova, Y. Yang, M. Häring, V. Dyachuk, C. Bock, M. Farlik, M. L. Piacentino, F. Boismoreau, M. M. Hilscher, C. Yokota, X. Qian, M. Nilsson, M. E. Bronner, L. Croci, W. Y. Hsiao, D. A. Guertin, J.-F. Brunet, G. G. Consalez, P. Ernfors, K. Fried, P. V. Kharchenko & I. Adameyko (2019) Spatiotemporal structure of cell fate decisions in murine neural crest. *Science (New York, N.Y.),* 364**,** NA-NA.

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

Song, H. H., J. Suehiro, Y. Kanki, Y. Kawai, K. Inoue, H. Daida, K. Yano, T. Ohhashi, P. Oettgen, W. C. Aird, T. Kodama & T. Minami (2009) Critical Role for GATA3 in Mediating Tie2 Expression and Function in Large Vessel Endothelial Cells. *Journal Of Biological Chemistry,* 284**,** 29109-29124.

Song, Q. & J. Su (2021) DSTG: deconvoluting spatial transcriptomics data through graph-based artificial intelligence. *Briefings in bioinformatics,* 22**,** NA-NA.

Soysal, S. D., S. Muenst, T. Barbie, T. Fleming, F. Gao, G. Spizzo, D. Oertli, C. T. Viehl, E. C. Obermann & W. E. Gillanders (2013) EpCAM expression varies significantly and is differentially associated with prognosis in the luminal B HER2+, basal-like, and HER2 intrinsic subtypes of breast cancer. *British journal of cancer,* 108**,** 1480-1487.

Spall, J. C. (2005) Monte Carlo computation of the Fisher information matrix in nonstandard settings. *Journal Of Computational And Graphical Statistics,* 14**,** 889-909.

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

Ståhl, P. L., F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm & M. Huss (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science,* 353**,** 78-82.

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

Stoltzfus, C. R., J. Filipek, B. H. Gern, B. E. Olin, J. M. Leal, Y. Wu, M. R. Lyons-Cohen, J. Y. Huang, C. L. Paz-Stoltzfus & C. R. Plumlee (2020) CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. *Cell reports,* 31**,** 107523.

Su, K., L. Huang, W. Li, X. Yan, X. Li, Z. Zhang, F. Jin, J. Lei, G. Ba & B. Liu (2013) TC-1 (c8orf4) enhances aggressive biologic behavior in lung cancer through the Wnt/β-catenin pathway. *journal of surgical research,* 185**,** 255-263.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander & J. P. Mesirov (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings Of the National Academy Of Sciences Of the United States Of America,* 102**,** 15545-15550.

Sun, D., Z. Liu, T. Li, Q. Wu & C. Wang (2022) STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing. *Nucleic Acids Research,* 50**,** e42-e42.

Sun, E., R. Ma & J. Zou (2023) TISSUE: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses. *bioRxiv***,** 2023.04. 25.538326.

Sun, S., M. M. Hood, L. J. Scott, Q. Peng, S. Mukherjee, J. Tung & X. Zhou (2017) Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic acids research,* 45**,** e106-e106.

Sun, S., J. Zhu, S. V. Mozaffari, C. Ober, M. Chen & X. Zhou (2018) Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. *Bioinformatics (Oxford, England),* 35**,** 487-496.

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

Sun, X., X. Liu, E. R. Starr & S. Liu (2020b) CCKergic tufted cells differentially drive two anatomically segregated inhibitory circuits in the mouse olfactory bulb. *Journal of Neuroscience,* 40**,** 6189-6206.

Svensson, V., S. A. Teichmann & O. Stegle (2018) SpatialDE: identification of spatially variable genes. *Nature methods,* 15**,** 343-346.

Swanson, O. K. & A. Maffei (2019) From hiring to firing: activation of inhibitory neurons and their recruitment in behavior. *Frontiers in molecular neuroscience,* 12**,** 168.

Tabata, H. (2015) Diverse subtypes of astrocytes and their development during corticogenesis. *Frontiers in neuroscience,* 9**,** 114.

Tan, L., Q. Li & X. S. Xie (2015) Olfactory sensory neurons transiently express multiple olfactory receptors during development. *Molecular systems biology,* 11**,** 844.

Tang, H., A. Kung & E. Goldberg (2008) Regulation of murine lactate dehydrogenase C (Ldhc) gene expression. *Biology of reproduction,* 78**,** 455-461.

Tang, Q., C. L. Ebbesen, J. I. Sanguinetti-Scheck, P. Preston-Ferrer, A. Gundlfinger, J. Winterer, P. Beed, S. Ray, R. Naumann & D. Schmitz (2015) Anatomical organization and spatiotemporal firing patterns of layer 3 neurons in the rat medial entorhinal cortex. *Journal of Neuroscience,* 35**,** 12346-12354.

Teo, A. K. K., S. J. Arnold, M. W. B. Trotter, S. Brown, L. T. Ang, Z. Z. Chng, E. J. Robertson, N. R. Dunn & L. Vallier (2011) Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes & Development,* 25**,** 238-250.

Tepe, B., M. C. Hill, B. T. Pekarek, P. J. Hunt, T. J. Martin, J. F. Martin & B. R. Arenkiel (2018) Single-cell RNA-seq of mouse olfactory bulb reveals cellular heterogeneity and activity-dependent molecular census of adult-born neurons. *Cell reports,* 25**,** 2689-2703. e3.

Terra, R., H. Y. Luo, X. Y. Qiao & J. P. Wu (2008) Tissue-specific expression of B-cell translocation gene 2 (BTG2) and its function in T-cell immune responses in a transgenic mouse model. *International Immunology,* 20**,** 317-326.

Teschendorff, A. E., T. Zhu, C. E. Breeze & S. Beck (2020) EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. *Genome biology,* 21**,** 1-33.

Tian, L., F. Chen & E. Z. Macosko (2022) The expanding vistas of spatial transcriptomics. *Nature Biotechnology***,** 1-10.

Toyoda, Y. & T. Ishikawa (2010) Pharmacogenomics of human ABC transporter ABCC11 (MRP8): potential risk of breast cancer and chemotherapy failure. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents),* 10**,** 617-624.

Tuch, B. B., R. R. Laborde, X. Xu, J. Gu, C. B. Chung, C. K. Monighetti, S. J. Stanley, K. D. Olsen, J. L. Kasperbauer, E. J. Moore, A. J. Broomer, R. Tan, P. M. Brzoska, M. W. Muller, A. S. Siddiqui, Y. W. Asmann, Y. Sun, S. Kuersten, M. A. Barker, F. M. De La Vega & D. I. Smith (2010) Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One,* 5**,** e9317.

Tufro, A., V. F. Norwood, R. M. Carey & R. A. Gomez (1999) Vascular endothelial growth factor induces nephrogenesis and vasculogenesis. *Journal Of the American Society Of Nephrology,* 10**,** 2125-2134.

Uchida, Y., S.-i. Nakano, F. Gomi & H. Takahashi (2011) Up-regulation of calsyntenin-3 by β-amyloid increases vulnerability of cortical neurons. *FEBS letters,* 585**,** 651-656.

Uraguchi, M., M. Morikawa, M. Shirakawa, K. Sanada & K. Imai (2004) Activation of WNT family expression and signaling in squamous cell carcinomas of the oral cavity. *J Dent Res,* 83**,** 327-32.

Usoskin, D., A. Furlan, S. Islam, H. Abdo, P. Lonnerberg, D. Lou, J. Hjerling-Leffler, J. Haeggstrom, O. Kharchenko, P. V. Kharchenko, S. Linnarsson & P. Ernfors (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience,* 18**,** 145-153.

Van den Berge, K., F. Perraudeau, C. Soneson, M. I. Love, D. Risso, J. P. Vert, M. D. Robinson, S. Dudoit & L. Clement (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biology,* 19**,** 24.

van Vliet, S., A. Dal Co, A. R. Winkler, S. Spriewald, B. Stecher & M. Ackermann (2018) Spatially Correlated Gene Expression in Bacterial Groups: The Role of Lineage History, Spatial Gradients, and Cell-Cell Interactions. *Cell systems,* 6**,** 496-507.e6.

Vandenbon, A. & D. Diez (2020) A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature communications,* 11**,** 4318.

Vanhatalo, J., V. Pietiläinen & A. Vehtari (2010) Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine,* NA**,** n/a-n/a.

Vellame, D. S., G. Shireby, A. MacCalman, E. L. Dempster, J. Burrage, T. Gorrie-Stone, L. S. Schalkwyk, J. Mill & E. Hannon (2023) Uncertainty quantification of reference-based cellular deconvolution algorithms. *Epigenetics,* 18**,** 2137659.

Vickovic, S., G. Eraslan, F. Salmén, J. Klughammer, L. Stenbeck, D. Schapiro, T. Äijö, R. Bonneau, L. Bergenstråhle & J. F. Navarro (2019) High-definition spatial transcriptomics for in situ tissue profiling. *Nature methods,* 16**,** 987-990.

Vizgen. 2021. Vizgen MERFISH Mouse Brain Receptor Map.

Vokes, S. A. & P. A. Krieg (2002) Endoderm is required for vascular endothelial tube formation, but not for angioblast specification. *Development,* 129**,** 775-785.

Von Ahrens, D., T. D. Bhagat, D. Nagrath, A. Maitra & A. Verma (2017) The role of stromal cancer-associated fibroblasts in pancreatic cancer. *Journal of hematology & oncology,* 10**,** 1-8.

Wälchli, T., J. M. Mateos, O. Weinman, D. Babic, L. Regli, S. P. Hoerstrup, H. Gerhardt, M. E. Schwab & J. Vogel (2014) Quantitative assessment of angiogenesis, perfused blood vessels and endothelial tip cells in the postnatal mouse brain. *Nature protocols,* 10**,** 53-74.

Wang, D., J. Wu, P. Liu, X. Li, J. Li, M. He & A. Li (2022a) VIP interneurons regulate olfactory bulb output and contribute to odor detection and discrimination. *Cell Reports,* 38**,** 110383.

Wang, G., J. Zhao, Y. Yan, Y. Wang, A. R. Wu & C. Yang (2023) Construction of a 3D whole organism spatial atlas by joint modeling of multiple slices. *BioRxiv***,** 2023.02. 02.526814.

Wang, X., W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, K. Evans, C. Liu, C. Ramakrishnan & J. Liu (2018a) Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science,* 361**,** eaat5691.

Wang, X., J. Park, K. Susztak, N. R. Zhang & M. Li (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications,* 10**,** 1-9.

Wang, X., M. Sang, S. Gong, Z. Chen, X. Zhao, G. Wang, Z. Li, Y. Huang, S. Chen & G. Xie (2022b) BET bromodomain inhibitor JQ1 regulates spermatid development by changing chromatin conformation in mouse spermatogenesis. *Genes & Diseases,* 9**,** 1062-1073.

Wang, Y. S., C. Cho, J. Williams, P. M. Smallwood, C. Zhang, H. J. Junge & J. Nathans (2018b) Interplay of the Norrin and Wnt7a/Wnt7b signaling systems in blood-brain barrier and blood-retina barrier development and maintenance. *Proceedings Of the National Academy Of Sciences Of the United States Of America,* 115**,** E11827-E11836.

Wang, Z., M. Gerstein & M. Snyder (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics,* 10**,** 57-63.

Weber, D., J. Heisig, S. Kneitz, E. Wolf, M. Eilers & M. Gessler (2015) Mechanisms of epigenetic and cell-type specific regulation of Hey target genes in ES cells and cardiomyocytes. *Journal Of Molecular And Cellular Cardiology,* 79**,** 79-88.

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

Wozny, C., P. Beed, N. Nitzan, Y. Poessnecker, B. R. Rost & D. Schmitz (2018) VGLUT2 functions as a differential marker for hippocampal output neurons. *Frontiers in cellular neuroscience,* 12**,** 337-337.

Wu, D. & G. K. Smyth (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research,* 40**,** e133.

Wu, S. Z., G. Al-Eryani, D. L. Roden, S. Junankar, K. Harvey, A. Andersson, A. Thennavan, C. Wang, J. R. Torpy & N. Bartonicek (2021) A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics,* 53**,** 1334-1347.

Xia, C., J. Fan, G. Emanuel, J. Hao & X. Zhuang (2019) Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences of the United States of America,* 116**,** 19490-19499.

Xu, D., F. Yang, K. Wu, X. Xu, K. Zeng, Y. An, F. Xu, J. Xun, X. Lv, X. Zhang, X. Yang & L. Xu (2020) Lost miR-141 and upregulated TM4SF1 expressions associate with poor prognosis of pancreatic cancer: regulation of EMT and angiogenesis by miR-141 and TM4SF1 via AKT. *Cancer biology & therapy,* 21**,** 354-363.

Yan, W., Y. Si, S. Slaymaker, J. Li, H. Zheng, D. L. Young, A. Aslanian, L. Saunders, E. Verdin & I. F. Charo (2010) Zmynd15 encodes a histone deacetylase-dependent transcriptional repressor essential for spermiogenesis and male fertility. *Journal of Biological Chemistry,* 285**,** 31418-31426.

Yang, S. & X. Zhou (2020) Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. *American journal of human genetics,* 106**,** 679-693.

Yang, Z. Q., A. B. Moffa, R. Haddad, K. L. Streicher & S. P. Ethier (2007) Transforming properties of TC‐1 in human breast cancer: Interaction with FGFR2 and β‐catenin signaling pathways. *International journal of cancer,* 121**,** 1265-1273.

Young, J. K., T. Heinbockel & M. C. Gondré‐Lewis (2013) Astrocyte fatty acid binding protein‐7 is a marker for neurogenic niches in the rat hippocampus. *Hippocampus,* 23**,** 1476-1483.

Yu, D. & M.-C. Hung (2000) Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. *Oncogene,* 19**,** 6115-6121.

Yu, J., J. Li, J. Shen, F. Du, X. Wu, M. Li, Y. Chen, C. H. Cho, X. Li & Z. Xiao (2021a) The role of Fibrinogen-like proteins in Cancer. *International journal of biological sciences,* 17**,** 1079.

Yu, X., F. Abbas-Aghababazadeh, Y. A. Chen & B. L. Fridley (2021b) Statistical and bioinformatics analysis of data from bulk and single-cell RNA sequencing experiments. *Translational Bioinformatics for Therapeutic Development***,** 143-175.

Yuan, Y., H. Gao, Y. Zhuang, L. Wei, J. Yu, Z. Zhang, L. Zhang & L. Wang (2021) NDUFA4L2 promotes trastuzumab resistance in HER2-positive breast cancer. *Therapeutic advances in medical oncology,* 13**,** 17588359211027836.

Zeisel, A., H. Hochgerner, P. Lönnerberg, A. Johnsson, F. Memic, J. Van Der Zwan, M. Häring, E. Braun, L. E. Borm & G. La Manno (2018) Molecular architecture of the mouse nervous system. *Cell,* 174**,** 999-1014. e22.

Zeng, H., E. H. Shen, J. G. Hohmann, S. W. Oh, A. Bernard, J. J. Royall, K. J. Glattfelder, S. M. Sunkin, J. A. Morris & A. L. Guillozet-Bongaarts (2012) Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell,* 149**,** 483-496.

Zeppilli, S., T. Ackels, R. Attey, N. Klimpert, K. D. Ritola, S. Boeing, A. Crombach, A. T. Schaefer & A. Fleischmann (2021) Molecular characterization of projection neuron subtypes in the mouse olfactory bulb. *Elife,* 10.

Zhang, J. M., G. M. Kamath & D. Tse (2019) Valid post-clustering differential analysis for single-cell RNA-Seq. *Available at SSRN 3378005*.

Zhang, L., M. Chen, Q. Wen, Y. Li, Y. Wang, Y. Wang, Y. Qin, X. Cui, L. Yang & V. Huff (2015) Reprogramming of Sertoli cells to fetal-like Leydig cells by Wt1 ablation. *Proceedings of the National Academy of Sciences,* 112**,** 4003-4008.

Zhang, X.-f., M. Dong, Y.-h. Pan, J.-n. Chen, X.-q. Huang, Y. Jin & C.-k. Shao (2017) Expression pattern of cancer-associated fibroblast and its clinical relevance in intrahepatic cholangiocarcinoma. *Human pathology,* 65**,** 92-100.

Zhao, E., M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. B. Taylor & P. Nghiem (2021) Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology***,** 1-10.

Zheng, B., K. Ohuchida, L. Cui, M. Zhao, K. Shindo, K. Fujiwara, T. Manabe, N. Torata, T. Moriyama, Y. Miyasaka, T. Ohtsuka, S. Takahata, K. Mizumoto, Y. Oda & M. Tanaka (2015) TM4SF1 as a prognostic marker of pancreatic ductal adenocarcinoma is involved in migration and invasion of cancer cells. *International journal of oncology,* 47**,** 490-498.

Zheng, C., L. Zheng, J.-K. Yoo, H. Guo, Y. Zhang, G. Xinyi, B. Kang, R. Hu, J. Y. Huang, Q. Zhang, Z. Liu, M. Dong, X. Hu, W. Ouyang, J. Peng & Z. Zhang (2017a) Landscape of

Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell,* 169**,** 1342-1356.

Zheng, G. X. Y., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson & J. H. Bielas (2017b) Massively parallel digital transcriptional profiling of single cells. *Nature Communications,* 8**,** 14049

Zhi, X., K. Lamperska, P. Golusinski, N. J. Schork, L. Luczewski, W. Golusinski & M. M. Masternak (2014) Expression levels of insulin-like growth factors 1 and 2 in head and neck squamous cell carcinoma. *Growth Horm IGF Res,* 24**,** 137-41.

Zhou, W., J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford, L. A. Bastarache, W. Q. Wei, J. C. Denny, M. X. Lin, K. Hveem, H. M. Kang, G. R. Abecasis, C. J. Willer & S. Lee (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics,* 50**,** 1335-1341.

Zhou, X., P. Carbonetto & M. Stephens (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics,* 9**,** e1003264.

Zhu, C., S. Preissl & B. Ren (2020) Single-cell multimodal omics: the power of many. *Nature methods,* 17**,** 11-14.

Zhu, Q., S. Shah, R. Dries, L. Cai & G.-C. Yuan (2018) Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature biotechnology,* 36**,** 1183-1190.

Zhu, Y., J. M. Herndon, D. K. Sojka, K.-W. Kim, B. L. Knolhoff, C. Zuo, D. R. Cullinan, J. Luo, A. R. Bearden & K. J. Lavine (2017) Tissue-resident macrophages in pancreatic ductal adenocarcinoma originate from embryonic hematopoiesis and promote tumor progression. *Immunity,* 47**,** 323-338. e6.

Zylka, M. J., X. Dong, A. L. Southwell & D. J. Anderson (2003) Atypical expansion in mice of the sensory neuron-specific Mrg G protein-coupled receptor family. *Proc Natl Acad Sci U S A,* 100**,** 10043-8.