UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title

Statistical and Machine Learning Methods for Understanding Spatial Biology and Cellular Communication

Permalink

https://escholarship.org/uc/item/61s9p7dr

Author

Wang, Yutong

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Statistical and Machine Learning Methods for Understanding Spatial Biology and Cellular Communication

by

Yutong Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Associate Professor Aaron Streets
Professor Sandrine Dudoit

Spring 2024

Statistical and Machine Learning Methods for Understanding Spatial Biology and Cellular Communication

Copyright 2024
by
Yutong Wang

Abstract

Statistical and Machine Learning Methods for Understanding Spatial Biology and Cellular Communication

by

Yutong Wang

Doctor of Philosophy in Biostatistics

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Yun S. Song, Chair

Recent advancements in spatial transcriptomics (ST) and single-cell RNA sequencing (scRNA-seq) technologies have revolutionized our understanding of cellular heterogeneity and spatial organization within tissues. This dissertation presents computational methods and applications for analyzing these data modalities to investigate the tumor microenvironment, cell-cell interactions, and immune cell migration across tissues. Our first work utilizes XYZeq data to analyze transcriptome at a single-cell resolution while maintaining spatial context. This includes applying non-negative matrix factorization to identify spatially variable gene modules and performing trajectory inference to examine gene expression profiles differentially expressed across the cell proximity in the tumor microenvironment. Next, we present ggPair (gene-gene Pair), a new computational method for predicting ligand-receptor pairs using ST data. Our approach diverges from traditional methods by employing a convolutional autoencoder for unsupervised feature extraction followed by a siamese neural network to predict gene-gene interactions, thereby enabling the discovery of previously unrecognized interacting genes. Moreover, focusing on T cell differentiation and migration, we detail our analysis investigating the transcriptional signatures that distinguish immune cell populations in the context of an autoimmune disease model, using scRNA-seq data and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) knock-out data. Our analysis differentiates transcriptional activities of T regulatory cells (Tregs) from T effector cells, providing insights into immune cell dynamics and migration in disease states. In summary, we provide different computational strategies to dissect complex biological data, extending our understanding of cellular heterogeneity, spatial organization, and the cellular communication.

To my mother, Yulin, and my great-grandmother, Ruilian,
two extraordinary women whose unwavering love, sacrifice, and support
have made me who I am and brought me to where I stand today.

Contents

Contents **ii**

List of Figures **v**

List of Tables **ix**

1 Introduction **1**
1.1 Spatial Transcriptomics . . . 1
1.1.1 Sequencing Technologies . . . 1
1.1.2 Existing Computational Methods Development . . . 2
1.2 Cell-Cell Communication . . . 5
1.3 Dissertation Outline . . . 7

2 Characterizing Spatial Heterogeneity with XYZeq **9**
2.1 Introduction . . . 10
2.2 Results . . . 10
2.3 Discussion . . . 18
2.4 Methods . . . 18
2.4.1 XYZeq Methodology . . . 18
2.4.2 XYZeq Decontamination Analysis . . . 19
2.4.3 Distinctions between Collision Rate and Contamination Rate . . . 19
2.4.4 XYZeq Single Cell Analysis . . . 20
2.4.5 10X Data Processing . . . 20
2.4.6 Heatmap for XYZeq . . . 21
2.4.7 XYZeq Gene Pairplot . . . 21
2.4.8 XYZeq Cell/Well Pairplot . . . 21
2.4.9 Heat Map Comparing 10X to XYZeq . . . 21
2.4.10 Gene Module Analysis of Top Contributing Genes . . . 21
2.4.11 Overlapping Analysis between the Gene Modules . . . 22
2.4.12 Defining the Proximity Score by Wells . . . 22
2.4.13 Trajectory Inference Analysis . . . 24

3 Cellular Interaction Inference **30**
3.1 Introduction . . . 31
3.2 Results . . . 33
3.2.1 Model Framework . . . 33
3.2.2 Performance Comparison to Existing Methods . . . 34
3.2.3 Leave-One-Out Validation for GPCRs and nonGPCRs . . . 35
3.2.4 Identification of Positive Ligand Candidates for Orphan GPCR . . . 37
3.3 Discussion . . . 38
3.4 Conclusion . . . 40
3.5 Methods . . . 41
3.5.1 Preprocessing of Spatial Transcriptomics Data . . . 41
3.5.2 Compilation of Ligand-Receptor Interaction Annotations . . . 42
3.5.3 Convolutional Autoencoder (CAE) for Unsupervised Feature Extraction 42
3.5.4 Siamese Neural Network (SNN) for Pairwise Classification . . . 43
3.5.5 Leave-One-Out Strategy and Hypergeometric Testing . . . 43
3.5.6 Implementation Details for Other Methods . . . 44
3.5.7 Tree Calculation . . . 45

4 Identifying T Cell Migration Signatures **46**
4.1 Introduction . . . 46
4.2 Results . . . 48
4.2.1 scRNA-seq Analysis Reveals Transcriptional Signatures . . . 48
4.2.2 Identifying Treg Migration Regulators Using CRISPR Data . . . 52
4.3 Discussion . . . 54
4.4 Methods . . . 56
4.4.1 10X Data Processing . . . 56
4.4.2 Demultiplexing Analysis with Hashtag-Oligo . . . 57
4.4.3 Preprocessing scRNA-seq Data . . . 58
4.4.4 Cell Type Annotation . . . 58
4.4.5 Cell-Type-Specific DE Analysis between Conditions (Tissues) . . . 59
4.4.6 Difference in the Number of DE Genes . . . 61
4.4.7 Interpretation of In Vivo CRISPR Loss-of-Function Screen . . . 61
4.4.8 Preprocessing CRISPR Data . . . 62
4.4.9 Identification of CRISPR Gene Hits . . . 62

Bibliography **64**

A Supplementary Material for Chapter 2 **79**
A.1 Experimental Methods . . . 79
A.2 Supplementary Figures . . . 83
A.3 Supplementary Tables . . . 99

B Supplementary Material for Chapter 3 **101**
B.1 Supplementary Figures 102
B.2 Supplementary Tables 110

C Supplementary Material for Chapter 4 **113**
C.1 Supplementary Figures 113
C.2 Supplementary Table 128

List of Figures

2.1 Schematic of the XYZeq workflow. 11
2.2 XYZeq enables single-cell and spatial transcriptome profiling simultaneously. . . 13
2.3 Spatially resolved single-cell transcriptomes captured from tissue. 14
2.4 Frequency and spatial mapping of single-cell clusters from tissue. 25
2.5 Pie chart of cell type composition for each XYZeq well, and the pairplot showing the frequency of hepatocytes, MC38, and myeloid cells in each well. 26
2.6 Expression of gene modules in space that track with cellular composition. 27
2.7 Heatmap representing the number of overlapping genes between each pair of modules in liver/tumor and spleen/tumor, and the spatial projection of two gene modules. 28
2.8 Differential gene expression within MSCs associated with their spatial proximity to tumor. 28
2.9 Continued figures about the differential gene expression within MSCs associated with their spatial proximity to tumor. 29
2.10 Projection in space of mean expression of Csmd1 (A) and Tshz2 (B) in MSCs. . 29

3.1 ggPair Method Overview. 33
3.2 Model performance comparison using 5-fold cross-validation on the mouse data. 35
3.3 Model performance comparison using 5-fold cross-validation on the human osteosarcoma data. 35
3.4 Model performance comparison using 5-fold cross-validation on the first human breast cancer data. 36
3.5 Histograms showing the distribution of predicted probabilities for true (orange) and false (blue) ligand-receptor interactions on the mouse data. 36
3.6 The average density of the ligand candidates ranked by the predicted class probabilities, stratified by the true interaction label for each gene pair. 37
3.7 Precision at the top k call(s) for GPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B). 37
3.8 Recall at the top k call(s) for GPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B). 38
3.9 The number of ligands with certain number of occurrences in the top k predictions for different orphan GPCR receptors. 39
3.10 Examples from the receptors with known interacting ligands. 40

3.11 Examples from the orphan GPCR receptors. 41

4.1 Schematic of the scRNA-seq experiment workflow. 48
4.2 Dot plots of cluster by marker gene in CNS (A) and SP (B). 49
4.3 Volcano plots highlighting the differentially expressed genes between CNS and SP, among Tregs. 50
4.4 Volcano plots highlighting the differentially expressed genes between CNS and SP, among T effector cells. 51
4.5 GO enrichment terms among Treg. 52
4.6 GO enrichment terms among T effector cells. 53
4.7 Top two PCs of the samples in Tregs (A) and T effector cells (B), colored by the tissue annotation. 54
4.8 Visualization of the positive (red) and negative (blue) regulator genes of Tregs migration from SP to CNS, based on the CRISPR experimental data.log 2FC stands for log 2 fold change, and FDR stands for false discovery rate. 55
4.9 Stratified annotations of the CRISPR gene hits in the volcano plot. 56
4.10 Visualization of the log fold change of sgRNAs targeting top 5 selected genes. . 57

A.1 Single cell mixed species experiment reveals strong correlation to estimated cell gradient proportions. 83
A.2 Quantification of cells captured per well from liver/tumor tissue. 84
A.3 The median cell number in wells across XYZeq array for HEK293T human (A) and liver/tumor mouse (B) cells. 85
A.4 Distinct cell types clusters identified from XYZeq of liver/tumor tissue. 86
A.5 Annotated cell-identity clusters. 87
A.6 Reproducibility of XYZeq across tissue slices. 88
A.7 tSNE visualization of Leiden clusters across four liver/tumor slices. 89
A.8 Cell type clusters captured from XYZeq found comparable to 10X platform. . . 89
A.9 Distinct spatial localization pattern across tissue for each cell type cluster. . . . 90
A.10 Pairplot showing the frequency of cell types found in each well across the XYZeq array. 91
A.11 XYZeq of spleen/tumor tissue reveals comparable data quality to liver/tumor tissue. 92
A.12 Visualization of human and mouse cell distribution on the XYZeq array, and the median cell number in wells across XYZeq array. 93
A.13 Identification and spatial mapping of cell type clusters from spleen/tumor tissue. 94
A.14 Cell type contribution and functional annotation of gene modules. 95
A.15 Cell type contribution and functional annotation of the immune regulation gene modules. 96
A.16 Expression of the cell migration gene module enriched in MSCs. 97
A.17 Tumor proximity score defined for both liver and spleen tissue. 98

B.1 The four spatial transcriptomic batches from mouse brain used in this study. . . 102
B.2 Batch correction of the mouse brain data using the GraphST algorithm aligns the batches. 103
B.3 Spatial gene expression of Itga9 gene in three human spatial transcriptomics data. 103
B.4 Spatial gene expression of 10 randomly selected genes in the spatial batch B02-D1. 103
B.5 Spatial gene expression of 10 randomly selected genes in the spatial batch B02-E1. 104
B.6 Spatial gene expression of 10 randomly selected genes in the spatial batch B03-C2. 104
B.7 Spatial gene expression of 10 randomly selected genes in the spatial batch B03-D2. 105
B.8 CAE Training loss curve on mouse data. 105
B.9 CAE loss curves on training (blue) and validation (yellow) data. 105
B.10 Spatial gene expression of randomly selected genes in the osteosarcoma MERFISH data. 106
B.11 Spatial gene expression of randomly selected genes in the first human breast cancer 10X Visium data. 106
B.12 Spatial gene expression of randomly selected genes in the second human breast cancer 10X Visium data. 107
B.13 Model performance comparison using 5-fold cross-validation on the mouse data. 107
B.14 Model performance comparison using 5-fold cross-validation on the second human breast cancer data. 108
B.15 The average density of the ligand candidates ranked by the predicted class probabilities, stratified by the true interaction label for each gene pair. 108
B.16 Precision at the top k call(s) for nonGPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B). 108
B.17 Recall at the top k call(s) for nonGPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B). 109

C.1 Violin plots across different samples (batches) in the CNS. 113
C.2 Violin plots across different samples (batches) in the SP. 114
C.3 Visualization of the global level 1 (A) and local level 2 (B) clusters for the scRNA-seq data in the CNS. 114
C.4 Visualization of the global level 1 (A) and local level 2 (B) clusters for the scRNA-seq data in the SP. 115
C.5 CNS: each entry in the heatmap corresponds to the percentage of cells in a given level 2 cluster (cluster) that express a specific marker gene (row). 116
C.6 SP: each entry in the heatmap corresponds to the percentage of cells in a given level 2 cluster (cluster) that express a specific marker gene (row). 117
C.7 Visualization of the cell type annotation for the scRNA-seq data in the CNS (A) and SP (B). 118
C.8 Visualization of key marker genes in the CNS scRNA-seq data. 119
C.9 Continued visualization of key marker genes in the CNS scRNA-seq data. 120
C.10 Visualization of key marker genes in the SP scRNA-seq data. 121
C.11 Continued visualization of key marker genes in the SP scRNA-seq data. 122

C.12 Top two PCs of the samples in Tregs (A) and T effector cells (B), colored by the number of single-cell level UMI counts per sample. 122
C.13 Top two PCs of the samples in Tregs (A) and T effector cells (B). 123
C.14 Hierarchical clustering of the samples in Tregs (A) and T effector cells (B). . . . 123
C.15 Dispersion estimates of the samples in Tregs (A) and T effector cells (B). 124
C.16 Hierarchical clustering of the aggregated gene expressions in Treg (A) and T effector cell (B) samples. 124
C.17 Genes related to cytokines, cytokine receptors, and cell surface GO terms are annotated and highlighted among all the DE genes between CNS and SP. 125
C.18 Volcano plots highlighting the differentially expressed genes between CNS and SP, among a random subsample of T effector cells. 126
C.19 The number of total reads, mapped and unmapped reads per CRISPR experiment.127

List of Tables

A.1 Lists of overlapping genes among top 200 contributing genes between liver and spleen, for tumor response modules (LM5) and immune regulation modules (LM13). 99
A.2 Lists of overlapping genes among top 200 contributing genes between liver and spleen, for immune regulation modules (LM19) and cell migration modules (LM10). 100

B.1 Comparison of methods based on AUPRC for GPCR, Non-GPCR, and Overall data. 110
B.2 Comparison of methods based on AUROC for GPCR, Non-GPCR, and Overall data. 111
B.3 Comparison of methods based on Accuracy for GPCR, Non-GPCR, and Overall data. 112

C.1 The list of 120 marker genes for cell type annotation in the CNS and SP. 128

Acknowledgments

I feel exceptionally grateful to have worked with and learnt from my PhD advisor, Yun S. Song. Yun has always been my role model in academia. Not only is Yun intelligent and diligent, but he also possesses a strong passion for science, constantly driven by a curiosity to explore fascinating biological and mathematical questions. Yun taught me how to approach scientific problems, generate research ideas, and think critically and rigorously.

I am grateful to Yun, Aaron Streets, and Sandrine Dudoit for serving on my dissertation committee and providing valuable feedback on my work. I also extend my thanks to Yun, Aaron, Maya Petersen and Lexin Li for serving on my qualifying exam committee and contributing to my academic growth.

Youjin Lee has been one of my close collaborators throughout my PhD, and has taught me so much about immunology. Her contagious enthusiasm for science is evident in every research discussion we have had. I am grateful for her trust and support, which have empowered me to explore the data and seek answers to biological questions.

I would like to express my gratitude to Aaron for teaching me how to effectively present results to readers and for providing invaluable career advice. My sincere thanks also go to my other collaborators, Jimmie Ye, Alexander Marson, and Eric Chow, for their support and guidance in the XYZeq project. I was fortunate to collaborate with fellow graduate students and postdocs on multiple projects, and I have learned a great deal while working with Derek Bogdanoff, George Hartoularos, and Hunter Nisonoff on the XYZeq project.

I am thankful for all the members of the Song Lab for fostering an inclusive and collaborative environment, brainstorming research ideas, and providing feedback on my work. It has been a pleasure working with Fanding Zhou and Antoine Koehl on the ggPair project. I would also like to thank Jeffrey P. Spence, Gonzalo Benegas, Alan Aw, Carlos Albors, Canal Li, Sebastian Prillo, Will DeWitt, Frank Liu, Jane Yu, Nicholas Bhattacharya, Sanjit Batra, Neil Thomas, Jeffrey Chan, Dan Daniel Erdmann-Pham, Jesús Martínez-Gómez, Yun Deng, Milind Jagota, Junhao (Bear) Xiong, Chengzhong Ye, Justin Hong, and Jonathan Fischer for their support and camaraderie.

I am grateful for the opportunity to have worked as a summer intern at the Chan Zuckerberg Biohub during the summer of 2022. Jonathan Liu and Angela Oliveira Pisco were two exceptional mentors, always supportive of my research ideas and professional development. I greatly enjoyed brainstorming ideas and discussing science with researchers from the Data Science - Quantitative Cell Science Platform, the Theory Group, and the Genomes Group. I was fortunate to have informational interviews with brilliant scientists at Biohub, including William Burkholder, Sandra Schmid, Adrian Jacobo, Amy Kistler, Joan Wong, Angela Detweiler, Alejandro Granados, Yang-Joon Kim. It was fascinating to hear about their scientific career journeys and receive advice on professional career development. My gratitude extends to many friends I made during the internship, especially Emanuel Flores Bautista, Pei Fu, and Aditi Basu Bal.

I appreciate all the love and support from the Biostatistics Community at UC Berkeley. I am fortunate to have met and befriended many wonderful colleagues here, including Waverly

Lingqing Wei, Philippe Boileau, James Duncan, Yang Li, and Sylvia Cheng, among many others. I am especially grateful for the weekly Zoom calls with Phil and James, which brought me so much laughter and energy during the pandemic. I am also thankful for the many friends in the Bio + ML community, including Chloe Hsu, Stacy Li, Tiffany Tang, and Valeh Valiollah Pour Amiri. I would like to express my gratitude to our staff in Biostatistics and Computational Biology for their help, including Lauren M Krupa, Xuan Quach, Martin Witte, Kate Chase, and Kris Thompson.

I am grateful to have had the opportunity to work and contribute as the Inaugural Diversity, Equity, and Inclusion (DEI) Student Fellow in Biostatistics. I would not have been able to spearhead numerous initiatives to increase accessibility and representation without the help and support of the DEI faculty advisor Corinne A Riddell, our department chair Alan Hubbard, and the DEI committee members. I am especially thankful for the support from our Biostatistics Division in allowing me to plan and organize the inaugural data science workshop to provide underrepresented minority students with equitable exposure to programming and data analysis. I am grateful for the contributions of Kaitlyn Lee and Maxwell Murphy in coordinating the workshop together, and for the help from Juan Alberdin, Christina Da Silva, and Burke Bundy in Biostatistics.

Graduate school has been a crucial stage for me to explore who I am and the different possibilities of life experiences. Reading The Brothers Karamazov by Fyodor Dostoevsky has taught me to be more grounded in the little scientific progress I make and to care more about the people around me. I appreciate the many enlightening conversations with members of the Wild West Book Club about Russian literature, poems, women's empowerment, and even seemingly useless niche details. I am especially grateful to Luojun Yang, Annie Yang-Perez, LikeRiver, Margaret, Qianhong Zhu, FAtumftRBG, Takiko, and Yiwen Zhang. I am also thankful for the friends I have made at the climbing gym and in the German-language community, including Lin Yan, Yue Ma, Qingyang Freya Zhou, Yini Nie, Alex Appleton, and Juyee Pan.

My deepest gratitude goes to my parents. My mother has always been incredibly persistent in supporting my learning and education from a young age. Without her unwavering support, I would not be where I am today. Her dedication has been instrumental in shaping me into the person I am today, and for that, I am forever grateful. Last but not least, I thank my best friend and partner, Scott Viteri. I am grateful for the countless hours we have spent co-working, discussing math and science, for the laughter we have shared, and for all the unconditional support and encouragement he has given me in my life.

Chapter 1

Introduction

1.1 Spatial Transcriptomics

1.1.1 Sequencing Technologies

Tissues are complex organizational structures composed of heterogeneous cell types, which work together to perform specific biological functions. Traditional single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by revealing the abundance and composition of different cell types and states within a tissue [1]. However, such sequencing techniques require the dissociation of cells from their native environment. Consequently, the process of preparing cells leads to the loss of spatial information which is essential for understanding how cells interact and function within the context of their local microenvironment. This limitation arises from the need to dissociate the tissues into individual cells, disrupting the original spatial organization of cells within the tissue, and then pooling the cells before separating them into individual reactions for sequencing. After sequencing alignment, the output scRNA-seq data is a count matrix of the number of observed messenger RNA (mRNA) transcripts per gene per single cell. The lack of spatial context prevents the understanding of the complex interplay between cells and their surrounding environment.

To address such challenges, spatial transcriptomics (ST) technologies have emerged to study gene expression profiles while preserving spatial context within tissues [2, 3]. These technologies can be broadly categorized into two main approaches: imaging-based methods and next-generation sequencing (NGS)-based methods [4].

Imaging-based methods, such as Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH) [5], Sequential Fluorescence In Situ Hybridization (seqFISH+) [6], Fluorescent in situ RNA sequencing (FISSEQ)[7] enable in situ transcriptomic profiling at subcellular resolution ($\sim$ 400 nm). These methods can be further divided into in situ sequencing (ISS)-based and in situ hybridization (ISH)-based methods [8]. ISS approaches directly read out the transcript sequence within the tissue by reverse transcribing RNA, amplifying it using rolling circle amplification, and sequencing it. In contrast, ISH methods

detect target sequences by hybridizing complementary fluorescent probes through sequential rounds of hybridization and imaging. While FISH-based methods can potentially cover the whole transcriptome, they require prior knowledge of target sequences to design the probe library. These methods can achieve high sensitivity (up to 80% detection efficiency).

NGS-based methods, such as 10X Genomics Visium, Slide-seq [9, 10], High-Definition Spatial Transcriptomics (HDST) [11], Deterministic Barcoding in Tissue for spatial omics sequencing (DBiT-seq) [12], and spatiotemporal enhanced resolution omics-sequencing (Stereo-seq) [13], sequence the whole transcriptome with spatial spot barcoding. These methods encode positional information onto transcripts before sequencing by barcoding them according to their location in a lattice of spots. For some sequencing methods, thin tissue sections are placed on an array of spatially barcoded features, such as densely packed beads (Slide-seq or HDST) or printed spots (10X Genomics Visium). Permeabilization of the tissue releases the mRNA from the cells, allowing it to diffuse toward the surface and hybridize to the spatially barcoded features. Hybridized material is subsequently processed into cDNA and sequenced using next-generation sequencing. DBiT-seq utilizes microfluidic barcoding to spatially barcode the biomolecules (mRNAs and proteins) in tissues. The spot diameter is $55\mu m$ in 10X Genomics Visium, $10\mu m$ in Slide-seq, $10\mu m$ in DBiT-seq, $3\mu m$ in HDST, and $0.5\mu m$ in Stereo-seq [14]. While 10X Genomics Visium is commercialized, other capture arrays can be difficult and time-consuming to manufacture, and there might be dead space between beads.

1.1.2 Existing Computational Methods Development

To facilitate the analysis and interpretation of ST data, several computational tools have been developed in popular programming languages such as R and Python. In R, packages like Seurat [15], SpatialExperiment [16] (an extension of SingleCellExperiment [17]), and Giotto [18] provide comprehensive frameworks for exploratory data analysis (EDA), data visualization, and downstream analysis. These packages offer functionalities for data preprocessing, normalization, dimensionality reduction, clustering, and the integration of spatial information with gene expression data. Similarly, in Python, tools such as Scanpy [19], Squidpy [20], scVI [21], STUtility [22], stLearn [23], and SpatialData [24] have been developed to address the unique challenges associated with ST data analysis. These packages provide efficient implementations of various algorithms for data preprocessing, visualization, and machine learning-based analysis.

In NGS-based ST data, each spot typically contains multiple cells, potentially of different cell types. To address this, several computational methods have been developed for cell type decomposition using ST data [25]. Stereoscope [26] uses a negative binomial-based probabilistic model, while Cell2location [27] leverages a Bayesian model to infer the spatial distribution of cell types and their gene expression profiles. RCTD [28] develops a supervised learning approach that fits a statistical model to assign cell types using annotated scRNA-seq data. NMFreg [29] and SPOTlight [30] utilize non-negative matrix factorization (NMF) regression to estimate cell type proportions. STRIDE [31] employs a topic modeling method

to deconvolve the spatial mixture. SpatialDWLS [32] proposes a dampened weighted least squares approach, and STdeconvolve [33] is a reference-free latent Dirichlet allocation (LDA) based model for deconvolving ST data. CARD [34] develops a NMF model with a conditional autoregressive modeling assumption to account for the spatial correlation structure.

In sub-cellular resolution ST data, cell boundaries are not explicitly recorded, posing a computational challenge in accurately assigning each pixel to individual cells. This task of cell segmentation is particularly difficult due to the heterogeneity of cellular morphologies, such as variations in cell shape, size, and density across different tissues. To address this challenge, several methods have been proposed. Mesmer [35] presents a deep learning pipeline that incorporates expert annotation and feedback for nuclear and whole-cell segmentation. CellPose [36] constructs an auxiliary vector field flow representation, which can be trained using a U-Net architecture neural network [37]. BIDCell [38] integrates reference scRNA-seq data and biologically-informed loss functions to train a deep learning model for segmentation. Baysor [39] introduces a Bayesian mixture model to calculate the joint likelihood of both detected transcripts and cellular morphology. GeneSegNet [40] employs a deep learning model with a recursive training strategy to account for noise in training labels. Finally, SCS [41] proposes a transformer-based model for cell segmentation that eliminates the need for manual annotation.

To identify genes with spatial expression variation across tissue, multiple statistical and computational methods have been proposed for spatially variable gene (SVG) identification [42]. SpatialDE [43] employs a Gaussian process regression model with different spatial covariance matrices. SPARK [44] directly models the count data with a generalized linear spatial model (GLSM) across different spatial kernels using a penalized quasi-likelihood (PQL) algorithm. MERINGUE [45] extends the spatial auto-correlation and cross-correlation metrics for SVG identification. SPARK-X [46] proposes nonparametric covariance tests with high computational efficiency. Trendsceek [47] develops a marked point processes model, while nnSVG [48] utilizes nearest-neighbor Gaussian processes (NNGP). BOOST-GP [49] leverages a zero-inflated negative binomial model to model ST data. Single-cell Graph Cuts Optimization (scGCO) [50] builds an optimization algorithm using a probabilistic graph model, called a hidden Markov Random Field (HMRF), with graph cuts to identify SVGs. Pattern Recognition Of Spatial Transcriptomics (PROST) [51] calculates a PROST Index score to quantify the spatial variation and significance of gene expression patterns.

Tissue structure segmentation is crucial to characterize the spatial heterogeneity across tissue domains. BayesSpace [52] proposes a Bayesian statistical model with a hidden Markov random field for unsupervised spatial clustering. Spot-based Spatial cell-type Analysis by Multidimensional mRNA density estimation (SSAM) [53] directly assigns cell type labels to each spatial pixel, without the need for cell segmentation, by leveraging both spatial and gene expression information. Belayer [54] models the spatial gene expression data as a piecewise linear function using conformal maps, enabling the identification of spatial domains with distinct gene expression patterns. SpaGCN [55] proposed a graph convolution network model to identify tissue domains and identify domain-specific SVGs using histology images and ST data.

In addition to single-slice analysis, another important computational challenge is the joint integration of multiple tissue slices. This integration can be performed in two ways: horizontally, across adjacent tissue sections, which are typically located next to each other on the same plane, or vertically, across serial tissue sections from different depths within the same tissue. While existing batch correction methods, such as Harmony [56] and scVI [21], have been successfully applied to scRNA-seq data, there is a growing demand for the development of spatial-aware batch integration techniques that can effectively handle the unique challenges associated with integrating ST data from multiple tissue slices. PASTE [57] employs Gromov–Wasserstein optimal transport to achieve probabilistic horizontal and vertical alignment. Gaussian Process Spatial Alignment (GPSA) [58] utilizes a two-layer deep Gaussian Process with warping functions to model the gene expression data. GraphST [59] proposes a graph self-supervised contrastive learning method in which neighboring spatial spots are connected in the graph. GraphST iteratively updates representations to smooth the batch-specific feature distribution and incorporates PASTE for tissue slice alignment.

The high dimensionality of ST data poses challenges for analysis and interpretation. Thus, spatially-aware dimensionality reduction techniques are crucial for downstream analysis. Nonnegative spatial factorization (NSF) [60] is a spatially-aware probabilistic dimensionality reduction model that employs transformed Gaussian processes to encourage sparsity and scalability. MEFISTO [61] models high-dimensional data with known spatial or temporal dependencies between samples, enabling spatio-temporally informed dimensionality reduction and the separation of smooth from non-smooth patterns of variation. Spatial Identification of Cells using Matrix Factorization (SPICEMIX) [62] uses latent variable modeling to elucidate the interplay of spatial and intrinsic factors of cell identity.

The methods discussed above have made significant contributions to the analysis and interpretation of ST data, addressing challenges such as cell type decomposition, cell segmentation, SVG identification, tissue structure segmentation, and multi-slice integration. These advancements have enabled researchers to gain deeper insights into the spatial organization and heterogeneity of tissues at various resolutions. However, despite the progress made by these computational methods, there remain gaps in our understanding of the tumor microenvironment, cell-cell interactions, and cell migration in the context of spatial data. While existing methods have focused on identifying interacting cell types based on known gene pair databases, there is a need for novel approaches that can discover previously unreported gene pairs that potentially interact. Additionally, integrating diverse data modalities, such as scRNA-seq and CRISPR screening data, can provide a more comprehensive understanding of cellular dynamics and regulatory mechanisms in complex biological systems.

This dissertation aims to address these gaps by developing new computational methods and applying them to various datasets to gain insights into the spatial context of the tumor microenvironment, cell-cell communication, and immune cell migration. By addressing these gaps and integrating diverse data modalities, the dissertation contributes to a more comprehensive understanding of cellular heterogeneity, spatial organization, cell-cell communication, and immune cell migration, with potential implications for disease pathogenesis

and therapeutic interventions.

1.2 Cell-Cell Communication

Intercellular communication is a fundamental process that enables cells within multicellular organisms to coordinate their activities, maintain tissue homeostasis, and respond to external stimuli. Such complex networks of interactions are essential for a wide range of biological processes, including development, immune responses, and tissue repair [63]. Cells communicate with each other through various mechanisms, such as signaling via ligands and receptors [64], structural interactions through junction and extracellular matrix proteins [65, 66], neuronal synaptic communications, and the transmission of ions and metabolites. Disrupting cell-cell interaction (CCI) can lead to a breakdown of many essential functions, resulting in various pathological conditions such as developmental disorders, autoimmune diseases, and cancer.

Technological advancements have been made in the field of CCI inference. The synaptic-resolution connectivity map, or connectome, based on three-dimensional electron microscopy reconstruction [67], allows researchers to study brain-wide signal propagation. Connectome analysis has been applied to various organisms, including the adult Drosophila central brain [68]. Novel techniques like uLIPSTIC [69] have been developed to record physical interactions across cell populations, regardless of the specific ligands or receptors involved. These advancements have greatly expanded our understanding of the complex network of cellular interactions in various biological systems.

Transcriptomics data, which provides a comprehensive view of gene expression within cells, has also emerged as a powerful tool for identifying and quantifying CCI. By analyzing the expression levels of genes involved in signaling pathways, ligand-receptor interactions, and other communication mechanisms, researchers can gain valuable insights into the complex networks of cellular crosstalk. For example, transcriptomics data can reveal which signaling molecules are being produced by different cell types, how they are received and processed by target cells, and how these interactions change in response to different stimuli or disease states. Moreover, recent advances in scRNA-seq have enabled researchers to study cell-cell communication while accounting for the heterogeneity of cellular responses and the dynamics of intercellular signaling networks. The integration of transcriptomics data with other omics data and functional studies allow us to better characterize CCI and its role in health and disease [70]. This is not only advancing our fundamental understanding of biology but also providing insights for new therapeutic strategies that target specific communication pathways to treat a wide range of diseases.

Computational tools for inferring CCI often leverage the cellular heterogeneity captured by scRNA-seq [71, 72, 73]. The initial steps involve pre-processing the scRNA-seq data, filtering highly variable genes, and log-normalizing the gene-by-cell count matrix, which transfers the count matrix into the ambient space [74]. In the ambient space, cell cluster assignment and cell type annotation are performed based on the known marker gene expression. Curated

databases provide a list of ligand-receptor gene candidates that could potentially interact across cells [75, 76]. Most computational methods focus on identifying gene pairs that are likely to interact between single-cell populations in the sender and receiver cell types [77]. These methods analyze and score the gene expression levels of candidate gene pairs from existing databases across cell types to identify potential interactions [73].

In addition to cellular heterogeneity across cell types and cell states, spatial heterogeneity plays a crucial role in understanding cell-cell communication in the context of tissue architecture. By integrating ST data with scRNA-seq data, scientists can gain insights into the spatial distribution of genes and infer cell-cell communication within the tissue microenvironment [72, 78]. This spatial information helps to reduce false-positive predictions of cell-cell communication that may arise from analyzing scRNA-seq data in isolation. Several computational methods have been developed to leverage the spatial information provided by ST data for context-dependent CCI inference. Niche-DE [79] identifies genes within specific cell types whose expression is significantly affected by the presence of other cell types in their local neighborhood. NCEM [80] employs graph neural networks to model the effects of a cell's spatial niche composition on its gene expression. SpaOTsc [81] and COMMOT [82] utilize structured optimal transport and collective optimal transport, respectively, to infer cell-cell communication in ST data. COMMOT also takes into account the competition between different ligand and receptor species. SpaTalk [83] builds a graph network that incorporates the target genes in the knowledge graph of the ligand-receptor-target signaling network.

Cell migration is crucial in biological processes ranging from embryonic development, tissue repair, and immune response to disease states such as cancer metastasis [84, 85]. It involves cells moving in a coordinated manner, typically as cohesive groups, which highlights the importance of CCI and communication during migration [86]. Cells employ a variety of signaling pathways and mechanisms to exchange information and coordinate their movement in response to external (bio)chemical and mechanical cues [87]. For instance, chemokines and other secreted signaling molecules can attract or repel specific cell types, directing their migration towards or away from targeted areas [88]. In the context of group migration, leader cells at the front sense chemotactic signals, while follower cells provide feedback and support through physical contacts and signaling pathways such as the Notch pathway [89]. Moreover, cell adhesion molecules and components in the extracellular matrix are fundamental in facilitating or restricting cell movement, and they are frequently modulated by intercellular communication [90]. Therefore, identifying and characterizing the key regulators of cell migration is essential for a deeper understanding of CCI mechanisms and their role in orchestrating collective cellular activities during morphogenesis, tissue regeneration, and cancer progression.

1.3 Dissertation Outline

This dissertation presents computational work on both algorithm development and application for analyzing various data modalities, including ST, scRNA-seq, and CRISPR data. The primary biological questions addressed include characterizing the tumor microenvironment in a spatial context, investigating cell-cell interactions based on spatial data, and studying cell migration across tissues.

Chapter 2 focuses on the analysis of datasets collected by XYZeq, a ST technology that sequences transcriptomes at single-cell resolution while preserving spatial information through barcoding. The chapter applies non-negative matrix factorization (NMF) to identify gene modules that are spatially variable and associated with specific cell types or cellular compositions. Additionally, trajectory inference analysis is performed on mesenchymal stem cells (MSCs) to reveal spatially-dependent gene expression profiles related to their proximity to the tumor. Our analysis uncovers the spatial distribution of distinct cell types, such as hepatocytes and tumor cells, and identifies a cell migration transcriptomic program in tumor-associated MSCs at the tumor-tissue interface.

Chapter 3 introduces a novel computational method called ggPair (**g**ene **g**ene Pair), designed to infer undiscovered ligand-receptor pairs that could potentially interact based on ST data. Unlike existing CCI inference methods that primarily focus on inferring pairs of interacting cell types based on known gene pair databases, ggPair aims to discover novel gene pairs that have not been previously reported in the literature. The method employs a two-step approach: first, it uses a convolutional autoencoder to perform unsupervised feature extraction from ST data. Second, it applies a siamese neural network on the extracted features to predict specific gene-gene interactions. Our model enables the discovery of novel signaling patterns that may be missed by existing techniques relying on distance-metric-based spatial statistics or graph neural networks. Downstream analysis demonstrates that ggPair significantly improves the prediction performance of ligand-receptor interactions compared to state-of-the-art methods.

Chapter 4 integrates scRNA-seq and CRISPR knock-out screening data to identify novel regulators of T cell migration into the central nervous system (CNS). We employ a well-characterized animal model of multiple sclerosis, known as experimental autoimmune encephalomyelitis (EAE), where T cells are the main pathogenic drivers of autoimmune inflammation of the central nervous system (CNS). Comparing the transcriptional profiles of regulatory T cells (Tregs) and effector T cells isolated from the spleen and CNS revealed distinct gene expression signatures, with CNS-infiltrating cells showing up-regulation of genes involved in cellular communication, signaling, and migration. The in vivo pooled CRISPR screen identified several genes, including those related to cytokine signaling, cell adhesion, and G protein-coupled receptor activity, as potential regulators of Treg migration into the CNS. Our findings provide insights into the mechanisms driving T cell dynamics in CNS autoimmunity and highlight potential therapeutic targets for modulating Treg migration and function in the context of autoimmune disorders and cancer.

Overall, this dissertation integrates computational methods and diverse data modali-

ties to address fundamental questions in cellular heterogeneity, spatial organization, cell-cell communication, and immune cell migration. The insights gained from these analyses contribute to our understanding of complex biological systems and have potential implications for disease pathogenesis and therapeutic interventions.

Chapter 2

XYZeq: Spatially-resolved single-cell RNA-sequencing reveals expression heterogeneity in the tumor microenvironment

This chapter includes content that was published in Science Advances (2021):

> Youjin Lee*†, Derek Bogdanoff*, Yutong Wang*, George Hartoularos, Jonathan M. Woo, Cody T. Mowery, Hunter M. Nisonoff, David S. Lee, Yang Sun, James Lee, Sadaf Mehdizadeh, Joshua Cantlon, Eric Shifrut, David N. Nguyen, Theodore L. Roth, Yun S. Song, Alexander Marson†, Eric D. Chow†, Chun Jimmie Ye†, XYZeq: Spatially-resolved single- cell RNA-sequencing reveals expression heterogeneity in the tumor microenvironment. Science Advances 7, eabg4755 (2021)
>
> ∗ indicates equal contribution, † indicates co-corresponding authors.

The content of this chapter is adapted and integrated into the larger argument of this dissertation. It demonstrates the power of integrating cutting-edge experimental techniques with computational approaches to gain deeper insights into the spatial organization and cellular heterogeneity of complex biological systems. By applying computational methods such as non-negative matrix factorization and trajectory inference, we uncover spatially variable gene modules and cell type-specific gene expression patterns in the tumor microenvironment. To better understand how individual cells function within an anatomical space, we developed XYZeq, a novel scRNA-seq workflow that uses combinatorial indexing in microwells to encode spatial metadata into scRNA-seq libraries. Using XYZeq, we profiled heterotopic mouse liver and spleen tumor models, capturing transcriptomes from tens of thousands of cells across eight tissue slices. Analyses of these data revealed the spatial distribution of distinct cell types, a cell migration transcriptomic program in tumor-associated mesenchymal stem cells (MSCs) at the tumor-tissue interface, and localized expression of tumor suppressor

genes by MSCs that vary with respect to proximity to the tumor core. XYZeq simultaneously maps the transcriptome and spatial localization of individual cells in situ, revealing how cell composition and cell states can be altered by location within complex pathological tissue. The methods and findings presented in this chapter are crucial for understanding the complex interplay between cells and their spatial context.

2.1 Introduction

Over the past decade, massively-parallel scRNA-seq has emerged as a powerful approach to deeply catalogue the remarkable cellular heterogeneity in complex tissues ([91], [92]). While scRNA-seq can profile the transcriptomes of thousands of cells in a single experiment, it requires the dissociation of tissue into single cell suspensions prior to library preparation and sequencing, eliminating any spatial information ([93], [94], [95], [96]). Several strategies have emerged to simultaneously obtain molecular and spatial information from complex tissue. Imaging-based strategy combines high resolution microscopy with fluorescent in situ hybridization (FISH) to achieve subcellular resolution and could profile the entire transcriptome ([97], [98], [99], [100]), but this requires lengthy iterative microscopy workflows and large probe panels. Another approach is to hybridize RNA directly from tissue slices onto a microarray containing spatially-barcoded oligo(dT) spots or beads to encode location information into RNA-sequencing libraries. These approaches can sample the entire transcriptome without the need for iterative rounds of hybridization ([101]) and recent improvements using DNA-barcoded beads (HDST and Slide-seqv1/v2) report spatial resolutions at or below the diameter of cells ([29], [102], [103]). However, because of the low numbers of mRNA molecules captured per bead, these spatial transcriptomic approaches often aggregate neighboring beads prior to downstream analysis, resulting in lower effective resolution and averaging of transcript abundances from multiple cells. As a result, annotation of specific cell types present within each spatial unit of analysis is accomplished by aggregating gene sets computationally defined from orthogonal scRNA-seq datasets ([104], [105]). While integration methods have demonstrated the ability to localize cell types within the spatial organization of complex tissue, they rely on having available data from two independent assays and have limited ability to infer how spatial context influences the cell state of individual cell types.

2.2 Results

To overcome these limitations, we have developed XYZeq, a method that uses two rounds of split-pool indexing to encode the spatial location of each cell from a tissue sample into combinatorially-indexed scRNA-seq libraries ([106], [107]). Critical for the performance of XYZeq, we fixed tissue slices with dithio-bis(succinimidyl propionate) (DSP), a reversible

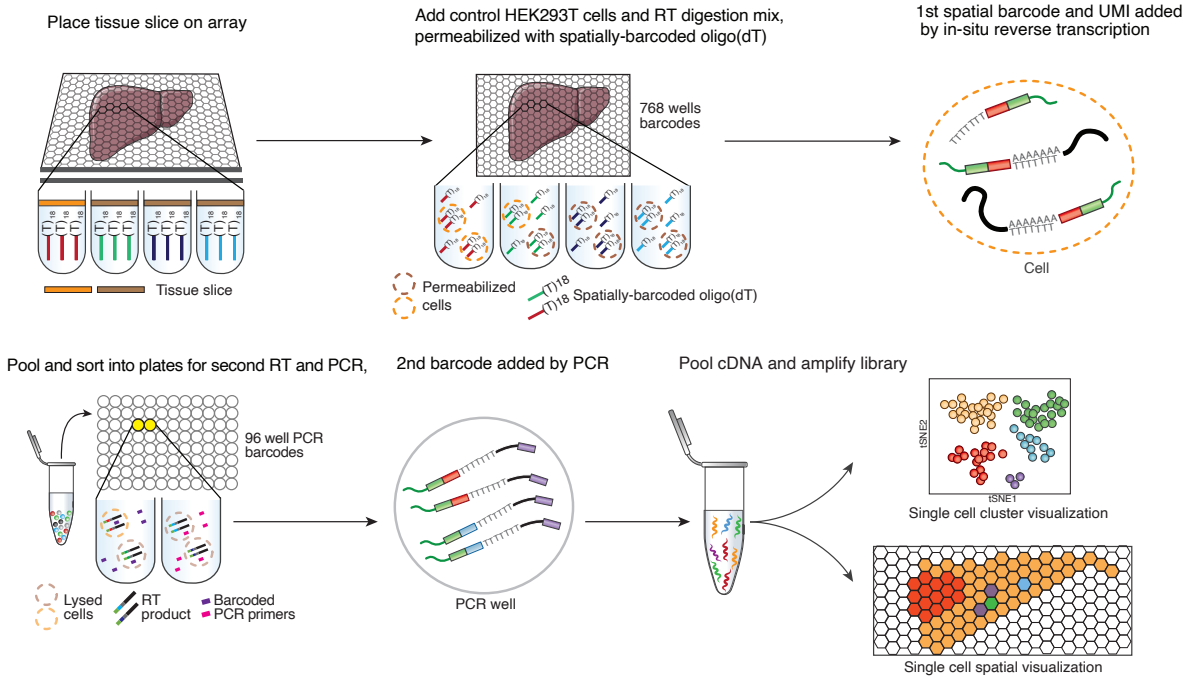


Figure 2.1: Schematic of the XYZeq workflow.

cross-linking fixative that has been shown to preserve histological tissue morphology while maintaining RNA integrity for single cell transcriptomics ([108]). In the first round of indexing, a fixed and cryo-preserved tissue section is placed on and sealed into an array of microwells spaced $500\mu m$ center-to-center. The microwells contain distinctly barcoded reverse transcription (RT) primers (spatial barcode). This step physically partitions intact cells from tissue into distinct in situ barcoding reactions. After reverse transcription, intact cells are removed from the array, pooled, and distributed into wells for a second round of PCR indexing, imparting each single cell with a combinatorial barcode (Fig. 2.1, 2.2A). After sequencing and demultiplexing, the spatial barcode maps each cell back to its physical location in the array (Fig. 2.2A). This combinatorial barcoding strategy theoretically could enable spatial transcriptomic analysis of large sets of single cells – with two rounds of split-pool indexing, 768 spatial RT-barcodes and 384 PCR-barcodes, up to $294,912$ unique single-cell barcodes can be generated.

To determine whether XYZeq can assign transcriptomes to single cells, we performed a mixed species experiment where a total of 11 distinct ratios of DSP-fixed human (HEK293T) and mouse (NIH/3T3) cell mixtures were deposited into each of the 768 barcoded microwells, creating a cell proportion gradient along the columns of the array (Fig. 2.2B, Methods). XYZeq was used to generate scRNA-seq data for $6,447$ cells. 94.8% of cell barcodes were assigned to a single species with an estimated barcode collision rate of 5.1% based on the

percentage of cell barcodes with reads mapping to both human and mouse transcriptomes (Fig. A.1A). We hypothesized that a portion of collisions were due to contamination from ambient RNA released by damaged cells. Using DeconX ([109]), a hierarchical Bayesian method that assumes the observed transcript counts of a cell is a mixture of counts from two binomial distributions, we removed contaminating transcripts, reducing the collision rate to 0.7% (Fig. 2.2C, Methods). After decontamination and removal of collision events, we obtained a median of 939 UMIs and 439 genes per human cell and 816 UMIs and 336 genes per mouse cell. Mapping each single cell to its originating microwell, we observed a high concordance between the observed and expected cell type proportions along the columns of the wells (Lin's Concordance Correlation Coefficient = 0.91; Fig. 2.2D and Fig. A.1B). Together, these results demonstrate that a minimal amount of barcode contamination takes place from single cells in each well and between neighboring wells on the array after pooling, indicating that the XYZeq workflow successfully produces spatially resolved scRNA-seq libraries.

We next applied XYZeq to a fixed and cryopreserved heterotopic murine tumor model established by intrahepatic injections of a syngeneic colon adenocarcinoma cell line, MC38, into immunocompetent mice. This model mimics tissue infiltrating features of metastatic cancer, and more importantly, is associated with a relatively well-defined tumor/liver boundary ([110], [111]). MC38 tumor cells also have immunomodulating properties with previous data showing immune cells infiltrating the tumor/tissue interface approximately 10 days post tumor inoculation ([112], [113]). Thus, we predicted that XYZeq could simultaneously capture the gene expression states and spatial organization of parenchymal liver cells, cancer cells, and tumor-associated immune cell populations. A $25\mu m$ slice of fixed frozen liver/tumor tissue from a C57BL/6 mouse was placed on top of the pre-frozen microwell array while a sequential $10\mu m$ slice was fixed for immunohistochemical staining (Fig. A.2A and Methods). We also deposited fixed human HEK293T cells into the same array at an average of 58 cells per well to serve as a mixed-species internal control to experimentally quantify collision rates. We performed XYZeq and observed an initial collision rate of 7.3% based on comparing the ratio of human versus mouse transcripts (Fig. A.2B). After computational decontamination and further quality control, which includes filtering cells based on cell counts and mitochondrial expression, the collision rate was reduced to 4.4% (Fig. 2.3A, Methods). After removing collisions, we obtained a total of $8,746$ cells and detected a median of $1,596$ UMIs and 629 unique genes per HEK293T cell and $1,009$ UMIs and 456 unique genes per cell from the heterotopic murine tumor model at 46% sequencing saturation (Fig. 2.3B). A hematoxylin and eosin (H&E) stained serial section of the tissue revealed a histological boundary between the tumor and adjacent liver tissue (Fig. 2.3C). As expected, we observed HEK293T human cells distributed across the entire array, while mouse cells were sequestered within the boundary of the murine tissue (Fig. 2.3D). Note, empty spatial wells with no cells detected were likely due to a limited number of cells sequenced ($\sim 10,000$). We obtained a median of 3 human cells/well and 9 mouse cells/well with a total of 13 cells/well expected (Fig. A.3).

XYZeq revealed distinct cell types within the murine liver and tumor. Semi-supervised Leiden clustering revealed thirteen cell populations in the murine tumor model (Fig. A.4A),

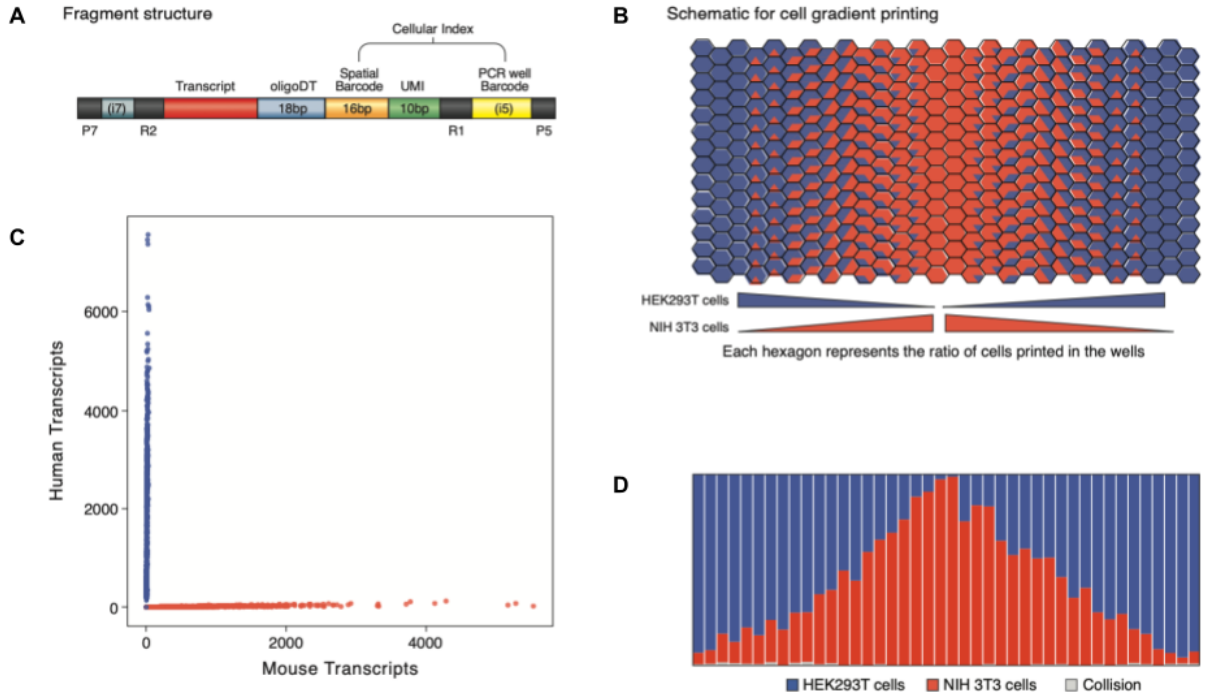


Figure 2.2: XYZeq enables single-cell and spatial transcriptome profiling simultaneously. (A) Schematic of XYZeq sequencing library structure. P5 and P7, Illumina adaptors; bp, base pairs; R1 and R2, annealing sites for Illumina sequencing primers. (B) Schematic representation of the mixed-species cell gradient pattern printed on the chip with 11 unique cell proportion ratios (see Methods for specific cell proportion ratios). (C) Scatterplot of mouse (x axis) and human (y axis) UMI counts detected from a mixture of HEK293T and NIH 3T3 cells after computational decontamination. Blue refers to human cells ($n = 4182$), red refers to mouse cells ($n = 2,220$), and gray refers to collisions ($n = 45$). (D) Proportion of HEK293T (blue) cells, NIH 3T3 (red) cells, or collisions (gray) detected by XYZeq for each column of the microwell array.

from which seven cell types were annotated based on markers that define each population: hepatocytes, tumor cells (MC38), Kupffer cells, liver sinusoidal endothelial cells (LSECs), mesenchymal stem cells (MSCs), lymphocytes, and myeloid cells (Fig. 2.4A). The annotation of MC38 cancer cells was supported by a high correlation of chromosomal copy numbers estimated from XYZeq scRNA-seq data and publicly available MC38 cytogenetic data (Pearson $r = 0.78$) ([114]). Notably, a partial amplification of chromosome 15 and a partial deletion of chromosome 14 observed in the XYZeq data were consistent with common chromosomal abnormalities seen in MC38 cells (Fig. A.4B). As a negative control, we saw low chromosomal copy number correlation when comparing MC38 cells to hepatocytes ([115]) and immune cells ([110]) (Pearson $r = 0.05$ and $r = 0.17$ respectively) (Fig. A.4B). A heatmap showing dif-

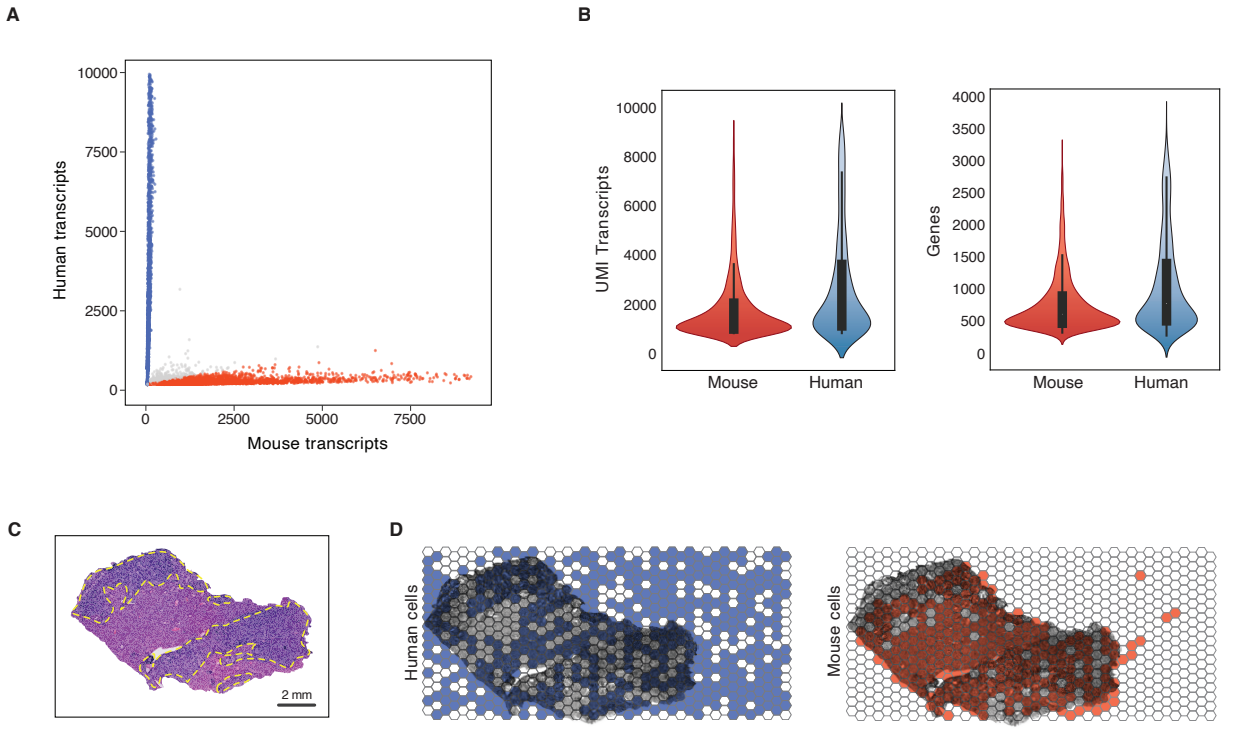


Figure 2.3: Spatially resolved single-cell transcriptomes captured from tissue. (A) Scatterplot of mouse (x axis) and human (y axis) UMI counts detected from liver/tumor tissues ($n = 4$) at 500 UMI cutoff after decontamination processing. Blue refers to human cells ($n = 2,657$), red refers to mouse cells ($n = 5,707$), and gray refers to collisions ($n = 382$). (B) Violin plots showing the number of detected UMIs (left) and genes (right) per mouse (red) and human (blue) cell. Median UMI counts for human cells: 1,596; mouse cells: 1,009. Median gene counts for human cells: 629; mouse cells: 456 across all liver/tumor slices. (C) H&E-stained image of the liver/tumor tissue slice. Tumor region, dark purple with yellow dotted outlines; liver region, pink. Scale bar, 2 mm. (D) Visualization of human (blue) and mouse (red) cell distribution on the XYZeq array overlayed on the H&E-stained slice.

ferentially expressed genes across seven cell types uncovered distinct clusters of cells defined by expression of canonical genes that are relatively exclusive to each cell type (Fig. 2.4B). Note, we estimated uniformly low contamination of each cell cluster (median under 1%) with the exception of hepatocytes, which had slightly more at 2.2% (Fig. A.4C, Methods). We found comparable median UMIs and genes detected across all cell clusters including immune cell populations that have been difficult to profile using other combinatorial indexing methods ([116]) (Fig. A.4D and A.4E). Cell types expected in healthy liver were identified using markers previously described, which included hepatocytes, Kupffer cells, and LSECs (26). Consistent with known heterogeneity of hepatocytes, we identified hepatocyte subsets

annotated by the expression of pericentral markers (Glul, Oat and Gulo) ([115]) (Fig. A.5). MC38 adenocarcinoma cells comprised a large uniform cluster and were distinguished by the expression of the known marker Plec ([111]). Myeloid cells were defined by canonical markers Cd11b and Cd74 ([117]), but other non-canonical markers were also observed, including Myo1f ([118]) and Tgfb ([119]). Lymphocytes showed a similar mix of broad and specific expression patterns of cell type markers, with expression of pan-lymphocyte marker Il18r1, T-lymphocyte marker Prkcq and cytotoxic T-cell marker Cd8b ([120], [121], [122]). Lastly, we detected a cluster of mesenchymal stem/stromal cells which expressed both broad mesenchymal cell markers Rbms3 and Tshz2 and stem/stromal cell markers Prkg1 and Gpc6 ([123], [124], [125], [126], [127]) (Fig. A.5).

We next assessed the reproducibility of XYZeq while comparing changes in the transcriptional landscape across the z-layer of the organ. Four non-sequential $25\mu m$ tissue slices from the same frozen liver tumor sample block were processed and analyzed. The average expression over all cells for genes detected across all slices were highly correlated between each pair of slices (average pairwise Spearman $r = 0.93$) (Fig. A.6). We noted that among the four tissue sections, slice 1 and slice 2, which were the two most proximal slices in their z-coordinates (separated by $80\mu m$), had the highest expression correlation (Spearman $r = 0.96$). In contrast, slice 1 and slice 4, which were the most distal in z-coordinates (separated by $830\mu m$), had the lowest correlations (Spearman $r = 0.91$). Further, clusters jointly annotated across all four slices consisted of cells from each slice, suggesting the observed heterogeneity is not due to batch effects (Fig. A.7).

We further compared the quality of the scRNA-seq data generated by XYZeq to other single cell technologies that are commercially available. To accomplish this, we compared the cell type clusters identified from XYZeq to those identified from an independent scRNA-seq dataset of the same liver tumor model generated using the 10X Genomics droplet-based Chromium system. Almost all cell populations detected by 10X were also observed by XYZeq, except neutrophils, erythroid progenitors, and plasma cells (Fig. 2.4C and Fig. A.8A), immune cell populations known to be sensitive to the cryopreservation required for XYZeq. Interestingly, 10X did not capture MSCs even though cells were isolated from fresh liver/tumor samples. In addition, B cells identified using the 10X platform correlated with the myeloid population detected by XYZeq, likely due to the transcript capture of Ly86, Cd74 and several Class II histocompatibility antigen genes (e.g., H2ab1 or H2dmb1). For the six cell types identified in both the 10X and XYZeq data, we generally observed high correlations in both the cell-type proportions (Lin's concordance correlation coefficient = 0.99; Fig. A.8B) and the pseudobulk expression profiles of each cell type (Pearson $r = 0.64 - 0.86, p < 0.01$, Fig. 2.4C).

Next, we turned to the critical question of whether XYZeq can determine the spatial location of each cell. To do this, we compared the spatial localization of each cell cluster to the images of H&E-stained sequential slices. First to determine that we could accurately define liver from tumor tissue, we confirmed that the density of hepatocytes and tumor cells across the spatial wells overlap with the histological annotation of the adjacent section (Fig. 2.4D). Projection of other cell types revealed distinct spatial organization patterns for myeloid cells,

lymphocytes, Kupffer cells, MSCs, and LSECs (Fig. 2.4D and A.9). Quantification of cellular composition occupying each spatial well revealed MSCs, lymphocytes, and myeloid cells were co-localized with tumor cells, while Kupffer cells and LSECs co-localized with hepatocytes, suggesting potential regions of cellular interaction in tumor infiltrated tissue (Fig. 2.5A). These qualitative observations were confirmed by pairwise correlation analysis of cell type proportion across all the wells ($0.37 \leq$ Pearson $r \leq 0.77$, $p < 0.05$; Fig. 2.5B and A.10).

To assess the generalizability of XYZeq to other tissues, we processed samples from the same heterotopic murine tumor model in the spleen. We recovered a total of $7,505$ cells at a median of $1,312$ UMIs and 661 unique genes per HEK293T cell and $1,169$ UMIs and 577 unique genes per mouse cell at an estimated collision rate of 1.36% (Fig. A.11A, Fig. A.11B). Similar to the liver tumor model, XYZeq was able to reconstruct the boundaries of the splenic mouse tissue with the MC38 tumor region annotated on a sequential H&E-stained slice (Fig. A.11C, D, Fig. A.12A). A median of 4 human cells/well and 7 mouse cells/well were detected (Fig. A.12B, C). Semi-supervised Leiden clustering revealed six distinct cell populations for the spleen tumor model including: B cells, T cells, myeloid cells, MSCs, endothelial cells, and MC38 tumor cells (Fig. A.13A). We observed that all four spleen slices contributed to each cell type cluster, suggesting that the annotated clusters are not due to batch effects (Fig. A.13B). A heatmap showing differentially expressed genes across the six cell types revealed distinct clusters of cells expressing canonical genes that are relatively exclusive to each type (Fig. A.13C). Cells from each type could be spatially mapped across the tissue (Fig. A.13D). Collectively, these results demonstrate that XYZeq can generate spatially resolved single cell RNA-seq data from different fixed frozen tissues.

The ability to obtain spatial and single-cell transcriptomic data simultaneously allowed us to assess the effects of cellular composition on gene expression patterns across space. We applied non-negative matrix factorization (NMF) to both the liver and spleen scRNA-seq data to define modules of co-expressed genes and associated the expression of each module in each cell type with its expression across spatial wells. Using our approach, we identified twenty modules of co-expressed genes in each tissue (Methods). As a proof of principle of the approach, we first identified liver/tumor module (LM) 14 from the liver data, which was predominantly expressed by the hepatocyte cluster in the tSNE space (Fig. 2.6A). As expected, the highest LM14 expressing wells were enriched for hepatocytes suggesting that the spatial variability of this module is largely driven by the frequency of hepatocytes (Fig. 2.6B).

Next, we reasoned that because both the liver and spleen were injected with the same cancer cell line, the invading tumors may induce a shared gene expression profile that vary over space, driven in part, by the cellular composition of the tumor microenvironment. To test this hypothesis, we first identified pairs of matching gene modules between the two tissues from the NMF analysis (Methods). We found four distinct liver/tumor modules (LM) that had at least 25% of genes overlapping with spleen/tumor modules (SMs) (Fig. 2.7A and Fig. A.14A). Gene Ontology (GO) analysis of the modules revealed the enrichment of genes implicated in tumor-response, immune regulation, and cell migration (Fig. A.14B, A.15, A.16B). Consistent with the enrichment analysis, many of the genes from these modules

have been implicated in tumorogenesis (complete gene lists in Table A.1, A.2). Unlike LM14, further analysis of these matching modules revealed a heterogenous composition of cell populations that contributed to the expression of specific module genes (Fig. A.14C and Methods). For example, the tumor response module LM5 and its matching modules SM2 and SM12 (Fig. 2.7A, A.14A), consisted of genes predominantly expressed in MC38 cells with some expression in myeloid cells and lymphocytes (Fig. 2.6C, A.14C, Methods: LM5 and SM12 plotted for tSNE plots). The immune regulation modules, LM13 and LM19 (matched with SM7 and SM20), consisted of genes expressed primarily in both conventional (e.g. myeloid and lymphocytes) and non-conventional (e.g., Kupffer cells from liver samples) immune cells (Fig. 2.6C, 2.7A, and A.14C). The expression of these overlapping modules were usually found highest in regions densely infiltrated with cancer cells (Fig. 2.7B-C: LM5 and SM12 plotted for tumor response modules and LM19 and SM7 plotted for immune regulation modules). Collectively, these results show that the joint analysis of scRNA-seq and spatial metadata from XYZeq can identify spatially variable gene modules due to differences in cellular composition across a tissue sample.

We next focused our analysis on LM10 (matched with SM15 and SM17), a module primarily expressed by MSCs and enriched for genes involved in cell migration (Fig. 2.7A, Fig. 2.8A, Fig. A.14C, and Fig. A.16A-B). Because MSCs are known to possess homing abilities to injured or inflamed sites ([128]), we hypothesized that LM10 could be differentially expressed in space based on the proximity of MSCs to their neighboring cells and tumor. To test this hypothesis, we first computed a tumor proximity score for each well based on its distance to and the composition of neighboring wells (Fig. 2.8B; see Methods and Fig. A.17 for score definition). Projecting the proximity score onto MSCs in tSNE space revealed that the transcriptional heterogeneity of the population is associated with spatial proximity (Fig. 2.9A). We then analyzed the MSC expression profiles using tradeSeq ([129]) to identify differentially expressed genes that tracked with the proximity score. We identified and clustered 177 genes from the liver/tumor tissue ($p < 0.05$) and 66 genes from the spleen/tumor tissue ($p < 0.05$) that are associated with the continuous, one-dimensional proximity score (Fig. 2.9B). The genes were broadly divided into three groups: intra-tumor, tumor-tissue boundary, and intra-tissue with statistically significant genes highlighted for the spleen/tumor tissue (Benjamini-Hochberg FDR < 0.05) (Fig. 2.9B). Interestingly, for MSCs found in the intratumor regions of the spleen, many of the differentially expressed genes are reported to regulate the extracellular matrix (ECM) (Fig. 2.9B, right panel) ([130], [131], [132], [133]), suggesting that MC38 cells may induce a local gene expression program in neighboring MSCs that could contribute to malignant remodeling of the ECM.

Lastly, we leveraged the scRNA-seq data from XYZeq to visualize how individual MSCs expressed Tshz2 and Csmd1, two genes of divergent function that are spatially variable with respect to the tumor in the spleen. Both genes are characterized as tumor suppressor genes and are often silenced in cancer cells to promote malignant growth and metastasis ([125], [134], [135]). However, we found that spleen/tumor MSCs expressed lower levels of Csmd1 but higher levels of Tshz2 in closer proximity to the tumor (Fig. 2.9C). The mean differential expression of these genes was specific to splenic MSCs and not expressed

by MC38 tumor cells. The expression pattern of each of these genes in space revealed a pattern consistent with the aforementioned spatial trajectory analysis, suggesting that their heterogeneous expression in MSCs may be determined by the location of the cells with respect to tumor (Fig. 2.10). Together, these results reveal that joint analysis of spatial and single-cell transcriptomic data from XYZeq can detect transcriptionally variable genes within specific cell types (e.g., MSCs) driven by their location within the complex tissue architecture.

2.3 Discussion

We introduce XYZeq, a new single-cell RNA-sequencing workflow that encodes spatial meta information at $500\mu m$ resolution. XYZeq enables unbiased single-cell transcriptomic analysis to capture the full spectrum of cell types and states while simultaneously placing each cell within the spatial context of complex tissue. In two murine tumor models, we demonstrate that XYZeq identifies both spatially variable patterns of gene expression determined by cellular composition and heterogeneity within a cell type determined by spatial proximity. Looking forward, XYZeq provides a scalable workflow that can easily be adapted to multiple z-layers of tissue and can potentially facilitate analysis of entire organs. Large scale integrated profiling of multiple modalities of single cells mapped to the structural features of their tissue will enable greater understanding of how the tissue microenvironment affect cellular infiltration and interaction in health and disease.

2.4 Methods

2.4.1 XYZeq Methodology

Liver/tumor organ was mounted on a Cyrostat (Leica) and sliced at $25\mu m$ for use as a XYZeq experimental sample or mounted on a histology slide at $10\mu m$ for immunohistochemical staining. On the day of experiment, XYZeq microwell array chips were spotted with reverse transcription cocktail mix that were spiked in with fixed HEK293T cells. The microwell array chips were brought down to -80C and tissue slice was placed on top of the array. A digital image was taken to document the orientation of the tissue before sandwiching a silicone gasket sheet between the XYZeq microwell array chip and a blank histology slide. The chip was placed in a Microarray Hybridization Chamber (Agilent) to ensure an air tight seal while undergoing tissue digestion and reverse transcription. In order to recover high quality RNA from fixed frozen tissue, the Microarray Hybridization Chamber housing the chip had to undergo a gradual step-wise temperature increase to 42C before the 20min incubation to undergo reverse transcription. The chip was removed from the chamber and placed in a 50ml conical tube with 50ml of 1x SSC buffer and 25% FCS. The tube was vortexed and spun down at 1000rcf for 10 minutes. Excess volume was removed and cells were filtered and stained for DAPI (Life Technologies) prior to sorting (BD Aria) into 96 well plates preloaded with 5ul of

second RT mix. Plates were reverse transcribed for 1.5 hours at 42C, followed by PCR using 2x Kapa Hotstart Readymix (Kapa Biosystems). PCR amplification was performed with indexing primer (5'-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCT ACACGACGCTCTTCCGATCT-3'; IDT). Contents of the PCR plate were pooled into 2ml Eppendorf tubes and cDNA was purified with AMpure XP SPRI bead (Beckman). cDNA was tagmented and amplified with Illumina Nextera library p7 index (IDT). Final library was analyzed by BioAnalyzer (Agilent) and quantified by Qubit (Invitrogen) and sequenced on a NovaSeq or HiSeq 4000 (Illumina) (read 1 : 26 cycles, read 2 : 98 cycles, index 1 : 8 cycles, index 2 : 8 cycles)

2.4.2 XYZeq Decontamination Analysis

In our analysis, we recognized some reads aligning to the mouse genes were present in cells that otherwise had high alignment to the human genome. We suspected these reads were ambient RNA contamination, and sought to remove them. We first removed mouse-aligned transcripts with an extremely high expression in human cell population ($n = 59$, log(counts $+1) > 6$). The human cell population was considered a control in the contamination detection, because any ambient RNA from lysed cells was expected to contaminate both mouse and human cells. DecontX ([109]) was then performed to estimate the contamination rate for different cell populations using the human-mouse mixture dataset, and therefore derive a decontaminated count matrix from the raw data. Briefly, the algorithm applies variational inference to model the observed counts of each cell as a mixture of true gene expression of its corresponding cell population and the contamination signature (from other cell populations), and then subtracts the contamination signature (Fig. A.4C). By considering the human-mouse mixed species experiment, we could remove those counts potentially contributing to collision, and effectively account for all potential transcripts in the lysed cells which contribute to ambient RNA. In Fig. A.4C, the initial estimated contamination rate for each mouse cell type are plotted with the median estimates ranging from 0.06% - 0.31% with the highest seen in the hepatocyte cell cluster with 2.18% initial contamination fraction. All the downstream analysis was performed based on the decontaminated data after contamination removal.

2.4.3 How Distinctions Were Made between Collision Rate and Contamination Rate

The collision rate is directly calculated from the gene expression of human-mouse mixture dataset based on the ratio between mouse-aligned and human-aligned transcripts, while the contamination rate for each cell is estimated as a cell-specific parameter in the Bayesian hierarchical model via variational inference from DecontX. In order to specify the contamination rate, each cell has a beta-distributed parameter modeling its proportion of transcript counts which come from its native expression distribution. The estimated contamination rate for each cell is the proportion of transcript counts which come from contamination in

the Bayesian model. Each transcript in a cell follows a multinomial distribution parameterized by the native expression distribution of its cell population or contamination from all the other cell populations, given a Bernoulli hidden state, indicating whether the transcript comes from its native expression distribution or from the contamination distribution.

2.4.4 XYZeq Single Cell Analysis

Single cell RNA sequence data processing was performed where sequencing reads were processed as previously described ([106]). Briefly, raw base calls were converted to FASTQ files and demultiplexed on the second combinatorial index using bcl2fastq v2.20. Reads were trimmed using trim galore v0.6.5, aligned to a mixed human (GRCh38) mouse (mm10) reference genome and UMI deduplicated. Reads were then assigned to single cells by demultiplexing on the first combinatorial index, prior to the construction of a gene by cell count matrix. The count matrix was processed using the Scanpy toolkit. Cells with less than 500 UMIs and greater than 10000 UMIs, as well as cells expressing less than 100 unique genes or more than $15,000$, were discarded. Cells with more than 1% mitochondrial read percentage were also discarded. Gene counts were normalized to $10,000$ per cell, log transformed, and further filtered for high mean expression and high dispersion using the filter genes dispersion function, with a minimum mean of 0.35, maximum mean of 7, and minimum dispersion of 1. Gene counts were then corrected using the regress out function with total counts per cell and the percentage mitochondrial UMIs per cell as covariates. Subsequent dimensionality reduction was done by scaling the gene counts to a mean of 0 and unit variance, followed by principal component analysis, computing of a neighborhood graph, and t-distributed stochastic neighbor embedding (tSNE). Leiden clustering was performed with a resolution of 0.8, and cells were grouped to reveal distinct murine cell types and human HEK293T cells.

2.4.5 10X Data Processing

Counts matrices were generated using the 'count' tool from Cellranger version 3.1.0, using the combined human and mouse reference dataset (version 3.1.0) and the 'chemistry' flag set to 'fiveprime'. The count matrix was processed using the Scanpy toolkit. Cells with less than 500 UMIs and greater than $75,000$ UMIs, as well as cell expressing less than 100 unique genes and greater than $10,000$, were discarded. Cells with more than 7.5% mitochondrial read percentage were also discarded. Gene counts were normalized to $10,000$ per cell, log transformed, and further filtered for high mean expression and high dispersion using the filter genes dispersion function, with a minimum mean of 0.2, maximum mean of 7, and minimum dispersion of 1. Gene counts were then corrected using the regress out function with total counts per cell and the percentage mitochondrial UMIs per cell as covariates. Subsequent dimensionality reduction was done by scaling the gene counts to a mean of 0 and unit variance, followed by principal component analysis, computing of a neighborhood graph, and tSNE. Leiden clustering was performed with a resolution of 1, and cells were grouped to reveal major murine cell types and human HEK293T cells.

2.4.6 Heatmap for XYZeq

Mouse cells were subsetted from the XYZeq processed data matrix. The processed gene expression values were plotted in a heatmap with a minimum fold change of 1.5 and hierarchically clustered using the heatmap function from Scanpy, with the default settings of Pearson correlation method and complete linkage.

2.4.7 XYZeq Gene Pairplot

Four slices of either liver/tumor or spleen/tumor tissue were processed using the XYZeq assay (with HEK293T cells spiked-in) and aligned to a joint human and mouse reference. All genes with at least one count in each slice were kept, and the counts across the common set of genes between pairwise slices were plotted in the lower triangle, with the Spearman correlation for the data shown in the upper triangle. Along the diagonal, histograms were plotted showing the distribution of counts per gene for all the non-zero genes for each slice.

2.4.8 XYZeq Cell/Well Pairplot

Pairplot showing the number of microwells containing pairwise combinations of cell types. For scatter plots, each point in the plot represents a well, and its coordinate positions indicates the number of cells of each cell type present in that well. Every dot on the scatter plot is a gene representing mean per gene for common genes across all cells in the slices. Along the diagonal of the figure are histograms, showing the univariate distribution of cell number per well for the given cell type.

2.4.9 Heat Map Comparing 10X to XYZeq

Mouse cells were subsetted from each of the processed data matrices. For pairwise mouse Leiden clusters found between XYZeq and 10X, the scaled and log transformed gene expression values of common genes were plotted. For each comparison, a Pearson correlation was calculated and plotted in the heatmap. Row/column labels were ordered according to their corresponding cell types. Correlation plot: Mouse cells were subsetted from each of the processed data matrices. Proportions for each cell type (as determined by the Lieden clustering and visualized using tSNE) were plotted, and the coefficient of determination was calculated by fitting to the model that assumes proportions are equal between the two assays.

2.4.10 Gene Module Analysis of Top Contributing Genes

To identify gene modules using NMF, genes expressed in fewer than five cells and cells expressing fewer than 100 genes were filtered out. Variance stabilizing transformation was performed on count data, and confounding covariates including number of counts per cell,

batch, and mitochondrial read percentage were regressed out by a regularized negative binomial regression model using the SCTransform ([136]) function in the Seurat R package. Pearson residual values from the regression model were centered, and all negative values were con-verted to zero. Nonsmooth NMF (nsNMF) was performed on the re-sulting expression data with a rank value of 20 using the nmf ([137]) function in NMF R package. In each module, genes were sorted by their magnitude in the corresponding coefficient matrix in a descend-ing order. Gene ontology enrichment analysis was performed for the sorted genes in each module using GOrilla ([138]). For each module, the top consecutive genes with higher coefficients in this module com-pared to all the other modules were further selected as genes contributing the most to the module ([139]) in the tissue-specific analysis. Binary spatial plots were generated by first calculating the median expression across all the cells for each well within each batch based on the log-normalized gene expression data. We then extracted the mean expression across all the genes within one module for each well and calculated the average of mean expression across selected module genes for each well weighted by the number of cells in each well. The wells with a mean expression across genes above the weighted average were labeled as highly expressing for that gene module, and all the other wells with nonzero expression of those selected module genes were labeled as lowly expressing that gene module. tSNE plots representing the gene modules were colored by their mean expression of genes within the annotated module.

2.4.11 Overlapping Analysis between the Gene Modules Identified in Liver and Spleen

Gene modules were first identified using nsNMF with a rank value of 20 for the two tissues, liver and spleen, respectively. The top 200 genes in each sorted gene list for a module were selected as having high association with the module. For each module in the liver tissue, the spleen module with the largest gene overlap was initially matched as functionally similar. We then removed those matched pairs with fewer than 25% overlapping genes out of top 200 genes in the liver module. In order to calculate cell type fractions that make up each module, the average gene expression for each gene across all the cells was calculated. Median expression across all the overlapping genes for each cell type was further computed, which was later transformed into fractions by dividing by the sum of median expression across all the cell types.

2.4.12 Defining the Proximity Score by Wells

We sought to define a score for each well of the hexagonal well array that would capture how centrally located a well was within either the tumor or non-tumor tissue domains. Central to the method was the determination of successive concentric "layers" of wells that were adjacent to a well in question: those corresponding to its immediate neighbors (layer 1), those wells exactly 2 wells away (layer 2), and so on, for n layers. In the spleen, we selected several wells on the far side of the tumor region and set the score of these wells

to 1. We then took 10 successive layers of wells and decreased the score linearly with each layer, with the wells in layers 10 and beyond set to 0. In the liver, MC38 cells were found in different locations, and therefore, unlike the spleen, there was no single unidirectional spatial dimension to place all MC38 cells at one end and all non-tumor tissue cells at the other. Therefore, we used an alternative approach to calculate these scores in the liver tissue. For each well $w_{x,y}$, annotated by their x, y position on the hexagonal well array, we calculated the proportion of hepatocytes, $p_{x,y}$, since the hepatocytes were the most abundant parenchymal cell type in, and strictly associated with, the non-tumor liver tissue:

$$\begin{aligned} t_{x,y} &= \text{number of total hepatocytes and MC38 cells in} w_{x,y} \\ h_{x,y} &= \text{number of hepatocytes in} w_{x,y} \\ p_{x,y} &= \frac{h_{x,y}}{t_{x,y}} \end{aligned}$$

Then, for each well in question $w_{x,y}$, we tabulated the surrounding wells in each of the successive concentric 10 layers. We denote these wells $w_{x',y'}$ to differentiate from the well in question. For each of those layers l, we took its constituent wells' $p_{x',y'}$ and calculated a cell number-weighted average $p_{x,y,l}$:

$$\begin{aligned} w_{x,y,l} &= \{w_{x',y'} \in \text{layer } l \text{ of } w_{x,y}\} \\ t_{x,y,l} &= \text{number of total hepatocytes and MC38 cells in } w_{x,y,l} \\ p_{x,y,l} &= \sum_{x',y'}^{w_{x,y,l}} \frac{t_{x',y'}}{t_{x,y,l}} p_{x',y'} \end{aligned}$$

Then, for the well in question $w_{x,y}$, we calculated a distance weighted average of all the $p_{x,y,l}$, and this became the proximity score $s_{x,y}$ for the well in question. The distance weights for each layer, μ_l, were based on an exponential decay, terminated to 10 terms and then normalized to 1 by dividing by the sum of all weights μ_s. We give equal weight to ¶$_{x,y}$ and the value for the layer 1 neighbors $p_{x,y,1}$. A decay factor d of 1.05 was chosen empirically, as it seemed to create the most uniform-like distribution of the scores across all wells.

$$\begin{aligned} d = 1.05, u_s &= \sum_{l=1}^{10} \frac{1}{d^l} \\ u_l &= \frac{1/d^l}{u_s} \\ s_{x,y} &= u_1 p_{x,y} + \sum_{l=1}^{10} u_l p_{x,y,l} \end{aligned}$$

These calculations were repeated for all wells containing at least 1 murine cell.

2.4.13 Trajectory Inference Analysis

Genes expressed in fewer than 5 cells, and cells expressing fewer than 100 genes were excluded. Variance stabilizing transformation was performed using the SCTransform ([136]) function in the R Seurat package. The resulting corrected count data in MSC in one tissue was used as the count matrix input in trajectory inference analysis, using the tradeSeq ([130]) package in R. Genes whose expression is associated with the proximity score were identified by the associationTest function in tradeSeq, based on a Wald test under the negative binomial generalized additive model. The p-values were corrected using Benjamimi-Hochberg multiple testing procedure, and genes with corrected p-values smaller than 0.05 were considered to be significantly associated with the proximity score. Continuous spatial plots were generated by first calculating the median expression across all the cells for each well within each batch based on the log-normalized gene expression data. We then extracted the mean expression across all the genes within one module for each well. Binary spatial plots were generated by calculating the average of mean expression across selected module genes for each well weighted by the number of cells in each well. The wells with a mean expression across genes above the weighted average were labeled as highly expressing for that gene module, and all the other wells with non-zero expression of those selected module genes were labeled as lowly expressing that gene module. tSNE plots representing the gene modules were colored by their mean expression of genes within the annotated module.

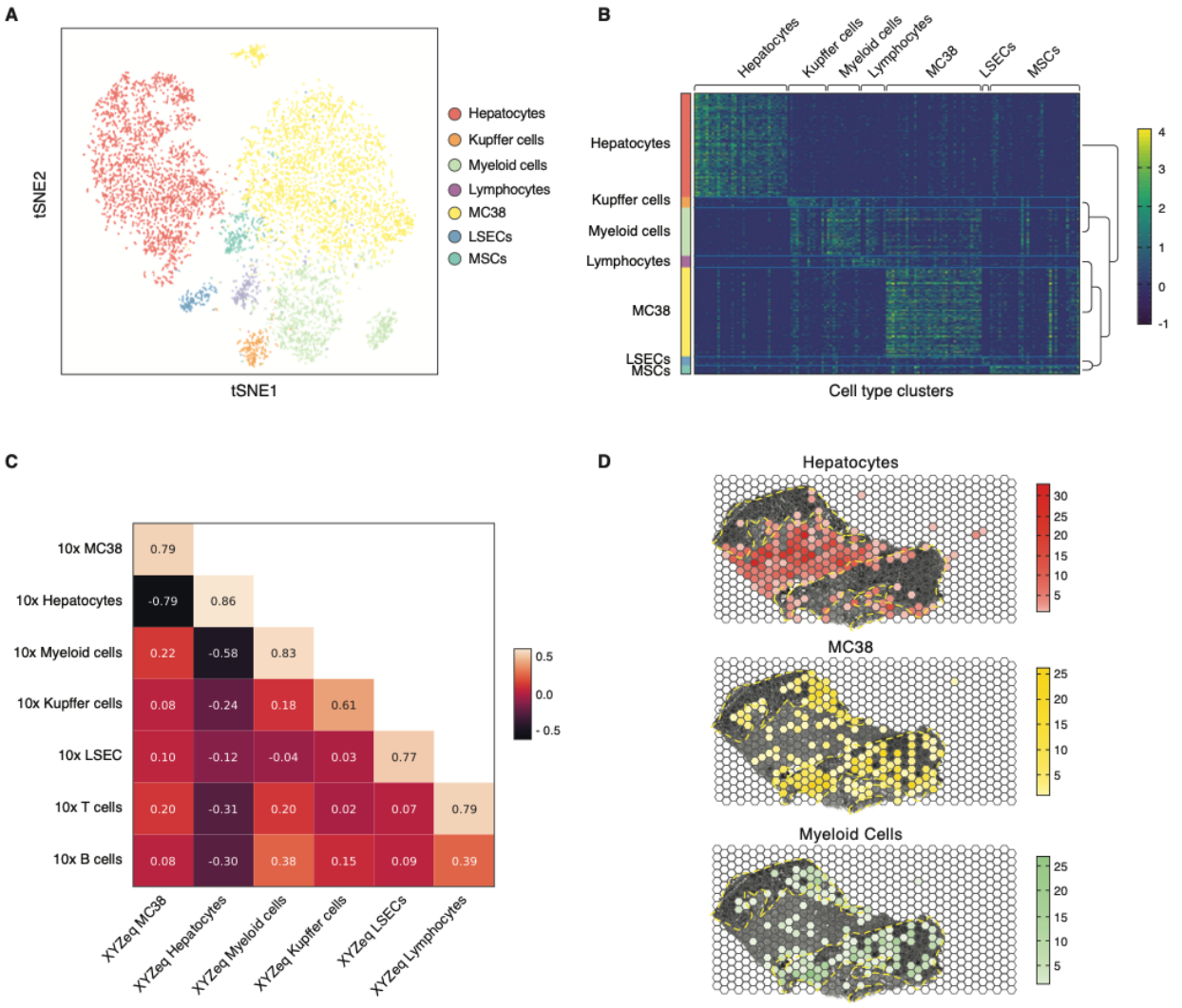


Figure 2.4: Frequency and spatial mapping of single-cell clusters from tissue. (A) tSNE visualization of the cell types identified from liver/tumor tissue. A total of 6,623 total cells were plotted. (B) Heatmap of scaled marker gene expression and hierarchical clustering of genes that define each cell type from liver/tumor tissue. Reference for color bar in (A). (C) Correlations of pseudobulk expression values for matching cell types between XYZeq and 10x Genomics Chromium. (D) Spatial localization of hepatocytes, MC38, and myeloid cells overlaid on a bright-field image of tissue. Yellow dotted outline indicates tumor regions.

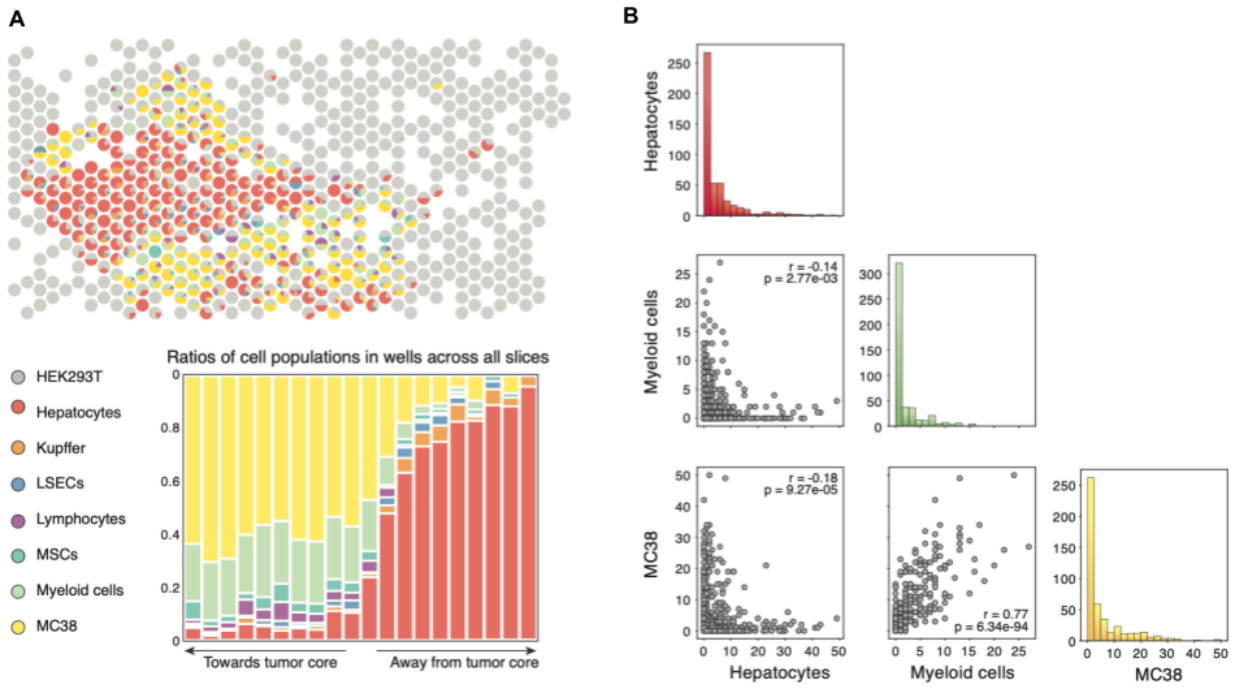


Figure 2.5: (A) Pie chart of cell type composition for each XYZeq well from a representative liver/tumor tissue slice (top) and bar chart illustrating combined cell type composition across all four slices of liver/tumor tissue, which tracks with proximity to the tumor (bottom) (see Methods for proximity score). (B) Pairplot showing the frequency of hepatocytes, MC38, and myeloid cells in each well. Scatterplots show the colocalization of two cell types in each well. Histograms show the distribution of number of cells (x axis) per well (y axis) for each cell type. Pearson correlation (r) and P values are annotated.

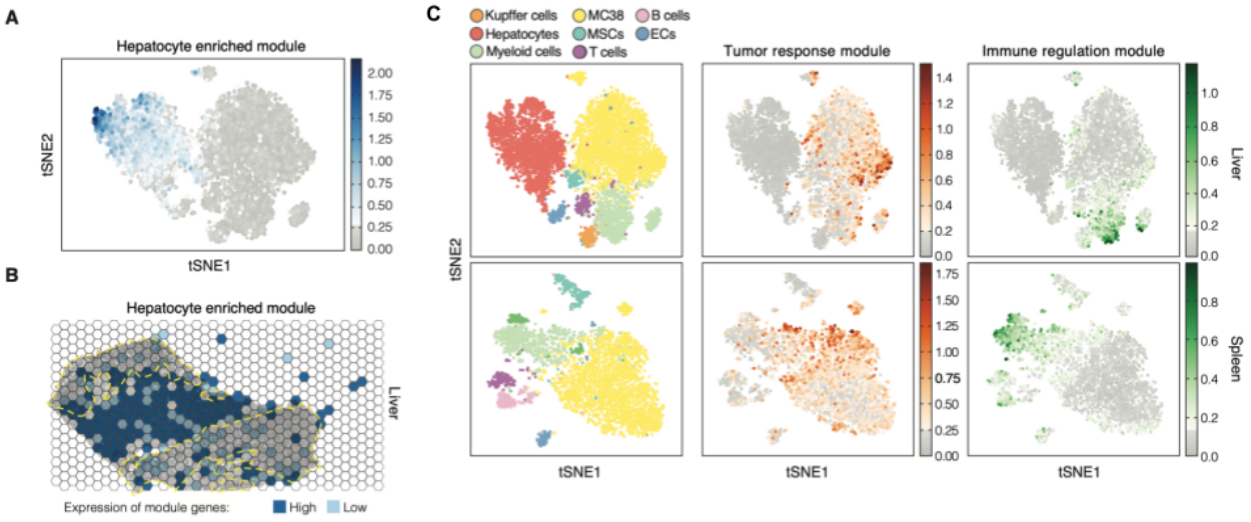


Figure 2.6: Expression of gene modules in space that track with cellular composition. (A) Projection of average expression of hepatocyte-enriched module (LM14) in tSNE space. Each dot is a cell and colored by the average expression of top contributing module genes (Materials and Methods). (B) Spatial expression of hepatocyte-enriched module (LM14). Each spatial well is colored by the average expression of the top contributing module genes weighted by the number of cells per well. Wells are binarized into high (above weighted average) versus low (all other nonzero expression). Yellow dotted outlines indicate tumor regions. (C) tSNE projection of XYZeq scRNA-seq data colored by annotated cell types in liver/tumor (top left) and spleen/tumor (bottom left) and mean gene expression of the top overlapping modules between liver/tumor (top row) and spleen/tumor (bottom row). Tumor response modules correspond to LM5 and SM12, and immune regulation modules correspond to LM19 and SM7. ECs, endothelial cells.

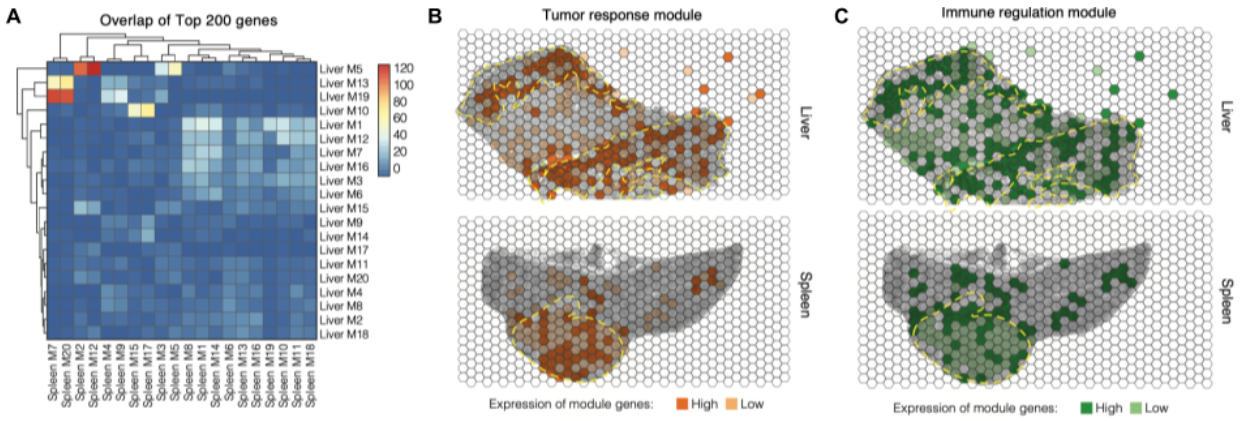


Figure 2.7: (A) Heatmap representing the number of overlapping genes between each pair of modules in liver/tumor and spleen/tumor. Each row is an LM, and each column is an SM. Spatial projection visualizes the mean expression of the tumor response modules (B) corresponding to LM5 and SM12 and the immune regulation modules (C) corresponding to LM19 and SM7. Each well in (B) and (C) is colored by the average gene expression of each module weighted by the number of cells per well (high versus low), and yellow dotted outline indicates tumor regions. Wells are binarized into high (above weighted average) versus low (all other nonzero expression).

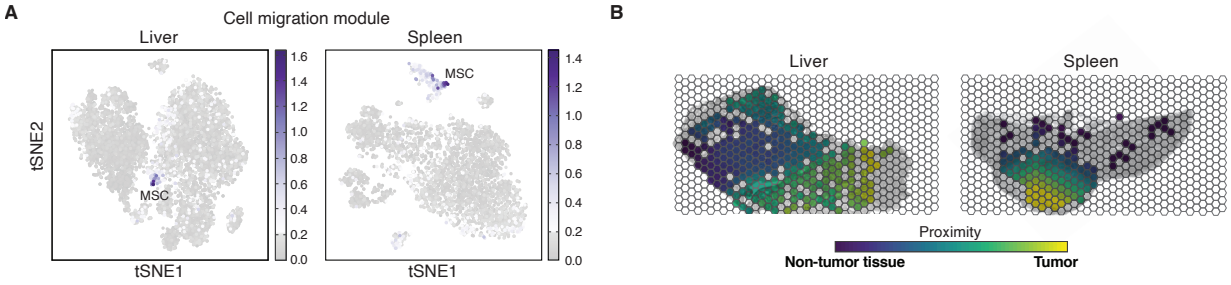


Figure 2.8: Differential gene expression within MSCs associated with their spatial proximity to tumor. (A) Average expression of the cell migration modules (LM10 and SM17) in tSNE space. Each dot is a cell colored by its mean expression of the top module genes between corresponding liver/tumor and spleen/tumor modules. (B) XYZeq array colored by the tumor proximity score. Values near 1 (yellow) indicate regions rich in tumor, values near 0 (purple) indicate regions rich in nontumor cells, and wells capturing the border between the two tissue types take on values around 0.5 (blue/green).

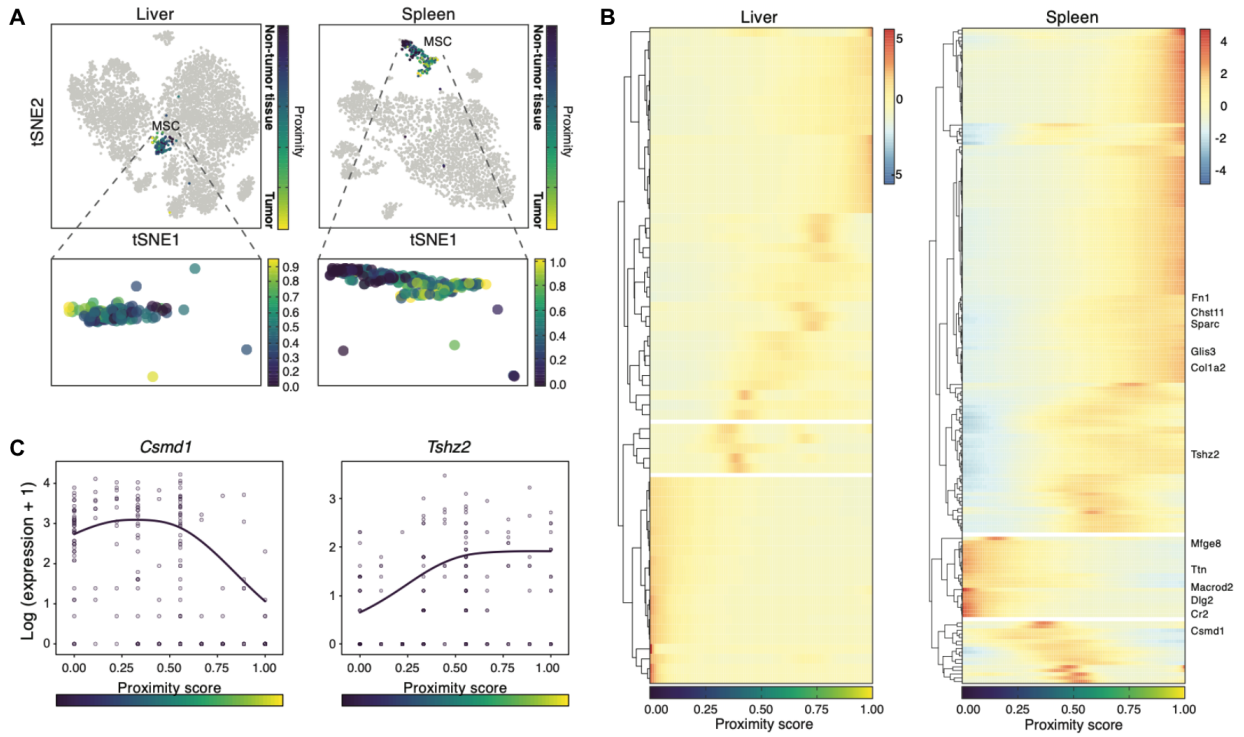


Figure 2.9: Differential gene expression within MSCs associated with their spatial proximity to tumor. (A) MSCs colored by the cell-specific proximity score in tSNE space. (B) Row-clustered heatmap showing the scaled, mean gene expression in MSCs of genes enriched in three spatial regions (intratumor, boundary, and intratissue) along the one-dimensional proximity score. For spleen/tumor, statistically significant genes enriched in the tumor and nontumor regions are highlighted. (C) Log expression (y axis) of Csmd1 (left) and Tshz2 (right) along the proximity score (x axis). Each dot corresponds to one MSC cell, and the regression line is fitted using the negative binomial distribution (Materials and Methods).

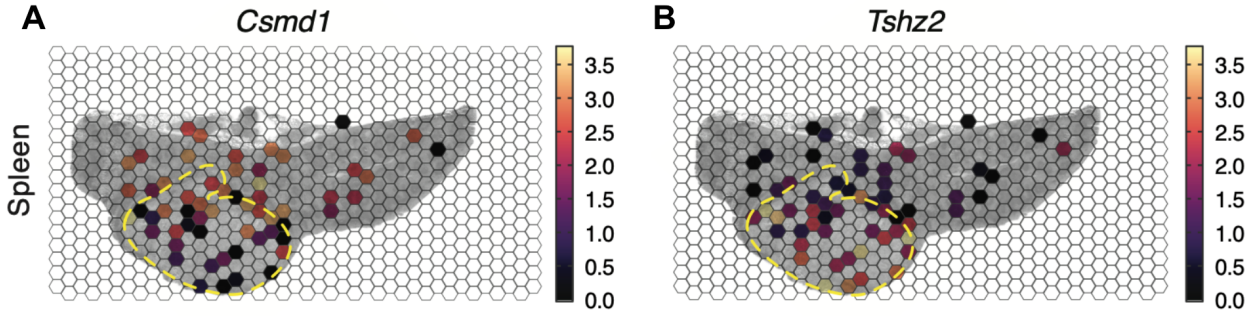


Figure 2.10: Projection in space of mean expression of Csmd1 (A) and Tshz2 (B) in MSCs. Yellow dotted outline indicates tumor region.

Chapter 3

Decoding cell-cell communication through deep learning reveals novel ligand-receptor pairs from spatial transcriptomics

This chapter includes content that is unpublished and in preparation:

> Yutong Wang*, Fanding Zhou*, Antoine Koehl, Yun S. Song. Decoding cell-cell communication through deep learning reveals novel ligand-receptor pairs from spatial transcriptomics.
>
> $*$ indicates equal contribution

Cell-cell communication, often mediated by ligand-receptor interactions, is crucial for orchestrating complex cellular interplay within tissues. Spatial transcriptomics provides new opportunities to map these networks by profiling gene expression across tissue sections. However, our understanding of the full repertoire of ligand-receptor interactions remains incomplete, and computational methods to predict these interactions directly from spatial gene patterns are limited. We introduce **ggPair** (**g**ene **g**ene **Pair**), a novel deep learning framework that combines unsupervised feature extraction with supervised similarity modeling to predict both known and previously uncharacterized ligand-receptor pairs. Our approach employs a convolutional autoencoder to learn data-driven spatial representations, followed by a siamese neural network for interaction classification, without relying on prior knowledge or predefined constraints. By learning the spatial gene expression patterns, ggPair enables the discovery of new gene pairs and significantly improves prediction performance over state-of-the-art techniques like graph networks and spatial correlation. Focusing on key drug target classes like G-protein coupled receptors, we apply statistical testing to rigorously validate

predictions, thereby uncovering novel signaling pathways and providing a powerful platform to unlock new biology and therapeutic opportunities.

3.1 Introduction

The cellular milieu is a complex, dynamic environment where cells engage in numerous interactions to maintain homeostasis. Cell-cell communication can occur through direct contact or over longer distances. Major mechanisms include tight junctions, membrane nanotubes, electrical synapses, and ligand-receptor signaling. Ligand-receptor interactions, in particular, allow cells to communicate through surface proteins binding signaling molecules. These can act locally through autocrine and paracrine signaling or globally via endocrine signaling.

Such localized autocrine and paracrine effects suggest leveraging spatial proximity relationships from transcriptomics data could be helpful to infer cellular communication. Computational methods for analyzing cell-cell communication through spatial transcriptomics (ST) can be organized into two primary categories: 1) those aimed at the identification of interacting cell types and the characterization of intercellular communication patterns, and 2) those focused on the prediction of specific gene-gene interactions that underlie signaling mechanisms between cells. Approaches like Giotto [140], NCEM [141], and COMMOT [142] focus on the first category, learning communication patterns between cells or cell types, often conditioned on known interacting gene pairs.

However, fewer computational methods address the second category of identifying ligand-receptor gene pairs mediating cell-cell signaling. This is especially important for elucidating orphan G-protein coupled receptors (GPCRs). While GPCRs are implicated in a broad range of biological processes and are the target of approximately 35% [143] Food and Drug Administration (FDA)-approved drugs, approximately 20% of human GPCRs remain 'orphaned', with unknown endogenous ligands. Remarkably, some of these orphan receptors exhibit evolutionary conservation [144], suggesting vital roles in regulating processes like brain development over extensive evolutionary timelines. Understanding these elusive interactions is essential for drug discovery and deciphering cellular signaling networks [145].

While databases like CellPhoneDB [75] and CellChat [76] have curated known ligand-receptor interactions through expert curation of the literature, they are inherently limited to prior knowledge and cannot predict novel undiscovered interactions. For instance, CellPhoneDB identifies interacting pairs from ligand-receptor candidates through permutation tests on scRNA-seq gene expression. The input of the statistical procedure is the single cell gene expression levels of ligands and receptors in two different cell types. The significant cell-type-specific interactions are identified by reshuffling cell-specific cell type labels and calculating a p-value for the mean ligand-receptor interactions. However, such statistical procedure only accesses already curated gene pairs, and lacks the capacity to identify novel ligand-receptor interactions beyond the scope of the existing literature.

For computational ligand-receptor interaction prediction, current state-of-the-art techniques rely primarily on spatial statistics or graph neural networks. Traditional correla-

tion metrics like spatial cross-correlation offer initial insights but are often constrained by their inability to capture complex patterns or to deal with data sparsity and noise. Recent approaches, such as SpatialDM [146], leverage established spatial statistics (global and local Moran's I [147]) for spatial transcriptomics data, focusing on identifying spatially co-expressed genes from curated ligand-receptor pairs. While SpatialDM has the potential to discover new pairs, its performance in this regard is not yet established. Moreover, its reliance on Euclidean distance assumes uniformity across tissue spaces, ignoring the spatial context and tissue heterogeneity. Leveraging graph neural networks, Graph Convolutional Neural Networks (GCNG) [148] represent an advanced approach, using supervised learning to predict interactions between gene pairs by creating gene-pair-specific graphs from spatial gene expression data. Each graph receives a binary label indicating whether the gene pair is expected to interact or not. However, GCNG imposes hardcoded Euclidean distances to define spatial neighbors when constructing the graph, and it leverages very limited labeled data (about $1-2$k positive interacting pairs) for supervised training. These methods, while effective in some settings, overlook the complex and dynamic nature of cellular interactions as they are limited by the pre-defined spatial constraints, and they are unable to incorporate the broader spatial and biological context provided by all the genes encoded in the genome.

Our method represents a significant advance by enabling fully data-driven discovery of ligand-receptor interactions without relying on prior knowledge or hand-engineered spatial relationships. We utilize a two-step approach (Fig. 3.1): first a Convolutional Autoencoder (CAE) [149] performs unsupervised feature extraction from ST data in an interaction label-agnostic manner. The CAE learns a latent representation capturing spatial gene expression patterns from all the 20k geness, including both ligand, receptor, and unlabeled non-ligand, non-receptor genes. We then employ a Siamese Neural Network (SNN) [150] on the extracted features to predict specific gene-gene interactions. This flexible modeling approach can uncover novel and complex signaling patterns missed by existing techniques relying on distance-metric-based spatial statistics or graph neural networks. In particular, by learning spatial representations without predefined neighborhood constraints, our method enhances discovery of novel ligand-receptor interactions unbiased by prior assumptions. This modeling flexibility allows direct characterization of gene-gene communication from spatial data, enabling new biological insights inaccessible to previous methods. We demonstrate that our approach significantly improves prediction performance over current state-of-the-art techniques for predicting ligand-receptor interactions interactions on ST data. Our approach bridges the gap between curated knowledge and flexible data-driven discovery, overcoming key limitations of current techniques for spatial omics analysis.

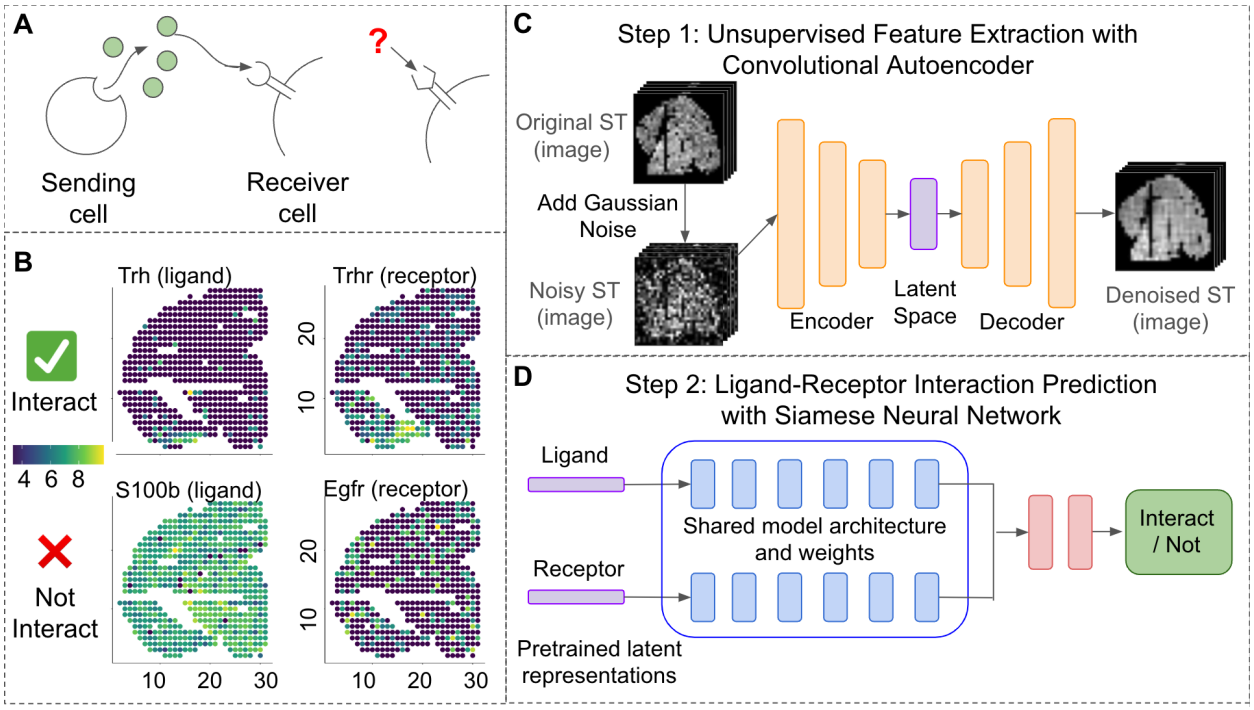


Figure 3.1: Method Overview. **A** The schematic of cell-cell communication through ligand-receptor interaction. The question of interest is to identify ligand-receptor pairs that interact with each other. **B** Examples of (non-)interacting ligand-receptor pairs with (dis)similar spatial gene expression. **C** - **D** ggPair utilizes a two-step approach. The first is to extract features for all the genes using a convolutional autoencoder **C**, and the second is to predict the gene-gene interaction using a Siamese Neural Network **D**.

3.2 Results

3.2.1 Model Framework

Our ggPair framework employs a CAE integrated with a SNN (Methods) for prediction of ligand-receptor interactions from ST data. The approach is designed to analyze multi-batch ST data, where each batch represents a separate biological replicate with all genes profiled across a 2-dim spatial grid. The CAE performs unsupervised feature extraction by considering each gene expression profile as an image. It is trained to denoise and extract meaningful latent representations for each gene while reconstructing the original spatial expression pattern. This allows the model to learn representations tailored to diverse gene families in a label-agnostic manner. The CAE model input is the noisy version of gene expression images, constructed by adding Gaussian noise to the original gene expression images. The CAE model output is the reconstructed gene expression images, and the model loss function is the mean squared error between the original and the denoised gene expression

images. The CAE model is first pre-trained on a single batch of data. It is then fine-tuned separately on each of the remaining batches, with 200 training epochs per batch to fully adapt to batch-specific distributions while avoiding overfitting.

The latent CAE embeddings encoding spatial gene expression are then fed into the SNN model for supervised pairing and classification tailored for similarity modeling. The SNN contains twin subnetworks processing latent representations of ligand and receptor genes separately. This is followed by merged layers to model pairwise interactions and predict interaction labels. The SNN architecture is particularly apt for this task due to its inherent ability to learn complex relationships between paired data, making ggPair proficient in discerning both established and novel ligand-receptor relationships from inherently noisy spatial omics data. The ensemble of model predictions across different Gaussian noise levels improves the model performance.

3.2.2 Performance Comparison to Existing Methods

We first used a mouse brain Spatial Transcriptomics dataset [151] to systematically evaluate ggPair for predicting known ligand-receptor interactions against current state-of-the-art methods like GCNG and SpatialDM. Evaluation of reconstruction quality revealed low mean squared error between the denoised and original images across batches (Fig. B.8), indicating effective feature extraction. Model predictions were aggregated across four biological replicates from different mice of the same cell line and the same age. To evaluate the robustness of our model, a rigorous 5-fold stratified cross-validation experiment was conducted across 20 random seeds, splitting receptor genes into training and testing sets to prevent data leakage. Specifically, each fold of the cross-validation used 20% of the receptors, and their associated ligands, as the testing set, while the remaining 80% constituted the training set. It is worth noting that the negative interaction pairs in each fold were newly sampled, ensuring different sets of negative examples across folds. This approach not only evaluated the model's predictive capabilities but also rigorously assessed its generalizability by ensuring that no receptor or its ligands appeared in both the training and testing sets of a given fold.

ggPair significantly outperformed the other methods across various metrics (Fig. 3.2), including Area Under the Precision Recall Curve (AUPRC) (Table B.1), Area Under the Receiver Operating Characteristic curve (AUROC) (Table B.2), and Accuracy (Table B.3). For the important clinical subset of GPCR interactions, ggPair achieved a mean AUPRC of 0.756 compared to 0.640 for GCNG and 0.530 for SpatialDM, demonstrating substantial gains. The non-overlapping 95% confidence intervals between ggPair and other methods reflect the statistical significance of these improvements in performance. Histograms of predicted probabilities further underscore ggPair's high confidence to distinguish between true and false interaction labels (Fig. 3.5).

We also evaluated the model performance using human ST data [152], to assess how well ggPair generalizes to ST data sequenced from human and across different sequencing platforms. Specifically, we used two human breast cancer 10X Visium ST datasets, and one human osteosarcoma MERFISH [5] dataset (Fig. B.3). ggPair is consistently outperform-

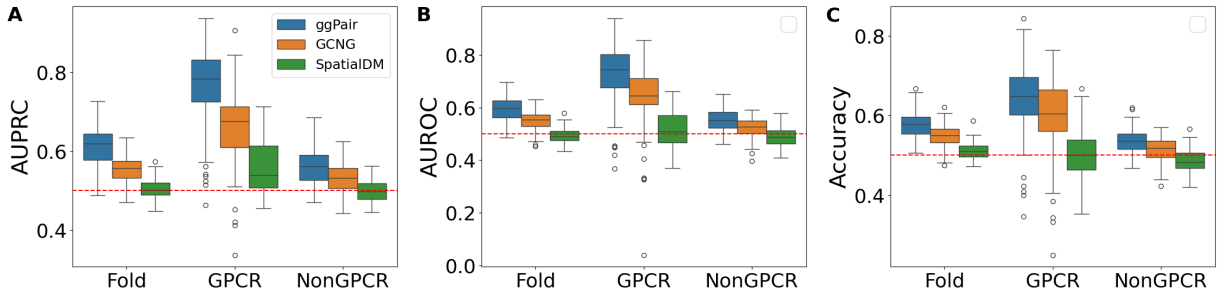


Figure 3.2: Model performance comparison using 5-fold cross-validation on the mouse data, ggPair in blue, GCNG in orange, and SpatialDM in green. AUPRC (A), AUROC (B), Accuracy (C).

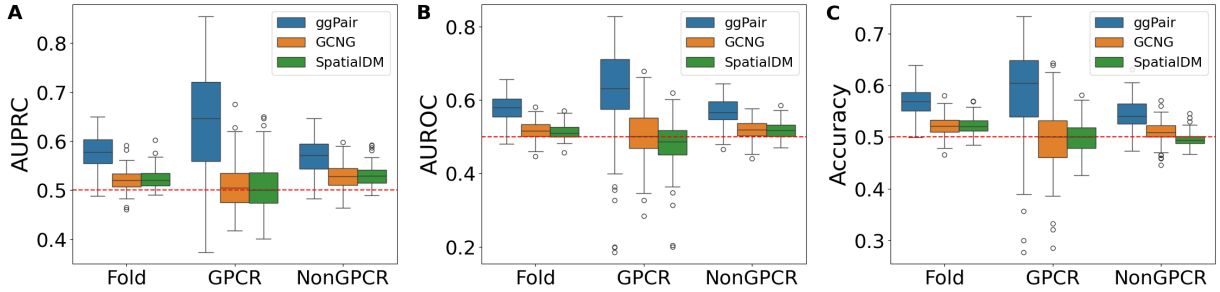


Figure 3.3: Model performance comparison using 5-fold cross-validation on the human osteosarcoma data, ggPair in blue, GCNG in orange, and SpatialDM in green. AUPRC (A), AUROC (B), Accuracy (C).

ing state-of-the-art methods on the five-fold cross-validation experiment, based on AUPRC, AUROC, and Accuracy metrics on three human spatial datasets (Fig. 3.3 - human osteosarcoma MERFISH data, Fig. 3.4 - the first human breast cancer 10X data, and Fig. B.14 - the second human breast cancer 10X data).

3.2.3 Leave-One-Out Validation for GPCRs and nonGPCRs

The number of negative interacting pairs is much more than that of positive interacting pairs. To account for such class imbalance, we developed a tailored Leave-One-Out Cross-Validation (LOOCV) approach for more realistic evaluation. In each LOOCV iteration, a single GPCR receptor was held out for testing while all other receptors were used for training. The left-out receptor was then paired with all potential ligand candidates for prediction. This process was repeated over 20 runs with different random seeds to ensure robustness. This approach mimics the real-world challenge where a model must predict interactions for completely novel receptors without any prior exposure in the training set, thereby testing the model's generalization capability to novel, unobserved cases.

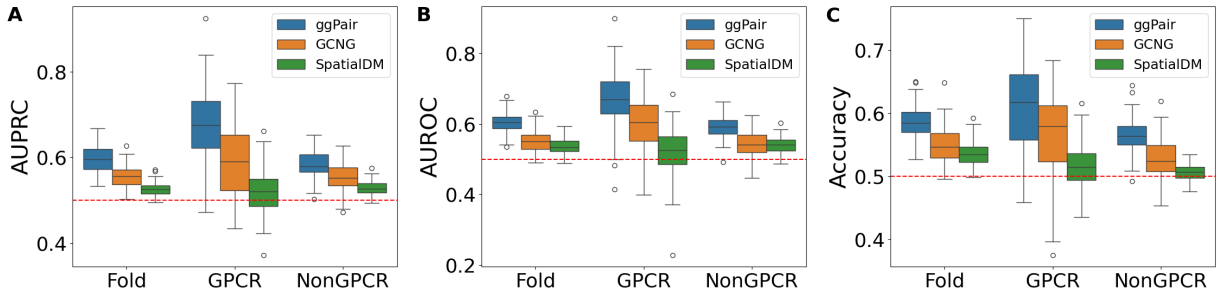


Figure 3.4: Model performance comparison using 5-fold cross-validation on the first human breast cancer data, ggPair in blue, GCNG in orange, and SpatialDM in green. AUPRC (A), AUROC (B), Accuracy (C).

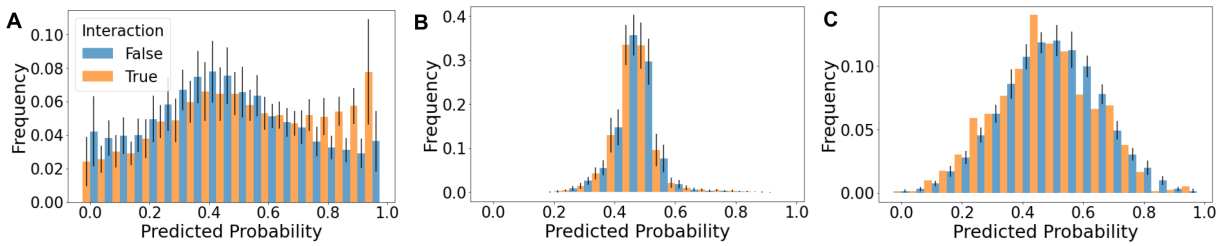


Figure 3.5: Histograms showing the distribution of predicted probabilities for true (orange) and false (blue) ligand-receptor interactions on the mouse data. Each histogram bin and the corresponding error bar represent the average and standard deviation of model predictions across 20 random seeds. ggPair (A), GCNG (B), SpatialDM (C).

Given the particular importance of elucidating GPCR signaling partners, we started by evaluating the performance of our model for individual GPCR receptors using LOOCV on the mouse data. ggPair consistently achieved high predicted class probabilities (i.e., with smaller ranks) for the true interacting ligands (Fig. 3.6). The statistical significance of these results was further validated by hypergeometric testing for over-representation of known ligands among the top 20 ranked candidates.

We next evaluated the models based on precision and recall at the top k calls, stratified based on the number of true interacting ligands per receptor. A high precision score for the top k calls in ggPair (Fig. 3.7, 3.8) could help experimentalists to prioritize experiments when validating the predicted interacting ligands with limited budget and resources. ggPair also has a better performance on nonGPCR receptors than GCNG (Fig. B.16, B.17). ggPair has a much higher recision and recall at the top k call among the groups of receptors whose true interacting ligands are more than 6.

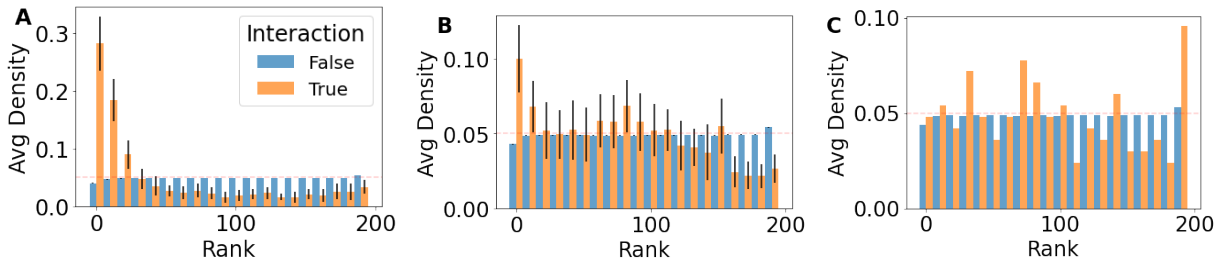


Figure 3.6: The average density of the ligand candidates ranked by the predicted class probabilities, stratified by the true interaction label for each gene pair. In each LOOCV iteration, every GPCR receptor was held out for testing while all the other receptors were used for training. The hold-out GPCR receptor was then paired with all the possible ligand candidates for prediction. ggPair (A), GCNG (B), SpatialDM (C).

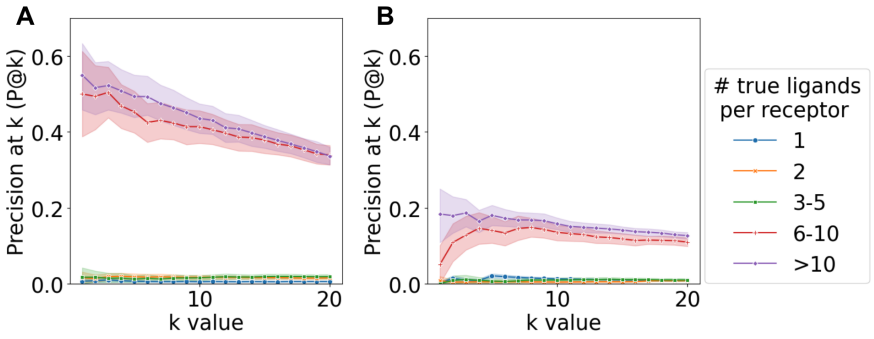


Figure 3.7: Precision at the top k call(s) for GPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B).

3.2.4 Identification of Positive Ligand Candidates for Orphan GPCR

We perform the LOOCV iterations where all the receptors were used in training and each orphan GPCR was held out for testing. The hold-out orphan GPCR receptor was paired up with all the possible ligand candidates for prediction. We repeat such LOOCV experiments 20 times.

Through the repeated LOOCV iterations, we observed distinct patterns in the frequency of ligands appearing within the top k ($k = 15$) calls for each orphan GPCR. Notably, while most ligands were rarely identified among the top k predictions, a few consistently appeared within this top k preferred ranking (Fig. 3.9) across most of the LOOCV iterations. The variability in ligand occurrences was quantitatively analyzed using binomial tests, under the null hypothesis that all the ligands have an equal chance of appearing in the top k calls. We calculated the probability of observing the recorded number of appearances within the top k

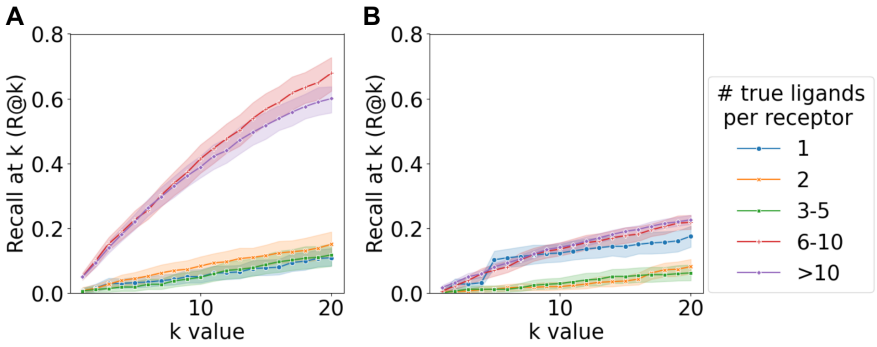


Figure 3.8: Recall at the top k call(s) for GPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B).

ranks under the assumption of equal likelihood for all the ligand candidates (Methods). This analytical approach allowed us to distinguish between ligands that frequently ranked high due to random chance and those that may have a statistically significant association with the orphan GPCR, thereby identifying potential interacting ligands worth further investigation for the orphan GPCR.

As a proof of principle, we performed the binomial tests on all the non-orphan receptors to check the performance of the binomial test procedure. Among all the 244 receptors, 109 receptors had one or more true interacting ligands with a significant p-value in the binomial test after multiple testing correction (Benjamini-Hochberg). ggPair is able to predict multiple true interacting ligands of different receptors correctly (Fig. 3.10). The statistically significant ligands for each orphan GPCR receptors were identified in the same approach (Fig. 3.11).

3.3 Discussion

Our study presents a novel deep learning approach for de novo discovery of ligand-receptor interactions from spatial gene expression patterns. By combining unsupervised representation learning with flexible similarity modeling, we achieve strong performance exceeding previous methods reliant on spatial correlation or graph networks. A key advantage of our technique is the ability to uncover complex signaling patterns without relying on prior knowledge or predefined constraints.

We focused specifically on the critical task of elucidating interactions for GPCRs. Despite their recognized significance in diverse physiological processes, nearly 20% of GPCRs remain "orphaned" with unknown ligands. Our method demonstrates reliable prediction for these GPCR receptors, as validated through rigorous statistical testing. This represents a major advance in overcoming a key challenge in cell signaling research. By expanding the repertoire of known GPCR interactions, our approach facilitates downstream drug target identification

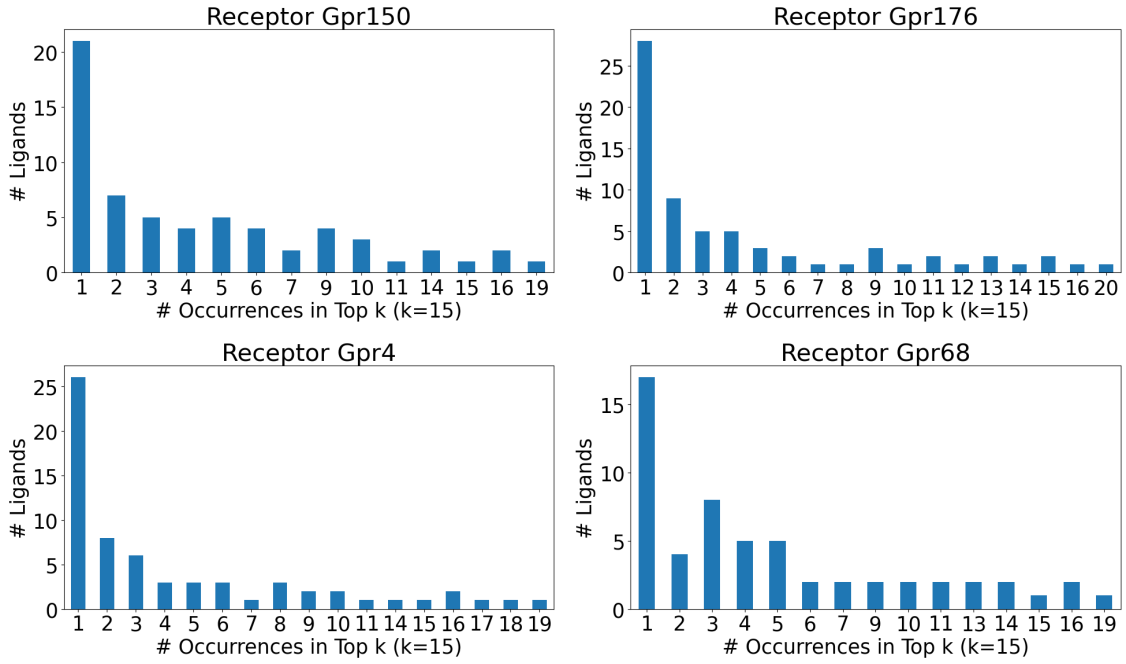


Figure 3.9: The number of ligands with certain number of occurrences in the top k predictions for different orphan GPCR receptors. The top k calls were ligands ranked by the predicted class probability using ggPair.

and mechanistic studies into biological regulation.

While we utilized a convolutional autoencoder architecture, other deep learning models like Vision Transformers (ViT) could be explored as well. Transformers have shown powerful performance on computer vision tasks given sufficient data and compute resources. As spatial transcriptomics datasets continue to grow in size and complexity, ViT models may offer advantages in learning representations tailored to gene expression patterns. There remains substantial room for innovating and advancing computer vision techniques.

In addition to predicting ligand-receptor interactions, our framework could be extended to reconstruct broader gene regulatory networks. The flexible pairing of unsupervised feature learning with similarity modeling is well-suited to identify novel relationships like transcription factor-target connections. This could shed light on the complex circuitry controlling diverse cell types and states. With appropriate training data, our methodology provides a robust platform for mapping diverse cellular interaction networks beyond ligand-receptor signaling.

Some limitations should be noted. Ligand-receptor interactions are often tissue-specific, so larger and more diverse spatial datasets will be needed. Our ground truth labels from curated databases may also have biases and incompleteness in representing true biological signaling networks. We use ligand and receptor gene expression as a proxy for actual protein

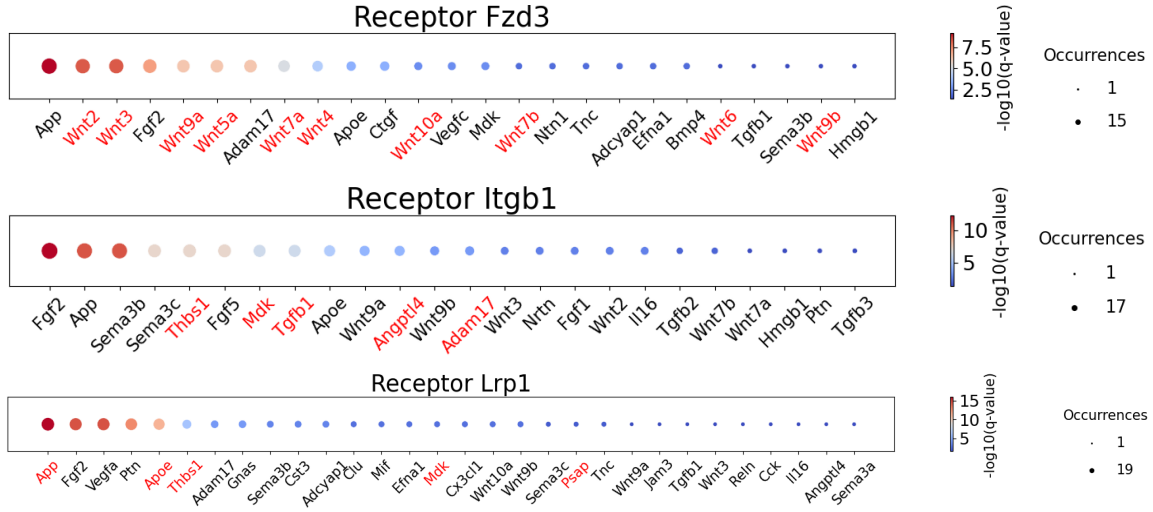


Figure 3.10: Examples from the receptors with known interacting ligands, the $-\log 10$(multiple testing corrected p-value) and the number of occurrences in the top k calls were plotted for all the significant ligands. The true interacting ligands were colored in red.

levels, but these proxies have caveats. For instance, receptor proteins on the cell surface may better indicate signaling activity. Multi-omics data could provide useful additional views, like protein levels for receptors. We also observed that gene-specific effects dominate variation, underscoring the need for larger spatial transcriptomics resources.

Furthermore, GPCR gene expression may not directly correlate with signaling impact due to amplification effects. Even lowly expressed GPCRs can trigger substantial downstream responses. Future work should explore other proxies like biosynthesis enzymes for GPCRs. Our current approach also does not consider tissue-specific effects and relies on ligand and receptor gene expression as surrogates. Incorporating proteomics or protein localization data could help mitigate these issues.

In summary, we present a powerful new framework for mapping cell-cell communication directly from gene expression patterns. By combining deep learning with statistical testing, our approach enables reliable prediction of ligand-receptor interactions while overcoming constraints of prior knowledge and hard-coded models. The ability to elucidate orphan GPCR signaling partners holds tremendous promise for unlocking new biology and therapeutic opportunities. This work crucially bridges the gap between curated knowledge and flexible data-driven discovery.

3.4 Conclusion

In conclusion, we present a novel computational approach that combines unsupervised representation learning via convolutional autoencoders with flexible similarity modeling using

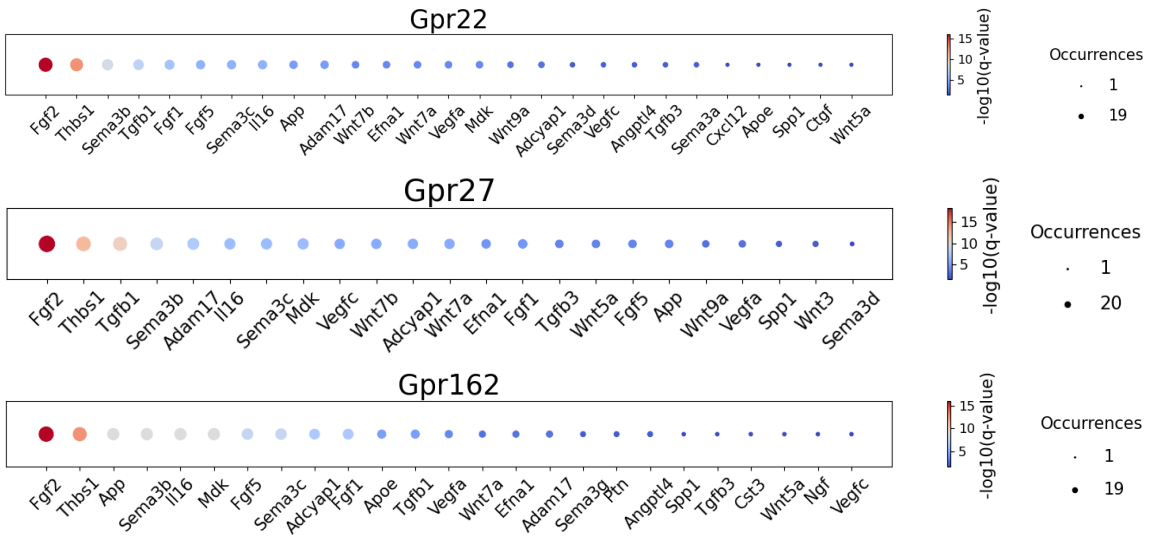


Figure 3.11: Examples from the orphan GPCR receptors, the $-\log 10$(multiple testing corrected p-value) and the number of occurrences in the top k calls were plotted for all the significant ligands.

Siamese neural networks. Our integrated ggPair framework enables reliable prediction of ligand-receptor interactions directly from spatial transcriptomics data, significantly outperforming previous methods reliant on spatial correlation or graph networks. By liberating the model from predefined constraints, this flexible data-driven approach can uncover complicated signaling patterns and novel orphan GPCR partners missed by existing techniques. Through rigorous cross-validation and statistical testing, we demonstrate the robustness of ggPair in elucidating cellular communication networks. Our work establishes a powerful platform to elucidate gene-gene interactions from spatial omics data, bridging the gap between curated knowledge and unbiased discovery. This sets the stage to unravel previously unexplored biological mechanisms and develop targeted therapeutic strategies.

3.5 Methods

3.5.1 Preprocessing of Spatial Transcriptomics Data

We utilized a mouse Spatial Transcriptomics (ST) dataset encompassing $10,327$ spots and $47,298$ genes. Four slices from C57BL/6 wild-type mouse brains of identical age (B02_D1, B02_E1, B03_C2, B03_D2) were selected to minimize confounding variables. Spatial coordinates were adjusted and aligned, followed by the exclusion of non-protein-coding genes, which was performed based on annotations in BioMart. The GraphST algorithm [59] was employed for spatial-aware batch correction, resulting in a refined set of $18,176$ protein-coding genes distributed over $2,277$ spatial spots. For each ST slice, we padded zero values to create a complete 32×32 grid. Noise augmentation was applied by adding Gaussian noise

with a mean of zero and a standard deviation scaled by a factor of 0.2, followed by value clipping.

Two human breast cancer 10X Visium datasets are from GSE176068 in the GEO database. After filtering out genes with more than 95% zero entries, the first (second) dataset includes 280 (280) unique ligands, 321 (306) unique receptors, and $1,129$ $(1,168)$ unique positive interacting ligand-receptor pairs. After normalizing the spatial coordinates, the human breast cancer datasets have 128×128 grids. The human osteosarcoma MERFISH data is from GSE152048, with 220 unique ligands, 250 unique receptors, and $1,026$ unique gene pairs after filtering. The spatial images have 64×64 pixels after normalization.

3.5.2 Compilation of Ligand-Receptor Interaction Annotations

To comprehensively annotate gene pairs with known ligand-receptor interactions, we combined mouse data from three curated databases: CellTalkDB [153], CellChatDB [76], and GPCRdb [154]. We further refined this dataset by excluding non-secreted ligands, which were identified using the secreted protein database from Uniprot. Additionally, genes that had more than 95% zero entries in the observed spatial transcriptomics data were removed. To maintain consistency with the nomenclature in our spatial transcriptomics dataset, gene names were standardized using synonyms when necessary. After these preprocessing steps, our dataset contained 801 annotated gene pairs that were positively identified as ligand-receptor interactions. To balance the classes, we included mismatched gene pairs in an equal ratio to interacting pairs for each unique receptor gene.

The approach for analyzing human data mirrored that of the mouse dataset, involving the combination of human data from CellPhoneDB [75], CellTalkDB, CellChatDB, and GPCRdb. The subsequent steps, including the exclusion of non-secreted ligands, removal of genes predominantly featuring zero entries, standardization of gene names, and balancing of classes, followed the same protocol as established for the mouse data.

3.5.3 Convolutional Autoencoder (CAE) for Unsupervised Feature Extraction

We designed a Convolutional Autoencoder (CAE) to perform unsupervised feature extraction from spatial transcriptomics data. Our CAE model was composed of an encoder and decoder, both containing Convolutional 2D layers and MaxPooling layers for down-sampling. The encoder compresses the input into a latent space, and the decoder reconstructs the input data from this latent representation. CAE input is the noisy spatial transcriptomics data, based on the summation of the original data plus Gaussian noise.

For the encoder, we used two convolutional layers with kernel size 3×3 followed by max-pooling layers. The decoder consists of two convolutional layers with the same kernel size, followed by up-sampling layers. All the convolutional layers use ReLU activation functions and are padded to maintain the input dimensions. We employed the Adam optimizer with a

learning rate of 1×10^{-4} and used mean squared error between the original and reconstructed images as the loss function. To counteract overfitting, we utilized early stopping based on the loss metric with a patience of 5 epochs. We trained the CAE model for 200 epochs with a batch size of 128, which we empirically found to be optimal to prevent overfitting. Following the initial pre-training, the CAE model was fine-tuned on the remaining three ST slices. For each slice, training and validation sets were created using a $80 - 20$ split. The latent embeddings from the encoder were extracted by isolating the output after the second max-pooling layer. For each processed ST slice, the latent embeddings were obtained using the trained encoder and reshaped into 2D arrays.

3.5.4 Siamese Neural Network (SNN) for Pairwise Classification

The Siamese Neural Network (SNN) used in this study is designed to perform binary classification tasks on a pair of input sequences. The model architecture consists of two identical subnetworks, each taking one input sequence. Each subnetwork is a convolutional neural network (CNN) composed of two convolutional layers followed by batch normalization and max-pooling operations. The convolutional layers use 32 and 64 filters of size 3, respectively, with ReLU activation. L2 regularization with a regularization constant of 0.001 is applied to the kernels of the convolutional layers to mitigate overfitting. Batch normalization is applied after each convolutional layer to stabilize the learning process, followed by max-pooling layers with a pool size of 2 for down-sampling. The architecture concludes with a flatten operation and a dense layer with 64 units and ReLU activation. The outputs from the two subnetworks are then concatenated and passed through two additional dense layers with 64 and 32 units and ReLU activation, respectively. Finally, a dense layer with a single unit and a sigmoid activation function produces the binary output.

Training the Siamese Network involves multiple considerations to ensure robustness and generalizability. The model is trained using a 5-fold cross-validation approach as described in the data splitting section. Within each fold, a validation set comprising 10% of the training data is randomly sampled to fine-tune the model parameters. The Adam optimizer is used with a learning rate that is experimentally determined. The loss function used for optimization is binary cross-entropy. The model is trained for a maximum of 200 epochs with early stopping based on the validation accuracy, which helps to prevent overfitting. The training is accelerated using the MirroredStrategy for distributed computing.

3.5.5 Leave-One-Out Strategy and Hypergeometric Testing

In order to rigorously evaluate the performance of our siamese neural network (SNN) in identifying ligand-receptor interactions, we employed a Leave-One-Out Cross-Validation (LOOCV) strategy specifically designed for GPCR genes. In each iteration of LOOCV, one unique GPCR gene was isolated for testing, while the remaining genes were used for training and validation. This allowed us to obtain an unbiased estimation of the model's performance on unseen GPCR genes. Furthermore, we incorporated hypergeometric testing to

assess the statistical significance of the predicted interactions. In addition to the Leave-One-Out Cross-Validation strategy, we employed hypergeometric testing to statistically validate the predicted interactions for each GPCR gene in the test set. The hypergeometric test assesses the statistical significance of having x successful matches among the top 20 ranked ligand candidates, given that n ligands are known to interact with the specific GPCR gene in question. Here, N represents the total pool of ligand candidates, and $k = 20$ is the number of top-ranked ligand candidates we considered for each GPCR gene. The hypergeometric probability $P(x)$ is calculated as follows:

$$P(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

We also implemented the LOOCV methodology for orphan GPCRs. During each LOOCV iteration, the dataset was partitioned such that all receptors, except one orphan GPCR, were included in the training set. The excluded orphan GPCR served as the test set and was paired with every possible ligand candidate for prediction purposes. This procedure was systematically repeated 20 times to enhance the robustness of our findings. To statistically assess the likelihood of each ligand's appearance within the top k predictions, we employed binomial tests. These tests were calculated under the null hypothesis, positing that each ligand had an equal chance of appearing within the top k ranks. The probability of success under this hypothesis was defined as $\frac{k}{N}$, where k is the number of top calls ($k = 15$), N is the total number of ligand candidates ($N = 205$), and the number of trials n reflects the total repetitions of the experiment ($n = 20$). Therefore, the formula for the probability of observing a certain number of occurrences in the top k ranks is

$$P(\#\text{ Occurrences in Top k}) \sim \text{Binomial}\left(n, \frac{k}{N}\right).$$

3.5.6 Implementation Details for Other Methods

We employed the autocrine+ GCNG model, which takes into account both autocrine and exocrine interactions using gene expression data in conjunction with a spatial distance matrix. We opted for this particular model based on its superior performance as demonstrated in the benchmarking section of the GCNG paper. Our model follows the same structure as the original autocrine+ GCNG, and we maintained the default hyperparameters. For the purposes of replication and benchmarking, we constructed the model using version v0.0.6 of skeptual.

We also applied SpatialDM, a method that utilizes global correlation statistics to identify spatially co-expressed ligand and receptor pairs, as our baseline. This approach is rooted in multiple hypothesis testing, exhibiting characteristics of unsupervised learning. For benchmarking purposes, rather than directly employing the CellChatDB database for candidate gene pairs, we implemented SpatialDM on the identical test sets detailed in our data splitting section. These sets contain both positive and negative ligand-receptor pairs. For every

pair, we utilize SpatialDM to compute the global Moran's R statistic. The weight matrix was determined using a parameter $l = 1.2$, and p-values were estimated via the asymptotic version with the method set to 'z-score'. Subsequently, the AUROC and AUPRC values were computed based on the ranks of the derived p-values.

3.5.7 Tree Calculation

Aligned sequences for all GPCR genes identified in mouse (Mus musculus) were downloaded from GPCRdb[155]. The alignment was restricted to structurally conserved positions to take advantage of generic numbering schemes[156] developed for GPCRs. The gene family tree was calculated using FastTree[157] using default parameters.

Chapter 4

Identification of novel regulators of T cell migration into the CNS using scRNA-seq and CRISPR data

This chapter includes content that is unpublished and in preparation:

> Youjin Lee*, Yutong Wang*, Rosmely Hernandez, Zach Steinhart, Tori Yamamoto, Connor MacLennan, Kunal Puri, Raymond A. Sobel, Yun S. Song†, Alexander Marson†, Elucidating the Biology of T-cell Migration into Tissue Microenvironments during Autoimmune Inflammation using CRISPR
>
> ∗ indicates equal contribution, † indicates co-corresponding authors. The author list is preliminary and subject to future revisions.

4.1 Introduction

Multiple sclerosis (MS) is an autoimmune disorder characterized by the infiltration of immune cells into the central nervous system (CNS), leading to inflammation, blood-brain barrier (BBB) breakdown, and demyelination of neurons [158, 159]. The damage to the myelin sheath, which is essential for efficient electrical impulse transmission along nerve fibers, results in various neurological symptoms [160]. To study the pathogenesis and potential treatments of MS, researchers often employ the experimental autoimmune encephalomyelitis (EAE) model in mice [161]. EAE closely mimics the pathology of MS and can be induced either actively, by immunizing mice with myelin components such as myelin oligodendrocyte glycoprotein (MOG) emulsion combined with pertussis toxin (PT), or passively, by transferring myelin-specific T cells [162]. The progression of EAE can be monitored through the observation of clinical symptoms, allowing for the investigation of disease mechanisms and the evaluation of therapeutic interventions.

The initiation and regulation of immune responses involve complex interactions among antigen-presenting cells (APCs) and T cells [163]. APCs are exposed to the MOG peptide, leading to T cell activation and differentiation into T cell subtypes within the lymph nodes. APCs present antigenic peptides via MHC class II molecules, providing the first activation signal to T cells [164]. The engagement of costimulatory molecules (B7.1 and B7.2) with CD28 on T cells delivers a survival signal, while cytokines such as IL-6, IL-12, and TGF-β direct T cell differentiation into various subtypes [165, 166]. Each T cell subtype, defined by distinct functions, transcription factors and cytokines, plays specialized roles in immune regulation [167]. For example, T helper 1 (Th1) cells, driven by IL-12 and characterized by the transcription factor T-bet, are crucial for mediating responses against intracellular pathogens. T helper 2 (Th2) cells, promoted by IL-4 and regulated by GATA3, typically respond to parasitic infections [168]. T helper 17 (Th17) cells, induced by a combination of IL-6 and TGF-β and regulated by the transcription factor RORγt, are involved in responses against extracellular bacteria and fungi but also in autoimmune pathologies like EAE [169]. Regulatory T cells (Tregs), characterized by the transcription factor Foxp3 and promoted by TGF-β, play a key role in maintaining immune tolerance and preventing autoimmune responses [170, 171]. These activated T cells then transmigrate through the BBB, where they encounter a unique microenvironment that triggers their reactivation and further differentiation.

Different T cell subtypes play distinct roles in the development and regulation of EAE. Th1 cells, characterized by IFNγ production, have been traditionally associated with EAE pathogenesis. However, IFNγ-deficient mice still develop severe disease, suggesting additional pathways contribute to disease severity [172, 173, 174]. In contrast, Th17 cells, which secrete IL-17, are critical for EAE induction, as evidenced by the protection against EAE observed in RORγt knock-out (KO), IL-17R KO, and IL-23R KO mice, which lack essential components of the Th17 signaling pathway [175]. This nuanced interplay of T cell subsets in EAE suggests that targeting specific T cell migration and function could offer new therapeutic strategies for CNS autoimmune diseases, a concept we aim to explore using scRNA-seq and CRISPR KO data to identify new regulators of T cell dynamics in the CNS.

Of the T helper cell subsets, Tregs play a crucial role in maintaining immune tolerance and suppressing inflammation in the CNS during MS and EAE [176]. They express the IL-2 receptor alpha chain (CD25) and play a protective role, as their depletion exacerbates EAE while their transfer ameliorates it [177]. The suppression of autoreactive effector T cells is mediated through various mechanisms that include secretion of immunosuppressive soluble factors (e.g., TGF-β, IL-10, and IL-35) and cell contact mediated regulation via high affinity T cell receptors (TCRs) and other membrane-bound molecules (e.g., CTLA-4, GITR, and LAG-3) that can disrupt activation and proliferation of immune cells [178]. Major efforts are under way to develop new Treg cellular immunotherapies. Recent advances in the use of large-scale in vitro expansion of Tregs followed by in vivo re-infusion raise the possibility that Treg-based immunotherapy may be a highly promising treatment for a variety of autoimmune disorders [179].

The activation and differentiation of myelin-reactive CD4+ T cells into effector (Th1

and Th17), and regulatory subsets, as well as their subsequent transmigration across the BBB into the CNS parenchyma are decisive events in the pathogenesis of MS and EAE. However, our understanding of the mechanisms that drive immunosuppression as well as transmigration into the CNS remains poorly understood. Recent studies have suggested impaired functioning Treg cells derived from MS patients [176]. To realize the full therapeutic potential of these cells, we must first define the gene networks that underpin and coordinate their migration into tissues. In this study, we used scRNA-seq and a targeted in vivo CRISPR KO screen to discover gene programs that drive the migration of Tregs into the CNS during EAE. By utilizing the CRISPR screen, we aim to identify novel regulators that specifically promote Treg infiltration into the CNS, providing insights into the mechanisms underlying T cell dynamics in the context of CNS autoimmunity.

Comparing scRNA-seq profiles of Tregs isolated from the spleen (SP) and CNS revealed upregulation of genes involved in cellular communication and signaling transmission pathways in CNS-infiltrating Tregs. The in vivo pooled CRISPR screen identified regulators of Treg migration into the CNS. Top hits included genes not previously implicated in T cell trafficking, representing potential novel therapeutic targets. Our study focuses on understanding the transcriptional signatures regulating the Treg migration and cell-cell communication within target tissues, with important implications for developing Treg-based immunotherapies for autoimmune disorders and cancer. Understanding the gene regulators guiding Treg homing to sites of inflammation may enable strategies to stabilize and enhance Treg function in specific tissue microenvironments, which could help restore immune homeostasis in autoimmunity while augmenting anti-tumor immunity in cancer.

4.2 Results

4.2.1 scRNA-seq Analysis Reveals Distinct T Cell Transcriptional Signatures in CNS and SP during EAE

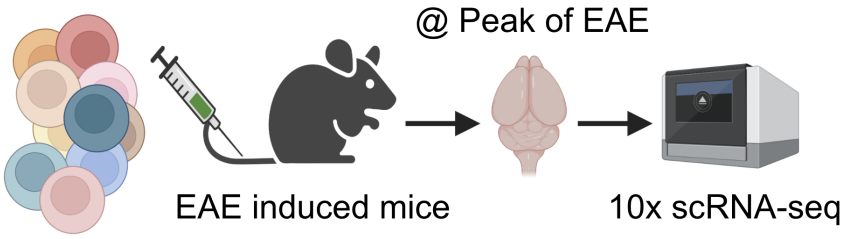


Figure 4.1: Schematic of the scRNA-seq experiment workflow.

To characterize the transcriptional differences between T cells in the CNS and SP during EAE, we performed scRNA-seq analysis on samples collected from both tissues (Fig. 4.1).

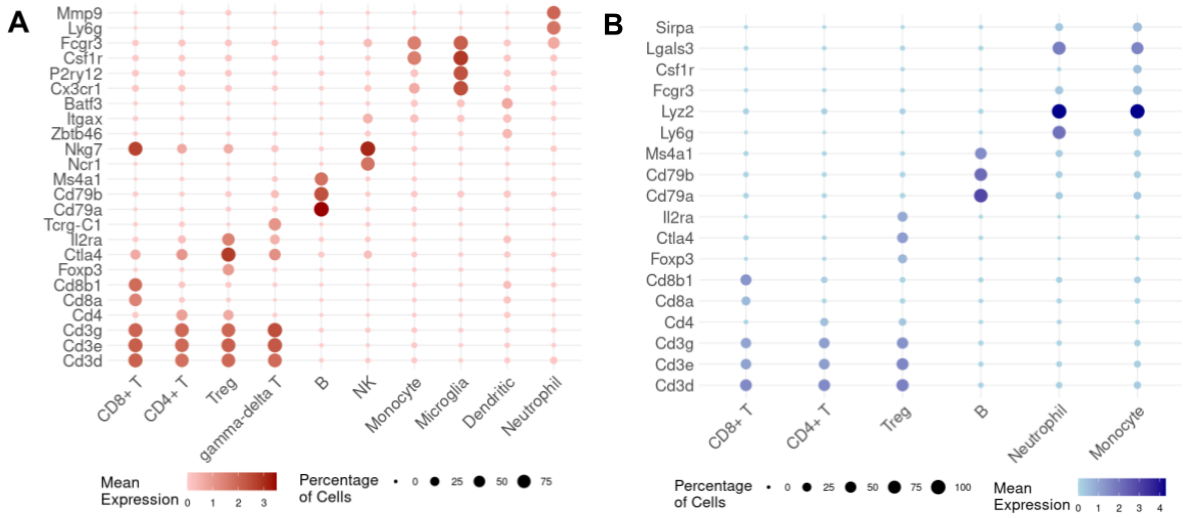


Figure 4.2: Dot plots of cluster by marker gene in CNS (A) and SP (B). Dot size represents percentage of cells expressing the gene, and color represents mean gene expression.

After integrating multiple samples within each tissue to reduce batch effects and clustering cell populations, we annotated multiple immune cell types in the CNS and SP based on the expression of canonical marker genes (Fig. 4.2, Treg: Cd4, Foxp3, Ctla4, and Il2ra).

Differential expression (DE) analysis was performed to investigate the transcriptional differences between cell populations in the CNS and SP, among Tregs (or T effector cells). Specifically, the pseudobulk data were generated by aggregating the raw gene expression counts of single cells per cell type and condition. The pseudobulk count data were normalized and transformed using DESeq2 [180], to account for technical variations and stabilize the variance across samples. Genes with a multiple-testing-corrected p-value less than 0.05 and an absolute $\log 2$ fold change greater than 0.5 were considered statistically significant (Fig. C.16). Among Tregs, 120 genes were more highly expressed in the CNS compared to the SP, while 72 genes were more highly expressed in the SP (Fig. 4.3). In T effector cells, 433 genes were more highly expressed in the CNS, and 269 genes were more highly expressed in the SP (Fig. 4.4). Among T effector cells, Th17-related genes were identified among the genes more highly expressed in the CNS than SP, including Il17a, Il17f, Il17re ($\log 2$ fold changes $= 9.19, 7.61, 4.90$, multiple-testing-corrected p-values $= 4.65 \times 10^{-6}, 6.84 \times 10^{-5}, 3.95 \times 10^{-3}$, respectively). These findings align with the current understanding of the critical role played by Th17 cells and IL-17 signaling in the pathogenesis of MS. The increased expression of Il17a and Il17f suggests a heightened presence and activity of Th17-like cells in the CNS during EAE, while the elevated expression of Il17re indicates an enhanced responsiveness of T effector cells in the CNS to IL-17 signaling. These results underscore the importance of the IL-17 pathway in driving inflammation and tissue damage in the CNS during EAE and provide further support for targeting this pathway as a potential therapeutic strategy in MS.

To gain insights into the biological processes associated with the differentially expressed

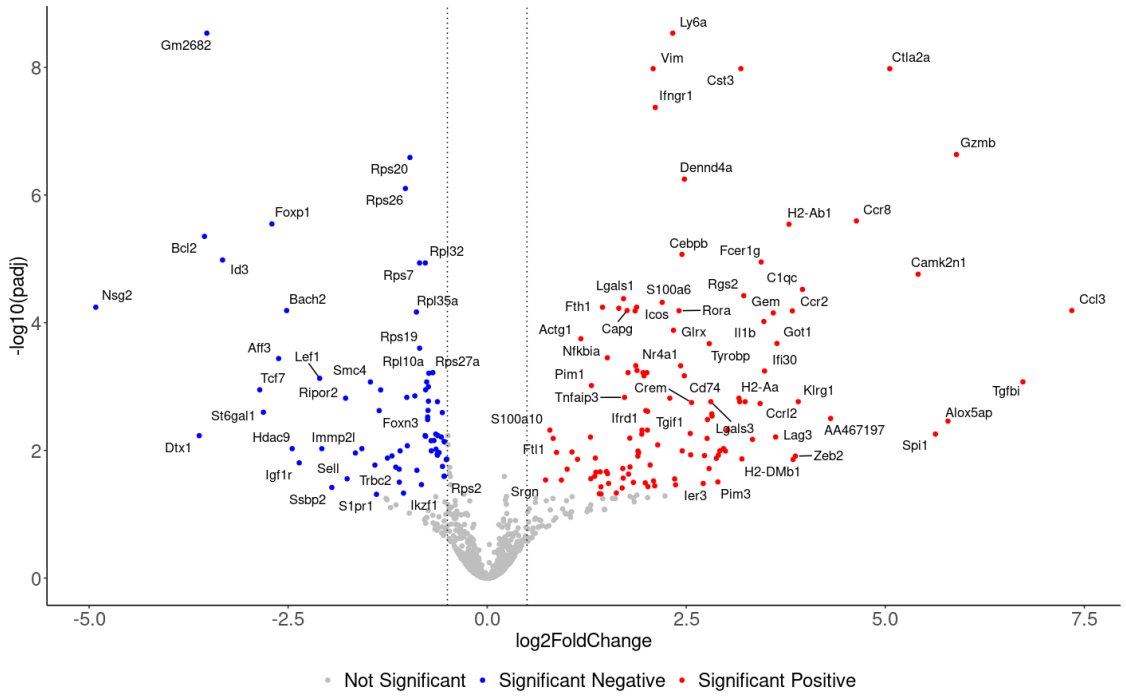


Figure 4.3: Volcano plots highlighting the differentially expressed genes between CNS and SP, among Tregs. Genes in red are more highly expressed in CNS than SP. Genes in blue are more highly expressed in SP than CNS. Genes in gray are not statistically significantly differentially expressed between the two tissue conditions. padj corresponds to the adjusted p-value after multiple testing correction with the Benjamini-Hochberg procedure.

genes, we performed Gene Ontology (GO) enrichment analysis. Genes more highly expressed in the CNS were primarily enriched in GO terms related to immune cell communication and migration, such as chemokine and cytokine binding, cytokine receptor activity, immune receptor activity, and cell adhesion molecule binding. This indicates the importance of these processes in driving T cell-mediated pathology within the target organ. These enriched GO terms were consistently observed in both Tregs (Fig. 4.5A) and T effector cells (Fig. 4.6A), highlighting the critical role of these processes in T cell function within the inflamed CNS. The upregulation of genes involved in chemokine and cytokine signaling suggests increased sensitivity of T cells to local inflammation in the CNS, which may facilitate their migration, recruitment and activation. Moreover, the enrichment of cell adhesion molecule binding implies increased interactions between T cells and the CNS microenvironment, potentially contributing to their retention and pathogenic effects. These findings suggest that cellular communication and signaling transmission play a crucial role in the migration of T cells from the SP to the CNS during EAE. In contrast, genes more highly expressed in the SP were mostly associated with general cellular functions, such as small ribosomal subunit, structural

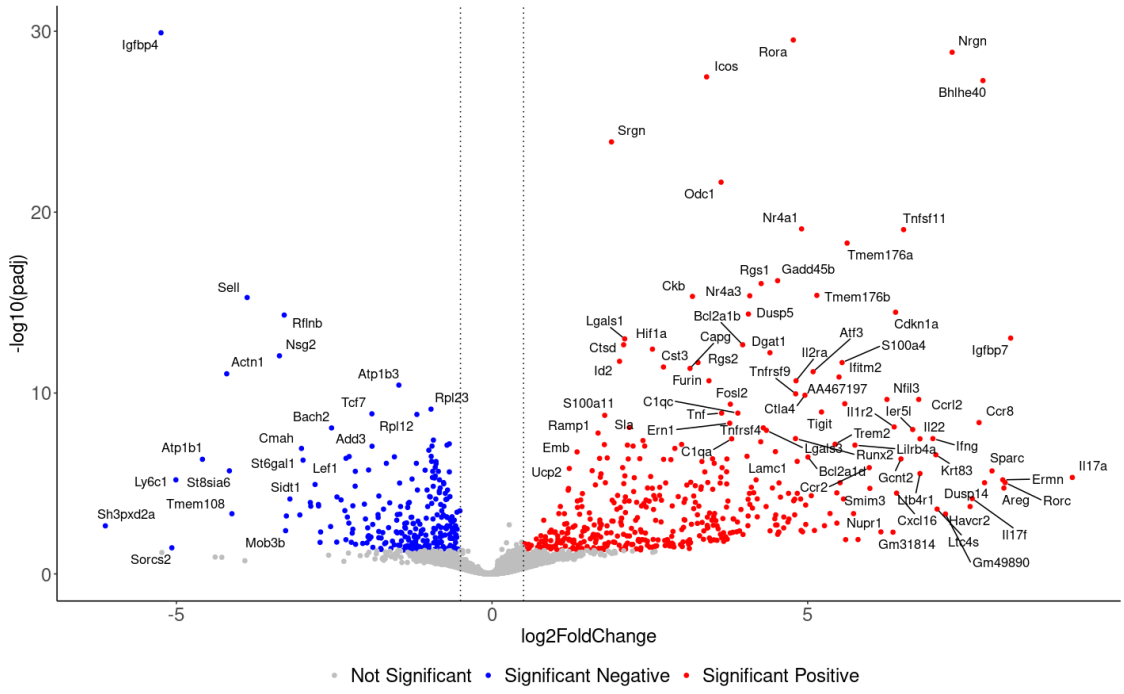


Figure 4.4: Volcano plots highlighting the differentially expressed genes between CNS and SP, among T effector cells. Genes in red are more highly expressed in CNS than SP. Genes in blue are more highly expressed in SP than CNS. Genes in gray are not statistically significantly differentially expressed between the two tissue conditions. padj corresponds to the adjusted p-value after multiple testing correction with the Benjamini-Hochberg procedure.

constituent of ribosome, rRNA binding, and mRNA binding, in both Tregs (Fig. 4.5B) and T effector cells (Fig. 4.6B). These SP-enriched genes suggest that T cells in the peripheral lymphoid organs, such as the spleen, are more engaged in homeostatic processes and protein synthesis, potentially reflecting their less activated state compared to those in the CNS.

To assess the major source of variation in the cell-type-specific transcriptional profiles, we performed Principal Component Analysis (PCA) on the pseudobulk data. The PCA results revealed that the tissue condition (CNS vs. SP) represented the primary source of variation in the first principal component (PC1) for both Treg and T effector cells, explaining 44% and 67% of the variance, respectively (Fig. 4.7). To further investigate the biological significance of this finding, we calculated the gene-specific PC coefficients along PC1 and compared them between differentially expressed (DE) and non-DE genes. Remarkably, the DE genes exhibited significantly more extreme gene-specific coefficients compared to non-DE genes. The Wilcoxon rank sum test p-value is less than 2.2×10^{-16} in all four tests, for both genes more highly expressed in the CNS than in SP, and genes more highly expressed in SP than in CNS, among Tregs and T effector cells, respectively. This consistency between

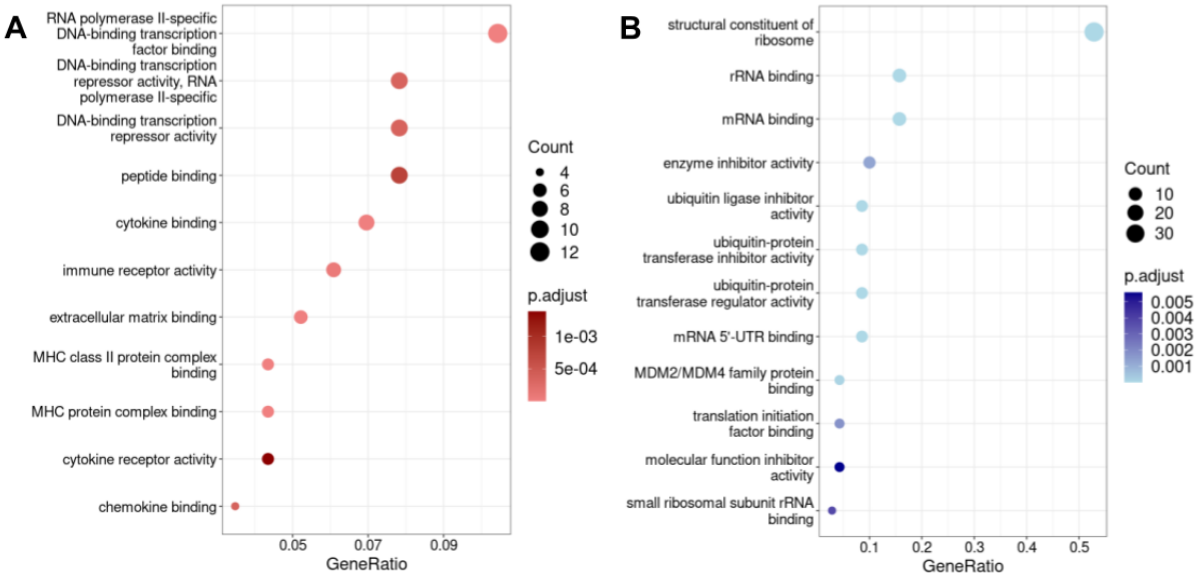


Figure 4.5: GO enrichment terms among Treg. Genes more highly expressed in CNS than SP (A, red). Genes more highly expressed in SP than CNS (B, blue). Each dot represents a GO term, the color indicates the level of enrichment significance. The size of each dot correlates with the gene count involved in that term.

the DE genes identified by DESeq2 and the gene-specific coefficients in PC1 serves as a post-analysis exploratory data analysis (EDA), validating the identified genes as potential targets that explain the major variation in transcriptional signatures between the CNS and SP within the same cell type, Tregs or T effector cells.

4.2.2 Identification of Regulators of Treg Migration Using In Vivo Pooled CRISPR Knockout Screening

Tregs play a crucial role in maintaining immune homeostasis and controlling immune responses. However, their immunosuppressive function can also hinder effective anti-tumor immunity. Conversely, Treg instability, characterized by the loss of the transcription factor Foxp3 and the acquisition of proinflammatory properties, has been shown to promote autoimmunity while potentially enhancing the immune response against tumors. To develop more effective Treg-based therapies for both autoimmune diseases and cancer, a comprehensive understanding of the pathways that regulate Treg migration is essential.

This motivated us to employ a CRISPR-based screening approach to identify regulators for Tregs migration in EAE, aiming to uncover new targets for modulating Treg stability and function. To identify genes that significantly influence Treg migration, we employed MAGeCK [181] to compare the enrichment and depletion of guide RNAs targeting specific genes in the CNS and SP samples. This approach utilizes a negative binomial model to

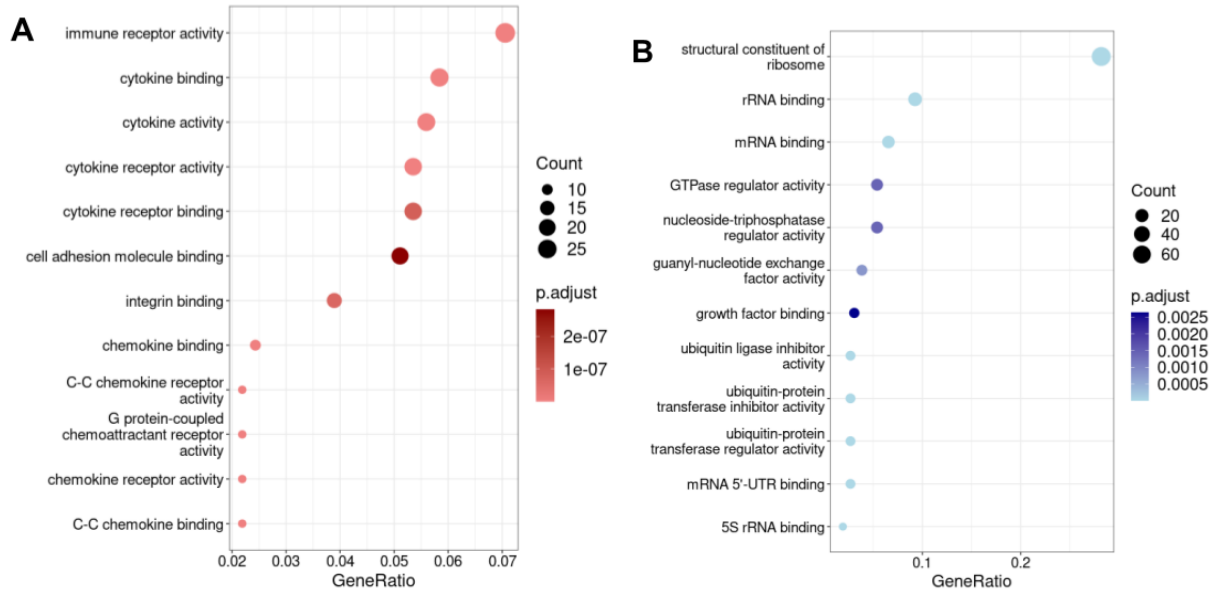


Figure 4.6: GO enrichment terms among T effector cells. Genes more highly expressed in CNS than SP (A, red). Genes more highly expressed in SP than CNS (B, blue). Each dot represents a GO term, the color indicates the level of enrichment significance. The size of each dot correlates with the gene count involved in that term.

estimate the significance of differences in sgRNA abundance between the two conditions, followed by a modified robust ranking aggregation (α-RRA) algorithm to prioritize genes consistently ranked higher than expected (Methods). Our analysis revealed several genes that significantly influence Treg migration from the spleen to the CNS (Fig. 4.8). These CRISPR gene hits were further annotated based on Gene Ontology (GO) terms, including cytokine-mediated signaling pathway (GO:0019221), cell-cell adhesion (GO:0098609), and G protein-coupled receptor activity (GO:0004930) (Fig. 4.9).

Among the identified regulators, we found several genes involved in cytokine-mediated signaling pathways (Fig. 4.9A). These findings suggest that specific cytokines and their associated signaling cascades play a critical role in modulating Treg migration. Additionally, genes related to cell-cell adhesion were also identified as significant regulators (Fig. 4.9B), highlighting the importance of cell-cell interactions in the process of Treg migration. Furthermore, our analysis revealed the involvement of G protein-coupled receptor activity in regulating Treg migration (Fig. 4.9C), indicating that signaling through these receptors may influence Treg behavior and function. To further validate our findings, we visualized the log fold change of sgRNAs targeting the top 5 selected genes (Fig. 4.10). This analysis demonstrated the consistency and robustness of the identified regulators across multiple paired experiments between CNS and SP. By modulating the activity of these regulators, we may be able to fine-tune Treg migration and enhance the efficacy of Treg-based immunotherapies. Further CRISPR activation (CRISPRa) screening and functional studies will be necessary to

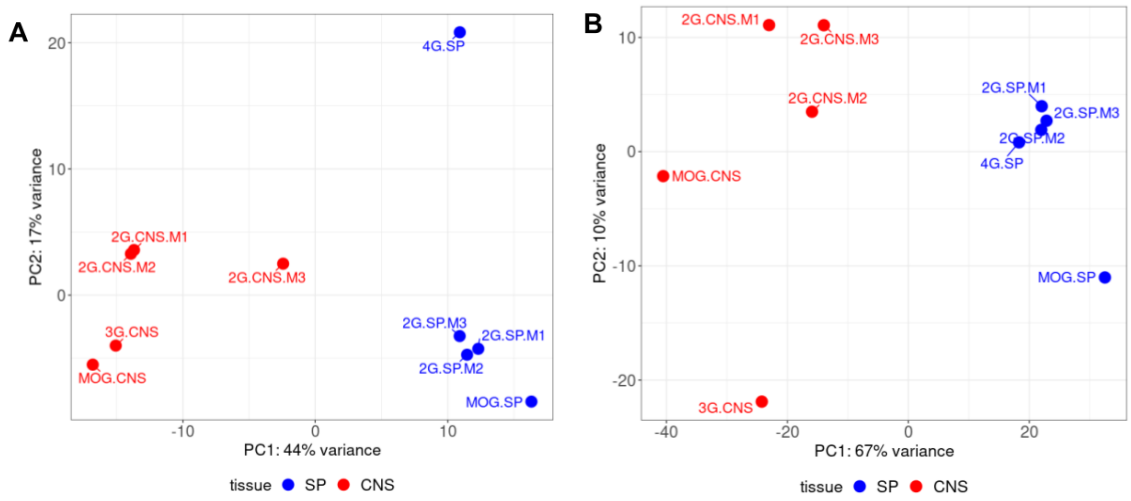


Figure 4.7: Top two PCs of the samples in Tregs (A) and T effector cells (B), colored by the tissue annotation.

validate the roles of these identified genes and elucidate their precise mechanisms of action in the context of Treg biology and immune regulation.

4.3 Discussion

In our study, we employed a CRISPR KO screen to identify genes that influence Treg migration from the spleen to the CNS in the EAE model. While our results suggest that genes whose knockout led to an enrichment of guides in the CNS could be negative regulators of Treg migration, it is important to consider alternative explanations. Compensatory mechanisms, and indirect effects on the immune microenvironment or other cell types could also contribute to the observed changes in Treg behavior. The knockout of a gene may trigger compensatory mechanisms within the cell, leading to changes in the expression or activity of other genes that could influence Treg migration. The knockout of a gene could also affect Treg migration indirectly by altering the immune microenvironment or the behavior of other cell types that interact with Tregs. To confirm the role of these genes as negative regulators, we are currently preparing experiments using CRISPRa and cDNA overexpression to upregulate the expression of candidate genes. If CRISPRa or cDNA overexpression of the candidate genes leads to a reduction in Treg migration to the CNS, it would provide stronger evidence for their role as true negative regulators. Combining CRISPR KO, CRISPRa, and cDNA overexpression experiments will help distinguish genuine regulators from genes with indirect or confounding effects, ultimately providing a more comprehensive understanding of the molecular mechanisms governing Treg migration in the context of EAE.

While our study provides valuable insights into the genes regulating Treg migration, it is important to acknowledge its limitations and discuss future directions. Firstly, our analysis

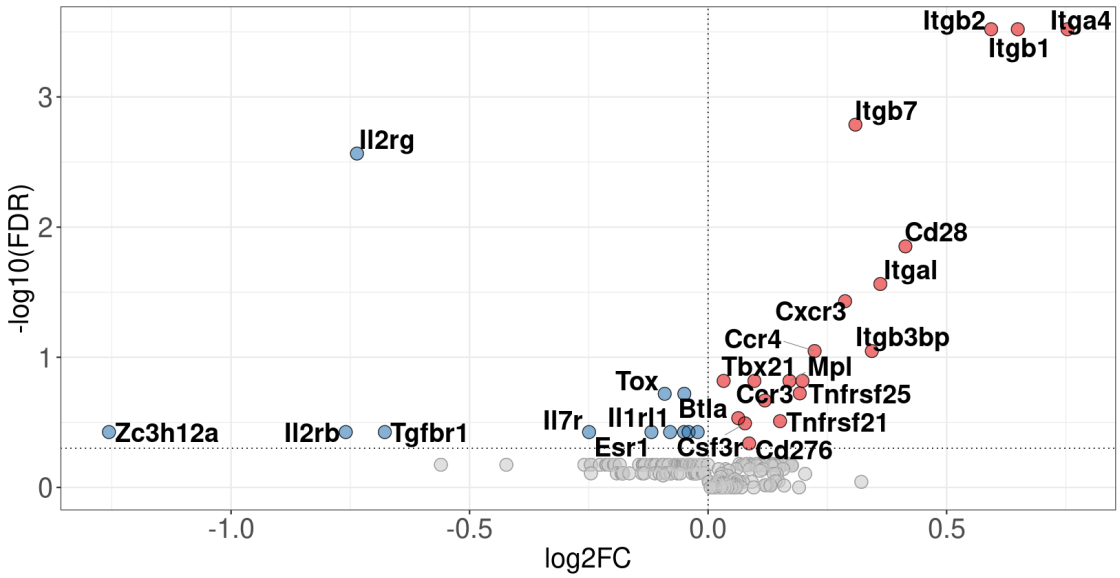


Figure 4.8: Visualization of the positive (red) and negative (blue) regulator genes of Tregs migration from SP to CNS, based on the CRISPR experimental data.log 2FC stands for log 2 fold change, and FDR stands for false discovery rate.

focused on a specific time point, comparing Treg populations in the CNS and spleen. To gain a more comprehensive understanding of the temporal dynamics of Treg migration, future studies should investigate multiple time points throughout the course of EAE. This approach would enable the construction of migration trajectories and help identify genes that regulate Treg trafficking at different stages of the disease. Secondly, our study primarily focused on gene-level perturbations. However, investigating the effects of sequence variant-level mutations, such as those in regulatory regions or coding regions, could provide a more nuanced understanding of the molecular mechanisms governing Treg migration. Targeted genome editing techniques, such as CRISPR-based base editors or prime editors, could be employed to introduce specific mutations and assess their impact on Treg function and migration.

Tregs play a crucial role in maintaining immune homeostasis and preventing autoimmunity, thus Treg biology is important in immunotherapy. In the context of cancer immunotherapy, Tregs can suppress anti-tumor immune responses, limiting the effectiveness of treatments such as checkpoint inhibitors or adoptive cell therapies. On the other hand, in autoimmune disorders like multiple sclerosis, enhancing Treg function could help reduce harmful immune responses and alleviate disease severity. Our study identifies genes that regulate Treg migration to the CNS in the EAE model, which could inform the development of targeted therapies aimed at modulating Treg trafficking and function. For example, inhibiting negative regulators of Treg migration could enhance their immunosuppressive activity

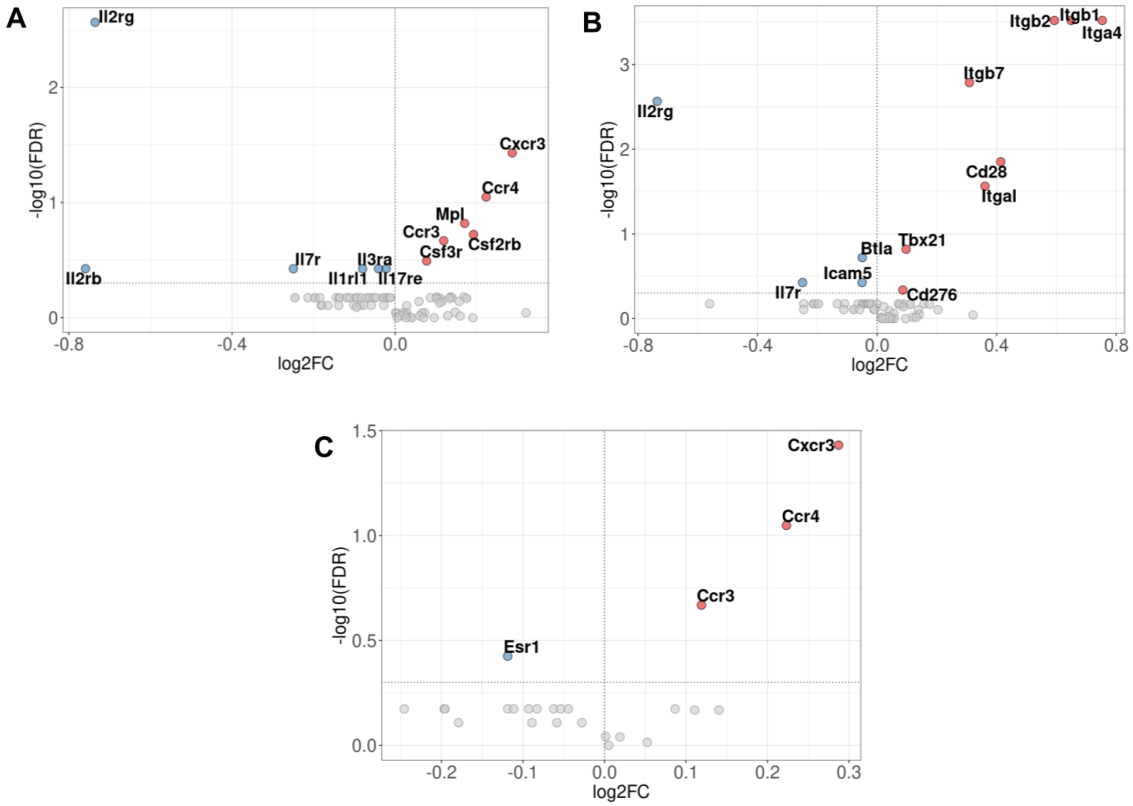


Figure 4.9: Stratified annotations of the CRISPR gene hits in the volcano plot. (A) Genes in the cytokine-mediated signaling pathway. (B) Genes involved in the cell-cell adhesion. (C) Genes related to the G protein-coupled receptor activity.

in the CNS, potentially reducing neuroinflammation and disease progression. Furthermore, our findings could be relevant for other autoimmune disorders characterized by abnormal Treg migration or function, such as rheumatoid arthritis or inflammatory bowel disease. By better understanding the molecular mechanisms governing Treg migration, we can develop more effective and targeted immunotherapies for a range of autoimmune and inflammatory conditions.

4.4 Methods

4.4.1 10X Data Processing

Mice from the same cell line were immunized with myelin oligodendrocyte glycoprotein (MOG) peptide, a myelin protein found in the brain and spinal cord, to induce experimental autoimmune encephalomyelitis (EAE), a widely used animal model for multiple sclerosis. Raw sequencing data were processed using the bcl2fastq Conversion Software (Illumina) and

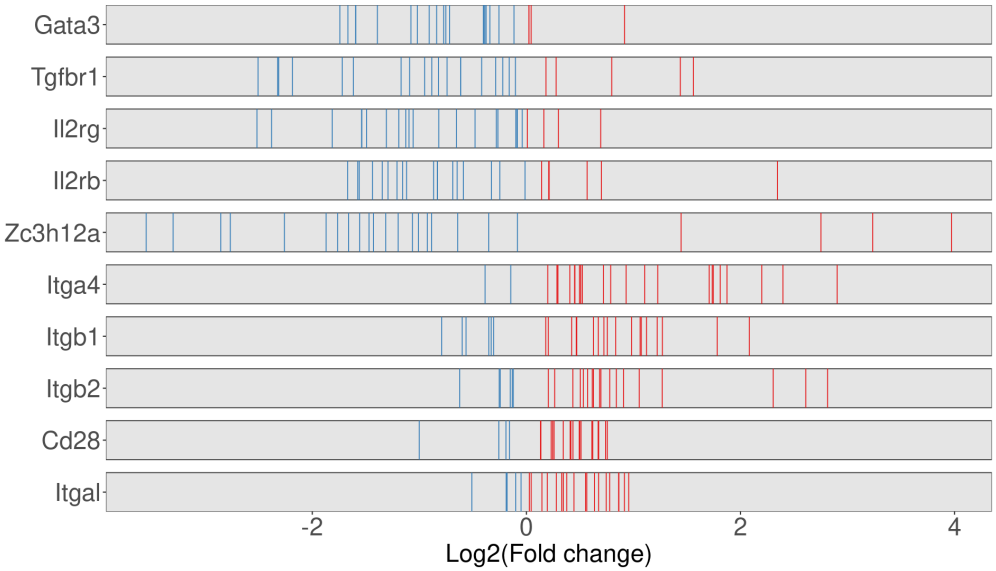


Figure 4.10: Visualization of the log fold change of sgRNAs targeting top 5 selected genes. Each row corresponds to one gene, and each colored line corresponds to one sgRNA for that gene, in one paired experiment between CNS and SP. Red corresponds to positive regulators, and blue corresponds to negative regulators.

the Cell Ranger pipeline (10x Genomics). The bcl2fastq software was used to convert BCL files to FASTQ files, with the use-bases-mask parameter set to Y26, I10, I10, Y90. This configuration specifies that the first 26 bases of Read 1 contain the cell barcode and Unique Molecular Identifier (UMI), the first 10 bases of the Index 1 and Index 2 reads contain the sample indices (not used for cell barcodes or UMIs), and the 90 bases of Read 2 contain the cDNA insert (gene transcript). The resulting FASTQ files were then processed using the Cell Ranger count pipeline, which performs alignment, filtering, barcode counting, and UMI counting. The pipeline was run with default parameters, using the mouse reference genome. The output of the Cell Ranger pipeline includes a gene-barcode matrix, which contains the counts of unique UMIs for each gene in each cell.

4.4.2 Demultiplexing Analysis with Hashtag-Oligo

scRNA-seq data were generated from both the CNS and spleen (SP) tissues across different mice to investigate the cellular and molecular changes associated with EAE. To increase the throughput and reduce costs associated with sequencing multiple samples separately, we employed Hashtag-Oligo (HTO) multiplexing to pool data from different samples (tissues and mice) into a single cell suspension. The TotalSeq™-A0316 anti-mouse Hashtag 16 Antibody, which is conjugated with a unique oligonucleotide barcode (the 'hashtag'), was used to tag

each cell with a unique identifier corresponding to its sample of origin. We used TotalSeq-A for this study, which is compatible with the 5' gene expression kit. After sequencing, these hashtag sequences were used to demultiplex the pooled data and assign each cell to its sample of origin. The demultiplexing process was performed using the scanpy.external.pp.hashsolo function in scanpy [182], based on a probabilistic demultiplexing method in HashSolo [183].

4.4.3 Preprocessing scRNA-seq Data

The UMI count matrix was loaded using the Read10X function in Seurat. Features detected in fewer than 3 cells were filtered out, and cells with at least 200 detected features were retained for further analysis. Cells with reads from multiple hashtags were annotated as 'doublet' and removed from the downstream data analysis. Multiple batches of scRNA-seq data were collected and processed separately (Fig. C.1, C.2). For the first batch (2G), which included samples from both the CNS and SP and different mice (M1, M2, M3), cells were filtered based on the following criteria: the number of features per cell greater than 100 and less than $2,000$, and the percentage of mitochondrial genes per cell less than 5%. The second batch, containing only CNS samples (3G_CNS), was filtered to include cells with the number of features per cell greater than 100 and less than $4,000$, and the percentage of mitochondrial genes per cell less than 10%. The third batch, consisting of SP samples (4G_SP), was filtered to retain cells with the number of features per cell greater than 100 and less than $2,000$, and the percentage of mitochondrial genes per cell less than 10%. In the fourth batch, which included CNS samples (MOG_CNS), cells were filtered based on the number of features per cell greater than 100 and less than $6,000$, and the percentage of mitochondrial genes per cell less than 10%. The final batch, containing SP samples (MOG_SP), was filtered to include cells with the number of features per cell greater than 100 and less than $3,000$, and the percentage of mitochondrial genes per cell less than 5%.

For each tissue-specific batch, the data were log-normalized with a scaling factor of 1×10^5 using the NormalizeData function in Seurat. Subsequently, highly variable genes were identified using the FindVariableFeatures function with the 'vst' method, selecting the top $2,000$ features.

4.4.4 Cell Type Annotation

Batch integration was performed separately on the CNS and SP samples to account for biological differences, such as cell type composition heterogeneity between the tissues. The integration process was carried out using the Seurat package in R. First, integration anchors were identified using the FindIntegrationAnchors function, which applies canonical correlation analysis (CCA) to identify shared sources of variation between the datasets. The first 50 dimensions of the CCA space were used to compute the integration anchors. Subsequently, the datasets were integrated using the IntegrateData function, which applies a scaling normalization to the datasets, allowing them to be jointly analyzed while reducing batch effects. The integrated data were then scaled using the ScaleData function to stan-

dardize the gene expression values across cells. To reduce the dimensionality of the dataset and capture the main sources of variation, principal component analysis (PCA) was performed on the integrated and scaled data using the RunPCA function. Finally, the t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm was applied to the integrated data using the RunTSNE function for visualization purposes only, utilizing the first 50 principal components.

For downstream analysis, we excluded two highly expressed non-coding RNA genes (Gm42418 and Malat1) across all cells, as they may not contribute to distinguishing different cell types or states. After filtering, there are 19,322 genes by 8,868 cells in CNS, and 18,420 genes by 57,050 cells in SP. To identify cell clusters, we applied the Louvain clustering algorithm with multilevel refinement using the FindNeighbors and FindClusters functions in Seurat. We applied a resolution parameter of 2 for the first level of clustering, resulting in 40 global clusters in CNS (Fig. C.3A), and 43 global clusters in SP (Fig. C.4A). To remove potential multiplets, we excluded cells with the top 2% gene counts within each cluster, because different cell populations may have varying average number of genes detected. Subsequently, we performed a second level of clustering within each global cluster using a resolution parameter of 1 in CNS, and 0.5 in SP, to further identify potential cell populations with mixed cell types. This second level of clustering leads to a total of 146 level 2 clusters (Fig. C.3B), and 121 global level 2 clusters in SP (Fig. C.4B).

We then annotated the cell types of each level 2 cluster based on the expression of 120 known marker genes (Table C.1, [184]). To visualize the expression patterns of marker genes across level 2 clusters, we generated a heatmap using the 'pheatmap' package in R (Fig. C.5, C.6). First, we calculated the percentage of cells expressing each marker gene within each level 2 cluster, resulting in a matrix where rows represent clusters and columns represent marker genes. The heatmap was generated using correlation as the clustering distance metric for both rows and columns, and complete linkage as the clustering method. We filtered out the level 2 clusters which have a high expression of marker genes from multiple cell types (such as B, T, or myeloids), because those clusters could be mixtures of different cell types. For CNS, the resulting scRNA-seq data includes 19,322 genes by 8,000 cells, with cell type annotation results in Fig. C.7A and key marker gene expression patterns in Figures C.8 and C.9. For SP, this leads to 18,420 genes by 34,475 cells, The cell type annotation results are shown in Fig. C.7B, and the key marker gene expression patterns are shown in Fig. C.10, C.11. Cells annotated as CD4+ T cells correspond to the cells with a high expression of CD4, but few expression of Treg marker genes (e.g. Foxp3, Il2ra, Ctla4). These CD4+ T cells correspond to T effector cells in study.

4.4.5 Cell-Type-Specific DE Analysis between Conditions (Tissues)

Conventional differential expression (DE) tests take each single cell as an independent sample, without accounting for the association across cells in the same biological sample or

condition. Directly performing such tests between conditions, without pseudobulk or mixed modeling, would lead to a large number of false positives [185, 186, 187]. In our analysis, differential gene expression analysis between conditions (tissues) was performed using a pseudobulk aggregation approach to account for within-sample correlations. Such analysis was implemented among Treg and T effector cell populations, respectively. It is worth noting that the raw gene expression count matrix is used for this part, without batch integration nor normalization. Batch integration is only helpful for cell-based analysis, such as trajectory inference, not gene-based analysis. Batch integration across conditions introduces dependence between samples in different conditions, which violates the independence assumptions in the DE statistical test.

For each gene, raw gene expression counts were aggregated by summing the counts of single cells per cell type and condition. The aggregated count data were normalized using the median of ratios method to account for the library depth difference, followed by a regularized log transformation using the 'rlog' function in DESeq2 [180] in R. This stabilized the variance while accounting for the variance-mean dependence. Sample-level PCA was performed on the transformed data to assess whether the tissue condition represents the major source of variation in the data, using the top 500 most variable genes. The results were consistent regardless of the chosen number of top variable genes. Both Treg and T effector samples were separately by the tissue condition on PC1, which explains 44% variance in Tregs, and 67% variance T effector cells respectively (Fig. 4.7). As quality control, samples are also annotated by the number of single-cell level UMI counts per sample (Fig. C.12), and the total sum of transformed data per sample (Fig. C.13). Hierarchical clustering was also performed across different samples. The clustering pattern highlights the difference between tissue conditions in CNS and SP (Fig. C.14).

DESeq2 was employed to model the normalized counts, estimate gene-wise dispersions, and shrink these estimates for more accurate dispersion modeling. The dispersion estimates against the mean of normalized counts ensured that the data followed the expected mean-variance relationship (Fig. C.15). To improve the accuracy of the log2 fold change estimates, the apeglm method [188] was employed using the 'lfcShrink' function. This method applies a heavy-tailed prior distribution to the log2 fold changes, which helps to remove noise and preserve large differences in gene expression. Finally, the negative binomial model was fitted, and the Wald test was conducted to test for DE. The resulting p-values were adjusted using Benjamini-Hochberg (BH) procedure to control the false discovery rate. The enriched GO terms were associated with the gene set, mapped from gene symbols to ENTREZ IDs using the org.Mm.eg.db database. To further investigate the biological significance of the differentially expressed genes, we compiled a list of cytokines, their corresponding signaling receptor genes [184], and genes associated with the GO term 'cell surface'. Notably, approximately 25% of the DE genes more highly expressed in the CNS compared to the SP were present in this list (30 out of 120 in Tregs, and 105 out of 433 in T effector cells; Fig. C.17). Meanwhile, only 3 out of 72 genes more highly expressed in the SP compared to the CNS in Tregs, and 18 out of 269 genes in T effector cells, were found in the list, serving as a negative control.

4.4.6 Difference in the Number of DE Genes between T Effector Cells and Tregs

We noticed that there are many more DE genes among T effector cells compared to Tregs. In the CNS, there are 433 DE genes more highly expressed than in the SP among T effector cells, and 120 DE genes among Tregs. Conversely, in the SP, there are 269 DE genes more highly expressed than in the CNS among T effector cells, and 72 DE genes among Tregs. This difference could partially be attributed to the cell number variation across samples from different tissues (CNS vs. SP). In the SP, the higher number of single cells annotated as T effector cells ($3,331$ in total) compared to Tregs (364 in total) could lead to increased statistical power to detect DE genes in T effector cells. In the CNS, 347 T effector cells and 521 Tregs were sequenced.

To investigate if this technical effect is driving the difference in the number of DE genes, we randomly subsampled 400 T effector cells in the SP and performed downstream CNS-vs-SP DE analysis among T effector cells. Pseudobulk DE was implemented using the subsampled cell population, identifying 197 genes more highly expressed in the CNS than SP and 111 genes more highly expressed in the SP than CNS among T effector cells (Fig. C.18). Comparing the DE results between the original and subsampled data revealed that the DE genes from the subsampled data were mostly covered by the DE genes from the original data. Specifically, 193 out of 197 genes more highly expressed in the CNS than SP and 110 out of 111 genes more highly expressed in the SP than CNS were identified in the original data. This confirmed that the cell number difference is a major contributor to the difference between the number of DE genes in T effector cells and Tregs. After subsampling, there are still more DE genes among T effector cells than Tregs. This may be due to the more specialized function of Tregs in maintaining immune tolerance, while T effector cells have a broader range of functions and can differentiate into various subsets (e.g., Th1, Th2, Th17). This functional diversity may contribute to a larger number of DE genes in T effector cells.

4.4.7 Interpretation of In Vivo CRISPR Loss-of-Function Screen

In a screening experiment, sgRNAs are designed, synthesized and cloned into a lentivirus library, which is subsequently transduced into cells at a low multiplicity of infection to ensure that only one sgRNA copy is integrated per cell. Our in vivo CRISPR loss-of-function screen began with a population of cells, where each cell contained a different single guide RNA (sgRNA) from our sgRNA library. These sgRNAs caused the knockout of different genes in each cell. We then exposed these cells to two different experimental conditions: CNS as the treatment group and spleen as the control group, and allowed them to grow and proliferate over time. At the conclusion of the experiment, we collected cells from both the CNS and spleen, then sequenced the sgRNAs to measure their representation in each population.

In a CRISPR loss-of-function screen, the concept of positive and negative selection plays a crucial role in identifying genes that influence cell growth or survival. When a gene is knocked out, and the corresponding cells exhibit reduced proliferation or increased cell death

compared to the control group, the gene is said to be under negative selection. This indicates that the gene is essential for maintaining normal cell growth and survival. In contrast, when a gene knockout leads to enhanced cell proliferation or survival, the gene is considered to be under positive selection, suggesting that it may have a suppressive effect on cell growth under normal conditions.

We hypothesized that genes whose knockout led to an enrichment of guides in the CNS were negative regulators of Treg migration into the CNS. This is because the knockout of these genes allowed Tregs to more easily migrate into the CNS, leading to an increased representation of the corresponding guides in the CNS population. On the other hand, genes whose knockout led to an enrichment of guides in the spleen were considered positive regulators of Treg migration into the CNS. The knockout of these genes reduced the ability of Tregs to migrate into the CNS, causing them to remain in the spleen and resulting in an increased representation of the corresponding guides in the spleen population. Additionally, genes whose knockout led to a depletion of guides in the spleen were also considered negative regulators of Treg migration into the CNS, as the knockout of these genes facilitated Treg migration out of the spleen and into the CNS.

4.4.8 Preprocessing CRISPR Data

Raw sequencing data from the CRISPR screen was initially processed using bcl2fastq software to convert base call files into FASTQ format. Several key parameters were specified to ensure high quality reads were obtained, including using both the template bases and index reads (use-bases-mask Y∗,I∗), generating separate FASTQ files for index reads (create-fastq-for-index-reads), removing any reads shorter than 35 bases after adapter trimming (minimum-trimmed-read-length 35 –mask-short-adapter-reads 35), only using data from specific tiles of the flow cell (tiles s_1, s_2, s_3, s_4), and allowing for one mismatch in the sample barcodes (barcode-mismatches 1). The resulting FASTQ files were then further processed using the MAGeCK software package [181] in Linux. First, the sequencing reads were mapped to the reference sgRNA library using the 'mageck count' command, together with a list of the control sgRNA sequences. The mapping ratio for each sample in the experiment was calculated for quality control (Fig. C.19).

4.4.9 Identification of CRISPR Gene Hits

To identify sgRNAs that were significantly enriched or depleted in the treatment group (CNS samples) compared to the control group (SP samples), the 'mageck test' command in MAGeCK was utilized. First, read counts from different samples were median-normalized to account for the difference in library sizes and read count distributions. Next, a negative binomial (NB) model was applied to estimate the variance of read counts and determine whether the sgRNA abundance differ significantly between the treatment and control conditions. The NB model assumes that the variance is some smooth function of the mean, which can be fitted using the observed mean and variance values of all the sgRNAs. To de-

cide whether each sgRNA has a significant impact on cell growth or survival, test statistics are calculated under the null hypothesis that there is no difference between the treatment and control conditions. However, the distribution of the test statistic differs depending on whether the gene is under positive or negative selection. This is because the read counts for each sgRNA in the treatment and control groups are assumed to follow a NB distribution, which is asymmetric. Consequently, the probability of observing a lower read count than expected in the treatment condition (i.e., $x < \mu_{ib}$ given $\mu_{ib} < \mu_{ia}$) is different from the probability of observing a higher read count than expected (i.e., $x > \mu_{ib}$ given $\mu_{ib} > \mu_{ia}$).

To identify genes with a significant fraction of sgRNAs ranked closer to the top of the sgRNA list (sorted by NB p-values), the α-RRA algorithm was utilized in MAGeCK. It assumes that if a gene has no effect on selection, sgRNAs targeting this gene should be uniformly distributed across the ranked list. α-RRA ranks genes by comparing the skew in rankings to the uniform null model and prioritizes genes whose sgRNA rankings are consistently higher than expected. Finally, a permutation test was performed to compute p-values, where sgRNAs were randomly assigned to genes while controlling for the number of sgRNAs targeting each gene.

Bibliography

[1] Byungjin Hwang, Jong Hoon Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental and Molecular Medicine* 50 (2018), pp. 1–14. DOI: 10.1038/s12276-018-0071-8.

[2] Vivien Marx. "Method of the Year: spatially resolved transcriptomics". In: *Nature Methods* 18 (2021), pp. 9–14. DOI: 10.1038/s41592-020-01033-y.

[3] Lior Moses and Lior Pachter. "Museum of spatial transcriptomics". In: *Nature Methods* 19 (2022), pp. 534–546. DOI: 10.1038/s41592-022-01409-2.

[4] L. Atta and J. Fan. "Computational challenges and opportunities in spatially resolved transcriptomic data analysis". In: *Nature Communications* 12 (2021), p. 5283. DOI: 10.1038/s41467-021-25557-9.

[5] Kok Hao Chen et al. "Spatially resolved, highly multiplexed RNA profiling in single cells". In: *Science* 348 (2015), aaa6090. DOI: 10.1126/science.aaa6090.

[6] Chee-Huat Linus Eng et al. "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". In: *Nature* 568 (2019), pp. 235–239. DOI: 10.1038/s41586-019-1049-y.

[7] Je Hyuk Lee et al. "Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues". In: *Nature Protocols* 10 (2015), pp. 442–458. DOI: 10.1038/nprot.2014.191.

[8] A. Rao, D. Barkley, G. S. França, et al. "Exploring tissue architecture using spatial transcriptomics". In: *Nature* 596 (2021), pp. 211–220. DOI: 10.1038/s41586-021-03634-9. URL: https://doi-org.libproxy.berkeley.edu/10.1038/s41586-021-03634-9.

[9] Samuel G. Rodriques et al. "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution". In: *Science* 363 (2019), pp. 1463–1467. DOI: 10.1126/science.aaw1219.

[10] Robert R. Stickels, Evan Murray, Purushothama Kumar, et al. "Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2". In: *Nature Biotechnology* 39 (2021), pp. 313–319. DOI: 10.1038/s41587-020-0739-1.

[11] Sanja Vickovic et al. "High-definition spatial transcriptomics for in situ tissue profiling". In: *Nature Methods* 16.10 (Oct. 2019). Epub 2019 Sep 9. PMID: 31501547; PMCID: PMC6765407, pp. 987–990. DOI: 10.1038/s41592-019-0548-y.

[12] Y. Liu et al. "High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue". In: *Cell* 183.6 (Dec. 2020). Epub 2020 Nov 13. PMID: 33188776; PMCID: PMC7736559, 1665–1681.e18. DOI: 10.1016/j.cell.2020.10.026.

[13] Ao Chen et al. "Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays". In: *Cell* 185.10 (May 2022). Open Access, 1777–1792.E21. DOI: 10.1016/j.cell.2022.04.003. URL: https://www.cell.com/cell/fulltext/S0092-8674(22)00399-3.

[14] H. Duan, T. Cheng, and H. Cheng. "Spatially resolved transcriptomics: advances and applications". In: *Blood Science* 5.1 (Nov. 2022). PMID: 36742187; PMCID: PMC9891446, pp. 1–14. DOI: 10.1097/BS9.0000000000000141.

[15] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13 (2021), pp. 3573–3587. DOI: 10.1016/j.cell.2021.04.048.

[16] Dario Righelli et al. "SpatialExperiment: infrastructure for spatially resolved transcriptomics data in R using Bioconductor". In: *Genome Biology* 23.1 (2022), pp. 1–27. DOI: 10.1186/s13059-022-02640-y.

[17] Robert A Amezquita et al. "Orchestrating single-cell analysis with Bioconductor". In: *Nature Methods* 17.2 (2020), pp. 137–145. DOI: 10.1038/s41592-019-0654-x.

[18] Robin Dries et al. "Giotto: a toolbox for integrative analysis and visualization of spatial expression data". In: *Genome Biology* 22.1 (2021), pp. 1–31. DOI: 10.1186/s13059-021-02286-2.

[19] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (2018), pp. 1–5. DOI: 10.1186/s13059-017-1382-0.

[20] Giovanni Palla et al. "Squidpy: a scalable framework for spatial omics analysis". In: *Nature Methods* 19.2 (2022), pp. 171–178. DOI: 10.1038/s41592-021-01358-2.

[21] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15.12 (2018), pp. 1053–1058. DOI: 10.1038/s41592-018-0229-2.

[22] Joseph Bergenstråhle, Ludvig Larsson, and Joakim Lundeberg. "Seamless integration of image and molecular analysis for spatial transcriptomics workflows". In: *BMC Genomics* 21.1 (2020), pp. 1–7. DOI: 10.1186/s12864-020-07146-0.

[23] Duy Pham et al. "stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues". In: *bioRxiv* (2020). DOI: 10.1101/2020.05.31.125658.

[24] L. Marconato, G. Palla, K.A. Yamauchi, et al. "SpatialData: an open and universal data framework for spatial omics". In: *Nature Methods* (2024). DOI: 10.1038/s41592-024-02212-x.

[25] B. Li, W. Zhang, C. Guo, et al. "Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution". In: *Nature Methods* 19 (2022), pp. 662–670. DOI: 10.1038/s41592-022-01480-9.

[26] Alma Andersson et al. "Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography". In: *Communications Biology* 3.1 (2020), p. 565. DOI: 10.1038/s42003-020-01247-y.

[27] Vitalii Kleshchevnikov et al. "Cell2location maps fine-grained cell types in spatial transcriptomics". In: *Nature Biotechnology* 40.5 (2022), pp. 661–671. DOI: 10.1038/s41587-021-01139-4.

[28] Dylan M Cable et al. "Robust decomposition of cell type mixtures in spatial transcriptomics". In: *Nature Biotechnology* 40.4 (2022), pp. 517–526. DOI: 10.1038/s41587-021-00830-w.

[29] Samuel G Rodriques et al. "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution". In: *Science* 363.6434 (2019), pp. 1463–1467.

[30] Marc Elosua-Bayes et al. "SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes". In: *Nucleic Acids Research* 49.9 (2021), e50. DOI: 10.1093/nar/gkab043.

[31] Dongqing Sun et al. "STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing". In: *Nucleic Acids Research* 50.7 (Apr. 2022), e42. DOI: 10.1093/nar/gkac150.

[32] Rui Dong and Guo-Cheng Yuan. "SpatialDWLS: accurate deconvolution of spatial transcriptomic data". In: *Genome Biology* 22 (2021), p. 145. DOI: 10.1186/s13059-021-02362-7.

[33] Benjamin F. Miller, Fei Huang, Lennart Atta, et al. "Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data". In: *Nature Communications* 13.1 (2022), p. 2339. DOI: 10.1038/s41467-022-30033-z.

[34] Yiming Ma and Xiang Zhou. "Spatially informed cell-type deconvolution for spatial transcriptomics". In: *Nature Biotechnology* 40 (2022), pp. 1349–1359. DOI: 10.1038/s41587-022-01273-7.

[35] Noah F Greenwald, Gregory Miller, Erick Moen, et al. "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning". In: *Nature Biotechnology* 40.4 (2022), pp. 555–565. DOI: 10.1038/s41587-021-01094-0.

[36] Carsen Stringer, Michalis Wang, Marius Michaelos, et al. "Cellpose: a generalist algorithm for cellular segmentation". In: *Nature Methods* 18.1 (2021), pp. 100–106. DOI: 10.1038/s41592-020-01018-x.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[38] Xin Fu, Yi Lin, Dahua Michael Lin, et al. "BIDCell: Biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data". In: *Nature Communications* 15.1 (2024), p. 509. DOI: 10.1038/s41467-023-44560-w.

[39] Vasiliy Petukhov, Richard J Xu, Ruslan A Soldatov, et al. "Cell segmentation in imaging-based spatial transcriptomics". In: *Nature Biotechnology* 40.3 (2022), pp. 345–354. DOI: 10.1038/s41587-021-01044-w.

[40] Yijin Wang, Wending Wang, Dongfang Liu, et al. "GeneSegNet: a deep learning framework for cell segmentation by integrating gene expression and imaging". In: *Genome Biology* 24.1 (2023), p. 235. DOI: 10.1186/s13059-023-03054-0.

[41] Hao Chen, Dongsheng Li, and Ziv Bar-Joseph. "SCS: cell segmentation for high-resolution spatial transcriptomics". In: *Nature Methods* 20.9 (2023), pp. 1237–1243. DOI: 10.1038/s41592-023-01939-3.

[42] Cheng Chen, Hyo Jeong Kim, and Pengyi Yang. "Evaluating spatially variable gene detection methods for spatial transcriptomics data". In: *Genome Biology* 25.1 (2024), p. 18. DOI: 10.1186/s13059-023-03145-y.

[43] Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. "SpatialDE: identification of spatially variable genes". In: *Nature Methods* 15.5 (2018), pp. 343–346. DOI: 10.1038/nmeth.4636.

[44] Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. "Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies". In: *Nature Methods* 17.2 (2020), pp. 193–200. DOI: 10.1038/s41592-019-0701-7.

[45] Brian Miller et al. "Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomics data with nonuniform cellular densities". In: *Genome Research* (May 2021).

[46] Jiaqiang Zhu, Shiquan Sun, and Xiang Zhou. "SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies". In: *Genome Biology* 22.1 (2021), p. 184. DOI: 10.1186/s13059-021-02404-0.

[47] Daniel Edsgärd, Per Johnsson, and Rickard Sandberg. "Identification of spatial expression trends in single-cell gene expression data". In: *Nature Methods* 15.5 (2018), pp. 339–342. DOI: 10.1038/nmeth.4634.

[48] Lukas M Weber et al. "nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes". In: *Nature Communications* 14.1 (2023), p. 4059. DOI: 10.1038/s41467-023-39748-z.

[49] Qiwei Li et al. "Bayesian modeling of spatial molecular profiling data via Gaussian process". In: *Bioinformatics* 37.22 (2021), pp. 4129–4136. DOI: 10.1093/bioinformatics/btab455.

[50] Kai Zhang, Wennan Feng, and Peng Wang. "Identification of spatially variable genes with graph cuts". In: *Nature Communications* 13.1 (2022), p. 5488. DOI: 10.1038/s41467-022-33182-3.

[51] Yuchen Liang et al. "PROST: quantitative identification of spatially variable genes and domain detection in spatial transcriptomics". In: *Nature Communications* 15.1 (2024), p. 600. DOI: 10.1038/s41467-024-44835-w.

[52] E. Zhao, M.R. Stone, X. Ren, et al. "Spatial transcriptomics at subspot resolution with BayesSpace". In: *Nature Biotechnology* 39 (2021), pp. 1375–1384. DOI: 10.1038/s41587-021-00935-2.

[53] J. Park, W. Choi, S. Tiesmeyer, et al. "Cell segmentation-free inference of cell types from in situ transcriptomics data". In: *Nature Communications* 12 (2021), p. 3545. DOI: 10.1038/s41467-021-23807-4.

[54] C. Ma et al. "Belayer: Modeling discrete and continuous spatial variation in gene expression from spatially resolved transcriptomics". In: *Cell Systems* 13.10 (2022), 786–797.e13. DOI: 10.1016/j.cels.2022.09.002.

[55] Jian Hu et al. "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network". In: *Nature Methods* 18 (2021), pp. 1342–1351. DOI: 10.1038/s41592-021-01255-8.

[56] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature Methods* 16.12 (2019), pp. 1289–1296. DOI: 10.1038/s41592-019-0619-0. URL: https://doi.org/10.1038/s41592-019-0619-0.

[57] R. Zeira, M. Land, A. Strzalkowski, et al. "Alignment and integration of spatial transcriptomics data". In: *Nature Methods* 19 (2022), pp. 567–575. DOI: 10.1038/s41592-022-01459-6.

[58] A. Jones, F.W. Townes, D. Li, et al. "Alignment of spatial genomics data using deep Gaussian processes". In: *Nature Methods* 20 (2023), pp. 1379–1387. DOI: 10.1038/s41592-023-01972-2.

[59] Yue Long et al. "Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST". In: *Nature Communications* 14 (2023), p. 1155. DOI: 10.1038/s41467-023-36796-3.

[60] F.W. Townes and B.E. Engelhardt. "Nonnegative spatial factorization applied to spatial genomics". In: *Nature Methods* 20 (2023), pp. 229–238. DOI: 10.1038/s41592-022-01687-w.

[61] B. Velten, J.M. Braunger, R. Argelaguet, et al. "Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO". In: *Nature Methods* 19 (2022), pp. 179–186. DOI: 10.1038/s41592-021-01343-9.

[62] Benjamin Chidester et al. "SPICEMIX enables integrative single-cell spatial modeling of cell identity". In: *Nature Genetics* 55 (2023), pp. 78–88. DOI: 10.1038/s41588-022-01256-z.

[63] E. Armingol, A. Officer, O. Harismendy, et al. "Deciphering cell–cell interactions and communication from gene expression". In: *Nature Reviews Genetics* 22 (2021), pp. 71–88. DOI: 10.1038/s41576-020-00292-x. URL: https://doi.org/10.1038/s41576-020-00292-x.

[64] Norbert Perrimon, Chrysoula Pitsouli, and Ben-Zion Shilo. "Signaling mechanisms controlling cell fate and embryonic patterning". In: *Cold Spring Harbor Perspectives in Biology* 4.8 (Aug. 2012), a005975. DOI: 10.1101/cshperspect.a005975.

[65] Dan Song et al. "Cell–cell communication: old mystery and new opportunity". In: *Cell Biology and Toxicology* 35 (2019), pp. 89–93. DOI: 10.1007/s10565-019-09470-y.

[66] Guillaume van Niel et al. "Challenges and directions in studying cell–cell communication by extracellular vesicles". In: *Nature Reviews Molecular Cell Biology* 23 (2022), pp. 369–382. DOI: 10.1038/s41580-022-00460-3.

[67] Michael Winding et al. "The connectome of an insect brain". In: *Science* 379 (2023), eadd9330. DOI: 10.1126/science.add9330.

[68] Louis K Scheffer et al. "A connectome and analysis of the adult Drosophila central brain". In: *eLife* 9 (Sept. 2020), e57443. DOI: 10.7554/eLife.57443.

[69] S. Nakandakari-Higa, S. Walker, M.C.C. Canesso, et al. "Universal recording of immune cell interactions in vivo". In: *Nature* 627 (2024), pp. 399–406. DOI: 10.1038/s41586-024-07134-4.

[70] Qinfeng Ma et al. "CellCommuNet: an atlas of cell–cell communication networks from single-cell RNA sequencing of human and mouse tissues in normal and disease states". In: *Nucleic Acids Research* 52.D1 (Jan. 5, 2024), pp. D597–D606. DOI: 10.1093/nar/gkad906. URL: https://doi.org/10.1093/nar/gkad906.

[71] Axel A. Almet et al. "The landscape of cell–cell communication through single-cell transcriptomics". In: *Current Opinion in Systems Biology* 26 (2021), pp. 12–23. ISSN: 2452-3100. DOI: https://doi.org/10.1016/j.coisb.2021.03.007. URL: https://www.sciencedirect.com/science/article/pii/S2452310021000081.

[72] E. Armingol, H. M. Baghdassarian, and N. E. Lewis. "The diversification of methods for studying cell–cell interactions and communication". In: *Nature Reviews Genetics* (2024). DOI: 10.1038/s41576-023-00685-8. URL: https://doi.org/10.1038/s41576-023-00685-8.

[73] Lihong Peng et al. "Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies". In: *Briefings in Bioinformatics* 23.4 (2022), bbac234. DOI: 10.1093/bib/bbac234. URL: https://doi.org/10.1093/bib/bbac234.

[74] T. Chari and L. Pachter. "The Specious Art of Single-Cell Genomics". In: *PLoS Computational Biology* 19.8 (Aug. 17, 2023), e1011288. DOI: 10.1371/journal.pcbi.1011288.

[75] M. Efremova, M. Vento-Tormo, and S.A. Teichmann. "CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes". In: *Nature Protocols* 15 (2020), pp. 1484–1506. DOI: 10.1038/s41596-020-0292-x.

[76] S. Jin, C.F. Guerrero-Juarez, and L. Zhang. "Inference and analysis of cell-cell communication using CellChat". In: *Nature Communications* 12 (2021), p. 1088. DOI: 10.1038/s41467-021-21246-9.

[77] D. Dimitrov et al. "Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data". In: *Nature Communications* 13.1 (June 9, 2022), p. 3224. DOI: 10.1038/s41467-022-30755-0.

[78] Z. Liu, D. Sun, and C. Wang. "Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information". In: *Genome Biology* 23 (2022), p. 218. DOI: 10.1186/s13059-022-02783-y. URL: https://doi.org/10.1186/s13059-022-02783-y.

[79] K. Mason, A. Sathe, P. R. Hess, et al. "Niche-DE: niche-differential gene expression analysis in spatial transcriptomics data identifies context-dependent cell-cell interactions". In: *Genome Biology* 25 (2024), p. 14. DOI: 10.1186/s13059-023-03159-6. URL: https://doi.org/10.1186/s13059-023-03159-6.

[80] D. S. Fischer, A. C. Schaar, and F. J. Theis. "Modeling intercellular communication in tissues using spatial graphs of cells". In: *Nature Biotechnology* 41 (2023), pp. 332–336. DOI: 10.1038/s41587-022-01467-z. URL: https://doi.org/10.1038/s41587-022-01467-z.

[81] Z. Cang and Q. Nie. "Inferring spatial and signaling relationships between cells from single cell transcriptomic data". In: *Nature Communications* 11 (2020), p. 2084. DOI: 10.1038/s41467-020-15968-5. URL: https://doi.org/10.1038/s41467-020-15968-5.

[82] Z. Cang et al. "Screening cell-cell communication in spatial transcriptomics via collective optimal transport". In: *Nature Methods* 20.2 (Feb. 2023), pp. 218–228. DOI: 10.1038/s41592-022-01728-4. eprint: 2023-01-23. URL: https://doi.org/10.1038/s41592-022-01728-4.

[83] X. Shao, C. Li, H. Yang, et al. "Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with SpaTalk". In: *Nature Communications* 13 (2022), p. 4429. DOI: 10.1038/s41467-022-32111-8. URL: https://doi.org/10.1038/s41467-022-32111-8.

[84] F. Qu, F. Guilak, and R.L. Mauck. "Cell migration: implications for repair and regeneration in joint disease". In: *Nature Reviews Rheumatology* 15 (2019), pp. 167–179. DOI: 10.1038/s41584-018-0151-0. URL: https://doi.org/10.1038/s41584-018-0151-0.

[85] A. J. Ridley et al. "Cell migration: integrating signals from front to back". In: *Science* 302.5651 (Dec. 5, 2003), pp. 1704–1709. DOI: 10.1126/science.1092053.

[86] P. Friedl and D. Gilmour. "Collective cell migration in morphogenesis, regeneration and cancer". In: *Nature Reviews Molecular Cell Biology* 10 (2009), pp. 445–457. DOI: 10.1038/nrm2720. URL: https://doi.org/10.1038/nrm2720.

[87] D. Boettiger. "Mechanical control of integrin-mediated adhesion and signaling". In: *Current Opinion in Cell Biology* 24 (2012), pp. 592–599. DOI: 10.1016/j.ceb.2012.08.002.

[88] T. Lämmermann, B. Bader, S. Monkley, et al. "Rapid leukocyte migration by integrin-independent flowing and squeezing". In: *Nature* 453 (2008), pp. 51–55. DOI: 10.1038/nature06887. URL: https://doi.org/10.1038/nature06887.

[89] R. Mayor and S. Etienne-Manneville. "The front and rear of collective cell migration". In: *Nature Reviews Molecular Cell Biology* 17 (2016), pp. 97–109. DOI: 10.1038/nrm.2015.14. URL: https://doi.org/10.1038/nrm.2015.14.

[90] J. Mouw, G. Ou, and V. Weaver. "Extracellular matrix assembly: a multiscale deconstruction". In: *Nature Reviews Molecular Cell Biology* 15 (2014), pp. 771–785. DOI: 10.1038/nrm3902. URL: https://doi.org/10.1038/nrm3902.

[91] Anoop P Patel et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190 (2014), pp. 1396–1401.

[92] Sidharth V Puram et al. "Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer". In: *Cell* 171.7 (2017), pp. 1611–1624.

[93] Christoph Ziegenhain et al. "Comparative analysis of single-cell RNA sequencing methods". In: *Molecular cell* 65.4 (2017), pp. 631–643.

[94] Iain C Macaulay, Chris P Ponting, and Thierry Voet. "Single-cell multiomics: multiple measurements from single cells". In: *Trends in Genetics* 33.2 (2017), pp. 155–168.

[95] Mario L Suva and Itay Tirosh. "Single-cell RNA sequencing in cancer: lessons learned and emerging challenges". In: *Molecular cell* 75.1 (2019), pp. 7–12.

[96] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. "Exponential scaling of single-cell RNA-seq in the past decade". In: *Nature protocols* 13.4 (2018), pp. 599–604.

[97] Kok Hao Chen et al. "RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells". In: *Science* 348.6233 (2015), aaa6090.

[98] Arjun Raj et al. "Imaging individual mRNA molecules using multiple singly labeled probes". In: *Nature methods* 5.10 (2008), pp. 877–879.

[99] Chee-Huat Linus Eng et al. "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". In: *Nature* 568.7751 (2019), pp. 235–239.

[100] Sheel Shah et al. "seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus". In: *Neuron* 94.4 (2017), pp. 752–758.

[101] Patrik L Ståhl et al. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". In: *Science* 353.6294 (2016), pp. 78–82.

[102] Sanja Vickovic et al. "High-definition spatial transcriptomics for in situ tissue profiling". In: *Nature methods* 16.10 (2019), pp. 987–990.

[103] Robert R Stickels et al. "Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2". In: *Nature biotechnology* (2020), pp. 1–6.

[104] Kaia Achim et al. "High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin". In: *Nature biotechnology* 33.5 (2015), pp. 503–509.

[105] Rahul Satija et al. "Spatial reconstruction of single-cell gene expression data". In: *Nature biotechnology* 33.5 (2015), pp. 495–502.

[106] Junyue Cao et al. "Comprehensive single-cell transcriptional profiling of a multicellular organism". In: *Science* 357.6352 (2017), pp. 661–667.

[107] Alexander B Rosenberg et al. "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding". In: *Science* 360.6385 (2018), pp. 176–182.

[108] Maryam Attar et al. "A practical solution for preserving single cells for RNA sequencing". In: *Scientific reports* 8.1 (2018), pp. 1–10.

[109] Siyuan Yang et al. "Decontamination of ambient RNA in single-cell RNA-seq with decontX". In: *Genome biology* 21.1 (2020), pp. 1–21.

[110] Jang Choon Lee et al. "Regulatory T cell control of systemic immunity and immunotherapy response in liver metastasis". In: *Science immunology* 5.52 (2020).

[111] Mahesh Yadav et al. "Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing". In: *Nature* 515.7528 (2014), pp. 572–576.

[112] Krithika N Kodumudi et al. "Immune checkpoint blockade to improve tumor infiltrating lymphocytes for adoptive cell therapy". In: *PloS one* 11.4 (2016), e0153053.

[113] Haidong Tang et al. "PD-L1 on host cells is essential for PD-L1 blockade-mediated tumor regression". In: *The Journal of clinical investigation* 128.2 (2018), pp. 580–588.

[114] Mirjana Efremova et al. "Targeting immune checkpoints potentiates immunoediting and changes the dynamics of tumor evolution". In: *Nature communications* 9.1 (2018), pp. 1–13.

[115] Tabula Muris Consortium et al. "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris". In: *Nature* 562.7727 (2018), pp. 367–372.

[116] Jiarui Ding et al. "Systematic comparative analysis of single cell RNA-sequencing methods". In: *BioRxiv* (2019), p. 632216.

[117] Maria Jo ao Carvalho Jord ao et al. "Single-cell profiling identifies myeloid cell subsets with distinct fates during neuroinflammation". In: *Science* (2019).

[118] Sung Vaugh Kim et al. "Modulation of cell adhesion and motility in the immune system by Myo1f". In: *Science* 314.5796 (2006), pp. 136–139.

[119] Xueyang Yu et al. "The cytokine TGF-β promotes the development and homeostasis of alveolar macrophages". In: *Immunity* 47.5 (2017), pp. 903–912.

[120] Hanna Helgeland et al. "Transcriptome profiling of human thymic CD4+ and CD8+ T cells compared to primary peripheral T cells". In: *BMC genomics* 21.1 (2020), pp. 1–11.

[121] Oliver J Harrison et al. "Epithelial-derived IL-18 regulates Th17 cell differentiation and Foxp3+ Treg cell function in the intestine". In: *Mucosal immunology* 8.6 (2015), pp. 1226–1236.

[122] Noah Isakov and Amnon Altman. "PKC-θ-mediated signal delivery from the TCR/CD28 surface receptors". In: *Frontiers in immunology* 3 (2012), p. 273.

[123] Lotta E Oikari et al. "Cell surface heparan sulfate proteoglycans as novel markers of human neural stem cell fate determination". In: *Stem cell research* 16.1 (2016), pp. 92–104.

[124] Daniel Fritz and Branko Stefanovic. "RNA-binding protein RBMS3 is expressed in activated hepatic stellate cells and liver fibrosis and increases expression of transcription factor Prx1". In: *Journal of molecular biology* 371.3 (2007), pp. 585–595.

[125] Miho Riku et al. "Down-regulation of the zinc-finger homeobox protein TSHZ2 releases GLI1 from the nuclear repressor complex to restore its transcriptional activity during mammary tumorigenesis". In: *Oncotarget* 7.5 (2016), p. 5690.

[126] Hema Kalyanaraman, Natasha Schall, and Renate B Pilz. "Nitric oxide and cyclic GMP functions in bone". In: *Nitric Oxide* 76 (2018), pp. 62–70.

[127] Natasha Schall et al. "Protein kinase G1 regulates bone regeneration and rescues diabetic fracture healing". In: *JCI insight* 5.15 (2020).

[128] Jaimin Baboo et al. "The impact of varying cooling and thawing rates on the quality of cryopreserved human peripheral blood T cells". In: *Scientific reports* 9.1 (2019), pp. 1–11.

[129] Qi Wang et al. "Interplay between mesenchymal stem cell and tumor and potential application". In: *Human Cell* 33.2 (2020), pp. 444–458.

[130] Koen Van den Berge et al. "Trajectory-based differential expression analysis for single-cell sequencing data". In: *Nature communications* 11.1 (2020), pp. 1–13.

[131] Juho Soikkeli et al. "Metastatic outgrowth encompasses COL-I, FN1, and POSTN up-regulation and assembly to fibrillar networks regulating cell adhesion, migration, and growth". In: *The American journal of pathology* 177.1 (2010), pp. 387–403.

[132] Yu Wang et al. "Systematic identification of the key candidate genes in breast cancer stroma". In: *Cellular and Molecular Biology Letters* 23.1 (2018), pp. 1–14.

[133] Jiao Li et al. "Stromal microenvironment promoted infiltration in esophageal adenocarcinoma and squamous cell carcinoma: A multi-cohort gene-based analysis". In: *Scientific reports* 10.1 (2020), pp. 1–12.

[134] Anika Escudero-Esparza et al. "Complement inhibitor CSMD1 acts as tumor suppressor in human breast cancer". In: *Oncotarget* 7.47 (2016), p. 76920.

[135] Santiago Ropero et al. "Epigenetic loss of the familial tumor-suppressor gene exostosin-1 (EXT1) disrupts heparan sulfate synthesis in cancer cells". In: *Human molecular genetics* 13.22 (2004), pp. 2753–2765.

[136] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome biology* 20.1 (2019), pp. 1–15.

[137] Renaud Gaujoux and Cathal Seoighe. "A flexible R package for nonnegative matrix factorization". In: *BMC bioinformatics* 11.1 (2010), pp. 1–9.

[138] Eran Eden et al. "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". In: *BMC bioinformatics* 10.1 (2009), pp. 1–7.

[139] Pedro Carmona-Saez et al. "Biclustering of gene expression data by non-smooth non-negative matrix factorization". In: *BMC bioinformatics* 7.1 (2006), pp. 1–18.

[140] Ruben Dries et al. "Giotto: a toolbox for integrative analysis and visualization of spatial expression data". In: *Genome Biology* 22.1 (2021), p. 78.

[141] D.S. Fischer, A.C. Schaar, and F.J. Theis. "Modeling intercellular communication in tissues using spatial graphs of cells". In: *Nature Biotechnology* 41 (2023), pp. 332–336. DOI: 10.1038/s41587-022-01467-z.

[142] Z. Cang, Y. Zhao, and A.A. Almet. "Screening cell–cell communication in spatial transcriptomics via collective optimal transport". In: *Nature Methods* 20 (2023), pp. 218–228. DOI: 10.1038/s41592-022-01728-4.

[143] K. Sriram and P. A. Insel. "G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs?" In: *JMol Pharmacol.* 93(4).4 (2018), pp. 251–258. DOI: 10.1124/mol.117.111062.

[144] Claudia Stäubert et al. "Superconserved receptors expressed in the brain: Expression, function, motifs and evolution of an orphan receptor family". In: *Pharmacology & Therapeutics* 240 (2022), p. 108217. ISSN: 0163-7258. DOI: https://doi.org/10.1016/j.pharmthera.2022.108217.

[145] Y. Fang, T. Kenakin, and Liu C. "Editorial: Orphan GPCRs As Emerging Drug Targets". In: *Frontiers in Pharmacology* 6 (2015).

[146] Z. Li, T. Wang, and P. Liu. "SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns". In: *Nature Communications* 14 (2023), p. 3995. DOI: 10.1038/s41467-023-39608-w.

[147] P. A. P. Moran. "Notes on Continuous Stochastic Phenomena". In: *Biometrika* 37.1/2 (1950), pp. 17–23.

[148] Y. Yuan and Z. Bar-Joseph. "GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data". In: *Genome Biology* 21 (2020), p. 300. DOI: 10.1186/s13059-020-02214-w.

[149] Jonathan Masci et al. "Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction". In: *Artificial Neural Networks and Machine Learning – ICANN 2011*. Springer. Berlin, Heidelberg, 2011, pp. 52–59.

[150] Davide Chicco. "Siamese neural networks: an overview". In: *Artificial Neural Networks, Methods in Molecular Biology* 2190 (2020), pp. 73–94. DOI: 10.1007/978-1-0716-0826-5_3.

[151] Wei-Ting Chen et al. "Spatial Transcriptomics and In Situ Sequencing to Study Alzheimer's Disease". In: *Cell* (2020). DOI: 10.1016/j.cell.2020.06.038.

[152] B. Li, W. Zhang, C. Guo, et al. "Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution". In: *Nature Methods* 19 (2022), pp. 662–670. DOI: 10.1038/s41592-022-01480-9. URL: https://doi.org/10.1038/s41592-022-01480-9.

[153] Xin Shao et al. "CellTalkDB: a manually curated database of ligand–receptor interactions in humans and mice". In: *Briefings in bioinformatics* 22.4 (2021), bbaa269.

[154] Albert J Kooistra et al. "GPCRdb in 2021: integrating GPCR sequence, structure and function". In: *Nucleic Acids Research* 49.D1 (2021), pp. D335–D343.

[155] Albert J Kooistra et al. "GPCRdb in 2021: integrating GPCR sequence, structure and function". en. In: *Nucleic Acids Res.* 49.D1 (Dec. 2020), pp. D335–D343. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1080.

[156] Vignir Isberg et al. "Generic GPCR residue numbers - aligning topology maps while minding the gaps". en. In: *Trends Pharmacol. Sci.* 36.1 (Jan. 2015), pp. 22–31. ISSN: 0165-6147, 1873-3735. DOI: 10.1016/j.tips.2014.11.001.

[157] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. "FastTree 2–approximately maximum-likelihood trees for large alignments". en. In: *PLoS One* 5.3 (Mar. 2010), e9490. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0009490.

[158] Alastair Compston and Alasdair Coles. "Multiple sclerosis". In: *The Lancet* 372.9648 (2008), pp. 1502–1517. DOI: 10.1016/S0140-6736(08)61620-7.

[159] Calliope A Dendrou, Lars Fugger, and Manuel A Friese. "Immunopathology of multiple sclerosis". In: *Nature Reviews Immunology* 15.9 (2015), pp. 545–558. DOI: 10.1038/nri3871.

[160] Klaus-Armin Nave and Hauke B Werner. "Myelination of the nervous system: mechanisms and functions". In: *Annual Review of Cell and Developmental Biology* 30 (2014), pp. 503–533. DOI: 10.1146/annurev-cellbio-100913-013101.

[161] Andrew P Robinson et al. "The experimental autoimmune encephalomyelitis (EAE) model of MS: utility for understanding disease pathophysiology and treatment". In: *Handbook of Clinical Neurology*. Vol. 122. Elsevier, 2014, pp. 173–189. DOI: 10.1016/B978-0-444-52001-2.00008-X.

[162] Cris S Constantinescu et al. "Experimental autoimmune encephalomyelitis (EAE) as a model for multiple sclerosis (MS)". In: *British Journal of Pharmacology* 164.4 (2011), pp. 1079–1106. DOI: 10.1111/j.1476-5381.2011.01302.x.

[163] Denise Chee Hsu. "Janeway's Immunobiology, 7th Edition". In: *Shock* 29.6 (June 2008), p. 770. DOI: 10.1097/01.SHK.0000286285.87596.06.

[164] Bruce Alberts, Alexander Johnson, Julian Lewis, et al. *Molecular Biology of the Cell*. Garland Science, 2008.

[165] Lieping Chen and Dallas B Flies. "Molecular mechanisms of T cell co-stimulation and co-inhibition". In: *Nature Reviews Immunology* 13.4 (2013), pp. 227–242.

[166] Roza I Nurieva et al. "Essential autocrine regulation by IL-21 in the generation of inflammatory T cells". In: *Nature* 448.7152 (2007), pp. 480–483.

[167] Jinfang Zhu, Hiroshi Yamane, and William E Paul. "Differentiation of effector CD4 T cell populations". In: *Annual Review of Immunology* 28 (2010), pp. 445–489.

[168] Kenneth M Murphy and Steven L Reiner. "The lineage decisions of helper T cells". In: *Nature Reviews Immunology* 2.12 (2002), pp. 933–944.

[169] Ivaylo I Ivanov et al. "The orphan nuclear receptor RORgammat directs the differentiation program of proinflammatory IL-17+ T helper cells". In: *Cell* 126.6 (2006), pp. 1121–1133.

[170] Dan R Littman and Alexander Y Rudensky. "Th17 and regulatory T cells in mediating and restraining inflammation". In: *Cell* 140.6 (2010), pp. 845–858.

[171] Shimon Sakaguchi et al. "FOXP3+ regulatory T cells in the human immune system". In: *Nature Reviews Immunology* 10.7 (2010), pp. 490–500.

[172] I. A. Ferber et al. "Mice with a disrupted IFN-gamma gene are susceptible to the induction of experimental autoimmune encephalomyelitis (EAE)". In: *Journal of Immunology* 156.1 (1996), pp. 5–7.

[173] D. O. Willenborg et al. "IFN-gamma plays a critical down-regulatory role in the induction and effector phase of myelin oligodendrocyte glycoprotein-induced autoimmune encephalomyelitis". In: *Journal of Immunology* 157.8 (1996), pp. 3223–3227.

[174] B. Becher, B. G. Durell, and R. J. Noelle. "Experimental autoimmune encephalitis and inflammation in the absence of interleukin-12". In: *Journal of Clinical Investigation* 110.4 (2002), pp. 493–497.

[175] Estelle Bettelli et al. "Reciprocal developmental pathways for the generation of pathogenic effector TH17 and regulatory T cells". In: *Nature* 441.7090 (2006), pp. 235–238.

[176] S. S. Duffy et al. "The role of regulatory T cells in nervous system pathologies". In: *Journal of Neuroscience Research* 96.6 (2018), pp. 951–968. DOI: 10.1002/jnr.24073.

[177] Linsey A Stephens, Kevin H Malpass, and Stephen M Anderton. "Curing CNS autoimmune disease with myelin-reactive Foxp3+ Treg". In: *European Journal of Immunology* 39.4 (2009), pp. 1108–1117.

[178] Shimon Sakaguchi. "Naturally arising CD4+ regulatory t cells for immunologic self-tolerance and negative control of immune responses". In: *Annual Review of Immunology* 22 (2004), pp. 531–562. DOI: 10.1146/annurev.immunol.21.120601.141122.

[179] James B Wing et al. "Regulatory T cells control antigen-specific expansion of Tfh cell number and humoral immune responses via the coreceptor CTLA-4". In: *Immunity* 41.6 (2014). Epub 2014 Dec 6, pp. 1013–1025. DOI: 10.1016/j.immuni.2014.12.006.

[180] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (2014), p. 550. DOI: 10.1186/s13059-014-0550-8. URL: https://doi.org/10.1186/s13059-014-0550-8.

[181] W. Li, H. Xu, T. Xiao, et al. "MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens". In: *Genome Biology* 15 (2014), p. 554. DOI: 10.1186/s13059-014-0554-4.

[182] F. Alex Wolf, Philipp Angerer, and Fabian J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (2018), p. 15. DOI: 10.1186/s13059-017-1382-0.

[183] Nils J. Bernstein et al. "Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning". In: *Cell Systems* 11.1 (2020), 95–101.e5. DOI: 10.1016/j.cels.2020.06.003.

[184] A. Cui, T. Huang, S. Li, et al. "Dictionary of immune responses to cytokines at single-cell resolution". In: *Nature* 625 (2024), pp. 377–384. DOI: 10.1038/s41586-023-06816-9. URL: https://doi.org/10.1038/s41586-023-06816-9.

[185] Jordan W. Squair, Mathieu Gautier, Claudia Kathe, et al. "Confronting false discoveries in single-cell differential expression". In: *Nature Communications* 12 (2021), p. 5692. DOI: 10.1038/s41467-021-25960-2. URL: https://doi.org/10.1038/s41467-021-25960-2.

[186] Keith D. Zimmerman, Mark A. Espeland, and Carl D. Langefeld. "A practical solution to pseudoreplication bias in single-cell studies". In: *Nature Communications* 12 (2021), p. 738. DOI: 10.1038/s41467-021-21038-1. URL: https://doi.org/10.1038/s41467-021-21038-1.

[187] Sini Junttila, Johannes Smolander, and Laura L Elo. "Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data". In: *Briefings in Bioinformatics* 23 (5 2022). DOI: 10.1093/bib/bbac286. URL: https://doi.org/10.1093/bib/bbac286.

[188] A. Zhu, J. G. Ibrahim, and M. I. Love. "Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences". In: *Bioinformatics* (2018). DOI: 10.1093/bioinformatics/bty895.

Appendix A

Supplementary Material for Chapter 2

A.1 Experimental Methods

Mice, Tumor Cell Line, and Tumor Inoculation

$6 - 12$ weeks old C57BL/6 female mice were purchased from Jackson Laboratories and housed in specific pathogen free conditions. MC38 colon adenocarcinoma cell line expressing luciferase was a gift from R. D. Beauchamp (Vanderbilt University). MC38 cell line was cultured in complete cell culture medium (RPMI 1640 with GlutaMAX, penicillin, streptomycin, sodium pyruvate, HEPES, NEAA, and 10% fetal bovine serum (FBS). Cell lines were routinely tested for mycoplasma contamination. For experiments, mice were given an anesthetic cocktail of Buprenorphine ($300\mu l$) and Meloxiacam ($300\mu l$) 30 minutes prior to the procedure. At the time of surgery, 1 drop of Bupivacaine was administered and mice were anesthetized with isoflurane prior to intrahepatic (or intrasplenic) injection of MC38 colon adenocarcinoma cells ($50\mu l$ at 10×10^6 cells/ml) using a $30\frac{1}{2}$ gauge needle. Incision was stapled closed and post-operative care was given to the mice. All experiments were conducted in accordance with the animal protocol approved by the University of California, San Francisco IACUC committee.

Cancer Model System

Intrahepatic and intrasplenic cancer model that we used for the paper is described in great detail in recently published report, Lee et al 2020 ([110]. Briefly, intrahepatic and intrasplenic tumors were generated by subcapsular injection of the tumor cells directly into the organs. To establish the ideal time point for sacrificing the mice, in vivo imaging was done on tumor inoculated mice. Intra-organ injected MC38 cells were modified to express the firefly luciferase. Mice were intraperitoneally infected with D-luciferin (150mg/kg; Gold Biotechnology) 7 minutes prior to imaging with the Xenogen IVIS Imaging system. Mice with detectable tumor nodules with at least 5mm fluorescence were sacrificed for tissue harvesting. Organs to be used for XYZeq were fixed with dithiobis(succinimidyl propionate) (DSP)

(Thermo Scientific) and cryopreserved while organs used for 10X Genomics Chromium Single cell sequencing were digested in RPMI complete medium that were supplemented with collagenase D (125U/ml;Roche) and deoxyribonuclease I (20mg/ml;Roche) then processed for single cell suspension using the gentleMACS tissue dissociator per manufacturer's protocol (Miltenyi).

10x Genomic Chromium Platform

Cells isolated from tissue were washed and resuspended in PBS with 0.04% BSA at 1000 cells/μl and loaded on the 10x Genomics Chromium platform per manufacturer's instructions and sequenced on NovaSeq or HiSeq 4000 (Illumina).

Tissue Harvesting and Cryopreservation

At day 10 post tumor inoculation, mice were sacrificed and harvested for the tumor injected liver (or spleen) and incubated for 30 minutes in ice cold DMSO-free freezing media (Bulldog Bio). This was followed by 30 minutes incubation in ice cold DSP (Thermo Scientific) supplemented with 10% FCS, then neutralized in ice cold 20mM of Tris-HCl, pH 7.5. The organs were placed in a cryomold, sealed airtight, and slowly frozen overnight in -80C.

Cells and Reagent Dispensing into Array

The sciFLEXARRAYER S3 (Scienion AG) was used to dispense cells and reagents to the microwell arrays. Drop stability and array quality were assessed for each experiment. Prior to dispensing into the microwell arrays slides, Autodrop detection was used to assess drop stability and quantify the velocity, deviations, and drop volume for each reagent. Volume entry was used to determine the number of drops required to reach the total designated well volume. Each well oligo(dT) primer 5' CTACACGACGCTCTTCCGATCTNNNNNNNNNN[16bp unique spatial barcode] TTTTTTTTTTTTTTTTTT-3', where "N" is any base; IDT) were spotted. During barcoding, the dewpoint control software monitored the ambient temperature and humidity allowing dynamic control of the temperature of the source plate to maintain nominal oligo concentrations through the duration of the run. Barcoded slides were dried in the wells prior to storage. Reaction mix (Thermo Fisher Scientific) were added to wells and automated with a 10% bleach wash between each probe to eliminate carry over contamination. Dissociation/permeabilization buffer was printed into each well on day of experiment and tissue section was loaded onto the microwell array slides. For all tissue experiments, DSP fixed HEK293T cells were added at $5\mu l$ (@10×10^6 cells/ml) to the RT digestion mix before being dispensed across all the wells in the microarray. The average number of HEK293T cells were 58 cells/well, however, the absolute number of cells per well likely varied across the array due to the cells being in suspension inside the dispensing nozzle. Cells harvested from the array after incubation was analyzed on ARIA (BD biosciences) and datasets were analyzed using FlowJo software (Tree Star Inc).

Array Fabrication

Photoresist masters are created by spinning on a layer of photoresist SU-8 2150 (Fisher Scientific) onto a 3-inch silicon wafer (University Wafer) at 1500rpm, then soft baking at 95C for 2hr. Then photoresist-layered silicon wafer is exposed to ultraviolet light (UV) for 30min over a photolithography masks (CAD/Art Sciences, USA) that was printed at 12,000 DPI. After ultraviolet exposure, the wafers are hard baked at 95C for 20min then developed for 2hrs in fresh solution of propylene glycol monomethyl ether acetate (Sigma Aldrich) to develop, followed by a manual rinse with fresh propylene glycol monomethyl ether acetate then baked at 95C for 2 min to remove residual solvent. Polymethylsiloxane (PDMS) mixture (Sylgard 184, Dow Corning Midland) with pre-polymer:curing-agent ratios of $10:1$ was poured over the SU-8 silicon wafer master. This was placed in a 100mm petri dish and was cured overnight in a 70°C oven. This PDMS negative mold was peeled off the SU-8 silicon master the following day. PDMS block was placed on a flat surface and Norland Optical Adhesive 81(NOA81) (Thorlabs) was poured into the mold to cover the entire surface. A slide was placed on top of the NOA-poured PDMS mold, and a transparent weight was placed on top. NOA was cured for 2 mins under UV light, flipping once half way thru the UV curing time. Finally, PDMS mold was detached from the cured NOA microwell array slide (referred to as microwell array chips). The dimensions of each hexagonal well is approximately $400\mu m$ in height and $500\mu m$ in diameter with the volume of 0.04mm^3 which can hold 40nl of liquid.

Cell Species Mixing Experiment

Mixture of HEK293T and NIH/3T3 cells were deposited into wells in a gradient pattern across the columns of the array with a total of 11 distinctive cell proportion ratios. Specifically columns on the array was spotted with human cells to mouse cells ratio of 100/0; 90/10; 80/20; 70/30; 60/40; 50/50; 40/60; 30/70; 20/80; 10/90; 0/100; 10/90; 20/80; 30/70; 40/60; 50/50; 60/40; 70/30; 80/20; 90/10; 100/0, with only human cells flanking the end columns and only mouse cells in the center columns. The ratio of UMI de-duplicated reads aligning to either human or mouse reference genomes were calculated for each cell, and those with less than 66% aligning to a single species were deemed barcode collision cells.

Acknowledgments

We thank C. Her for isolating hepatocytes; Jonathan Fischer for helpful discussion on differential gene expression analysis, J. Bolen and K. Wong for assistance in immunohistochemical staining; and A. Shaw and C. Apostolopoulou for tissue related discussion and imaging. J. Haliburton for discussion on microwell array fabrication.

Funding

Y.L. was supported by the T32 Ruth L Kirschstein Fellowship T32AI00733429. C.T.M. is supported by NIH grants T32GM007618, T32DK007418, and F30AI157167, and is a UCSF ImmunoX Computational Immunology Fellow. G.H is supported by National Science Foundation Graduation Research Fellow under grant number 1650113. D.N.N. is supported by NIH grants L40Al140341 and K08Al153767 and CIRM Alpha Stem Cell Clinic Fellowship. E.D.C. is supported by the NIH grant RM1 HG009490. A.M., C.J.Y, and Y.C. are supported by NIH grant DP3DK111914-01 and funding from the Northern California JDRF Center of Excellence. A.M. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund, is an investigator at the Chan Zuckerberg Biohub and has received funding from the Innovative Genomics Institute (IGI), the American Endowment Foundation, the Cancer Research Institute (CRI), the Lloyd J. Old STAR award, a gift from the Jordan Family, a gift from Barbara Bakar, and is a member of the Parker Institute for Cancer Immunotherapy (PICI). C.J.Y. is further supported by the NIH grants R01AR071522, R01AI136972, and the Chan Zuckerberg Initiative. C.J.Y is an investigator at the Chan Zuckerberg Biohub and is a member of the Parker Institute for Cancer Immunotherapy (PICI). Y.W. and Y.S.S. are supported in part by NIH grant R35-GM134922. Y.S.S is a Chan Zuckerberg Biohub Investigator.

Author Contributions

Y.L., D.B., C.J.Y, and E.D.C conceived the idea. Y.L. and D.B. fabricated the microwell array slides, developed library preparation pipeline, and completed the development of the method. E.D.C., A.M., C.J.Y., and Y.S.S. supervised the work. H.N. and Y.W. carried out differential gene expression analysis under supervision of Y.S.S.. J.M.W. and Y.L. developed the tissue fixing and freezing method. J.C. and D.B. developed the spotting protocol on the sciFLEXARRAYER. Y.W. performed decontamination analysis under supervision of C.J.Y. and Y.S.S.. Y.W. and D.B. performed the gene module analysis and trajectory inference analysis under supervision of C.J.Y. and Y.S.S.. D.B., Y.W., and G.H. developed the analytical pipeline and performed all sequencing analysis under supervision of C.J.Y. and Y.S.S.. J.L. and S.M. developed the mouse liver tumor model and performed the tumor injections. E.S. designed primers. D.N.N. provided input on the design of hexagonal wells. C.T.M., D.L., Y.L., and Y.S. prepared the 10X libraries. Y.L., Y.W., A.M., E.D.C., Y.S.S., and C.J.Y. wrote the manuscript. Competing interests: Y.L., D.B., A.M., E.D.C., C.J.Y., and J.M.W. are listed as inventors on a patent application related to this work. A.M. is a co-founder of Spotlight Therapeutics and Arsenal Biosciences. A.M. has served as an advisor to Juno Therapeutics, was a member of the scientific advisory board at PACT Pharma, and was an advisor to Trizell. A.M. owns stock in Arsenal Biosciences, Spotlight Therapeutics and PACT Pharma. The Marson lab has received research support from Juno Therapeutics, Epinomics, Sanofi, Gilead, Anthem and GlaxoSmithKline. C.J.Y. is a co-founder of and

owns equity in Dropprint Genomics.

A.2 Supplementary Figures

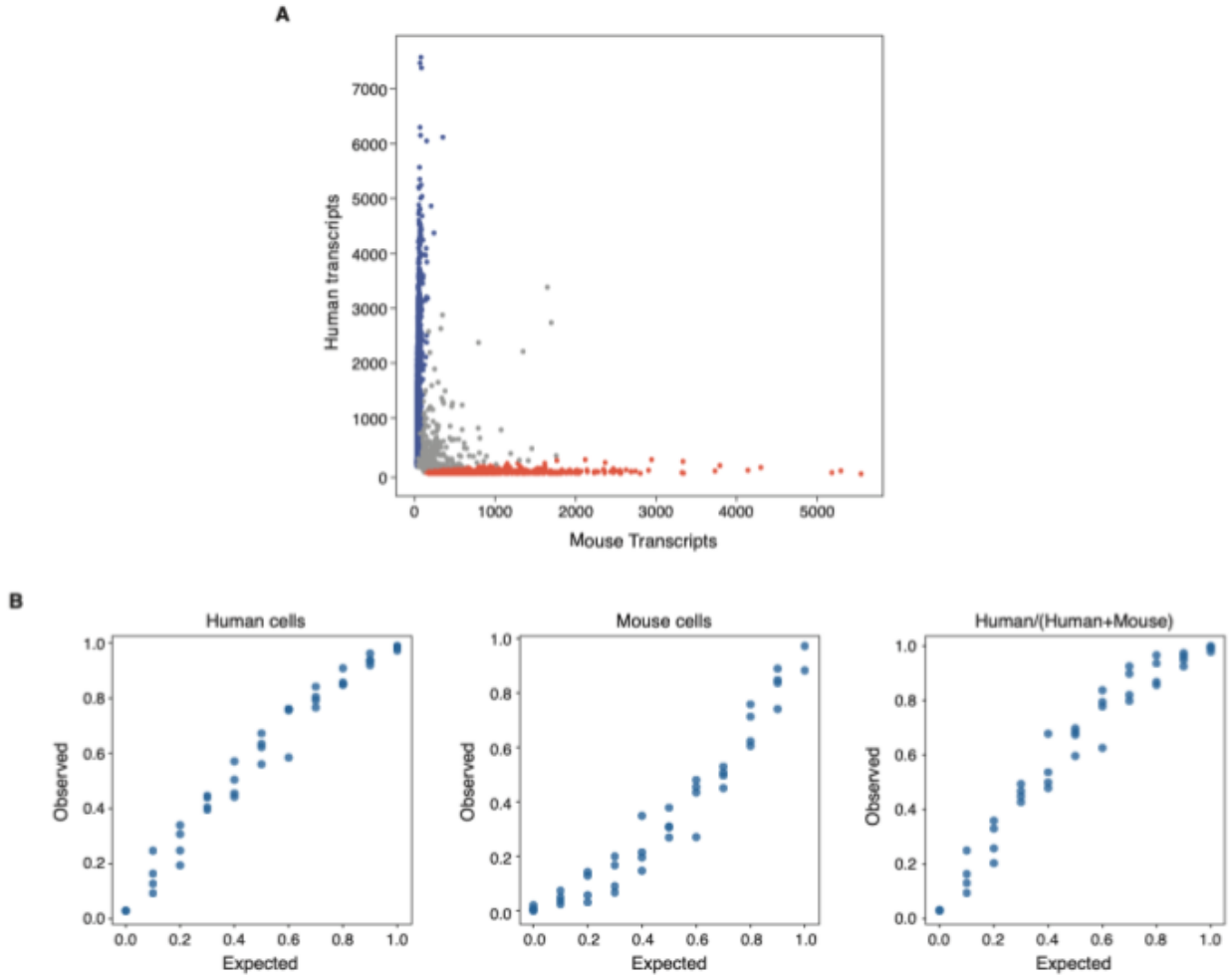


Figure A.1: Single cell mixed species experiment reveals strong correlation to estimated cell gradient proportions. (A) Scatter plot of mouse (x-axis) and human (y-axis) UMI counts detected from a mixture of HEK293T and NIH3T3 cells. Blue refers to human cells ($n = 4,389$) and red refers to mouse cells ($n = 1,728$) and gray refers to collisions ($n = 330$). (B) Scatter plots revealing high concordance between observed and expected cell type proportions in each column of the XYZeq array. Lin's Concordance Correlation = 0.91.

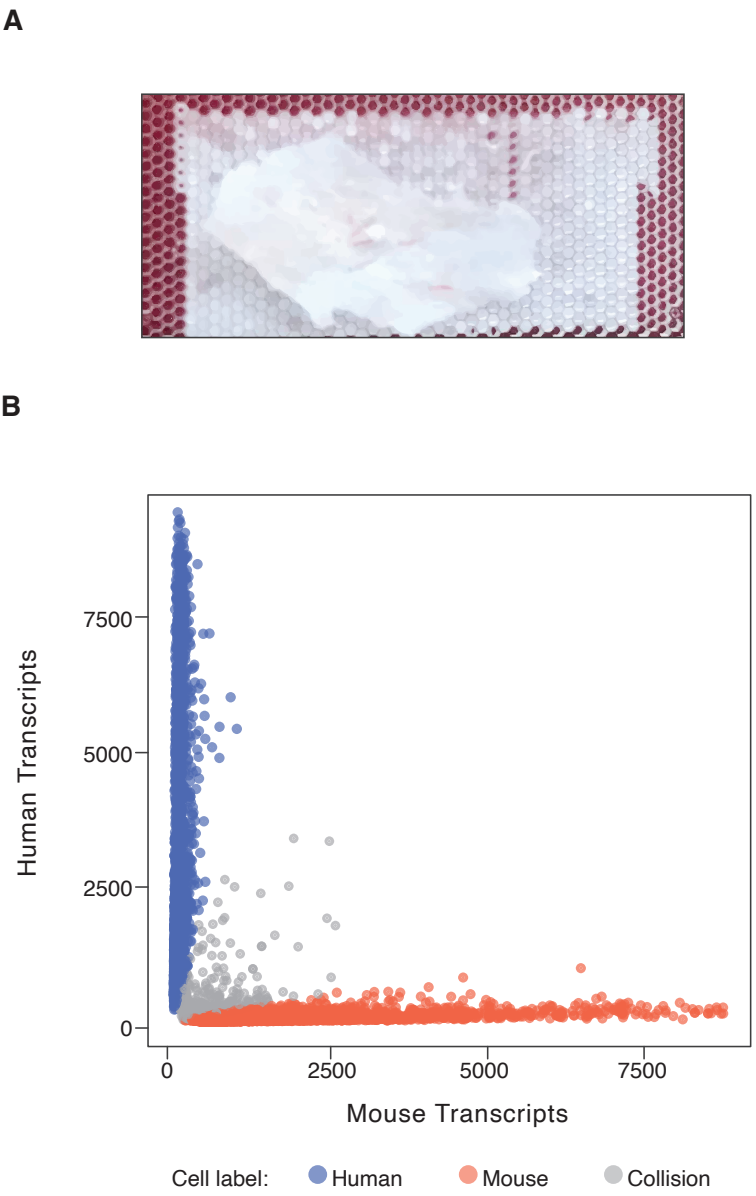


Figure A.2: Quantification of cells captured per well from liver/tumor tissue. (A) Image of the liver/tumor tissue slice on top of the XYZeq microarray with reagents spotted in wells (frozen: white). Photo credit: Youjin Lee (UCSF). (B) Scatter plot of transcripts ($n = 4$) from human (blue: $n = 2,667$), mouse cells (red: $n = 6,854$), and collisions (gray: $n = 747$).

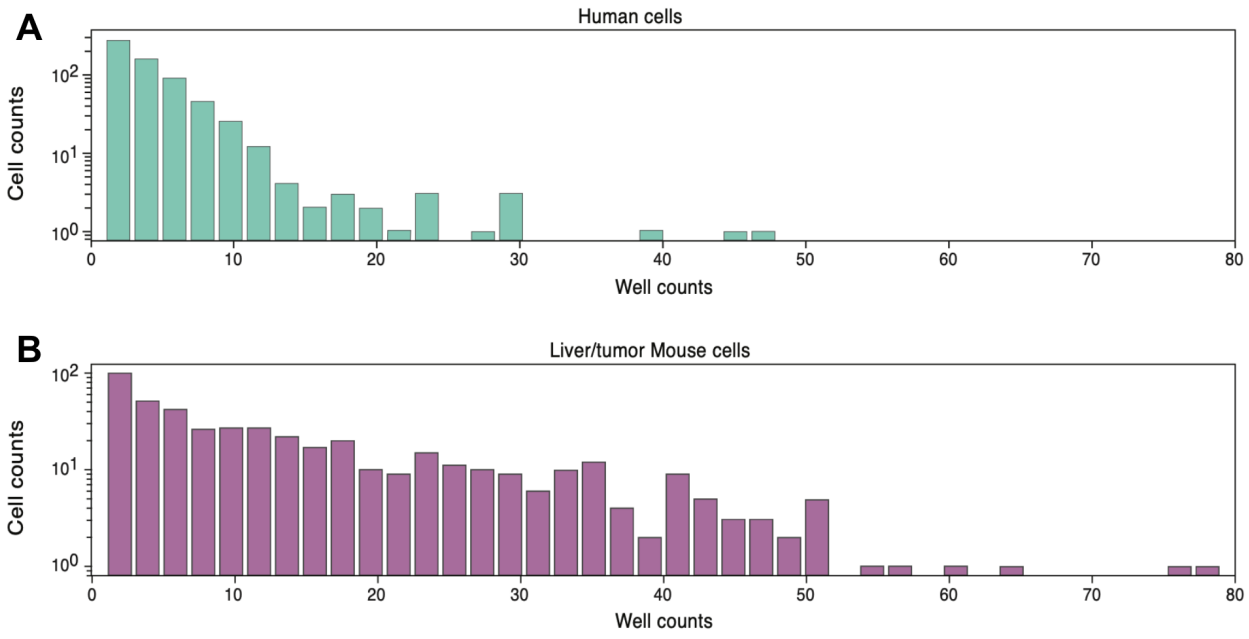


Figure A.3: The median cell number in wells across XYZeq array for HEK293T human (A) and liver/tumor mouse (B) cells.

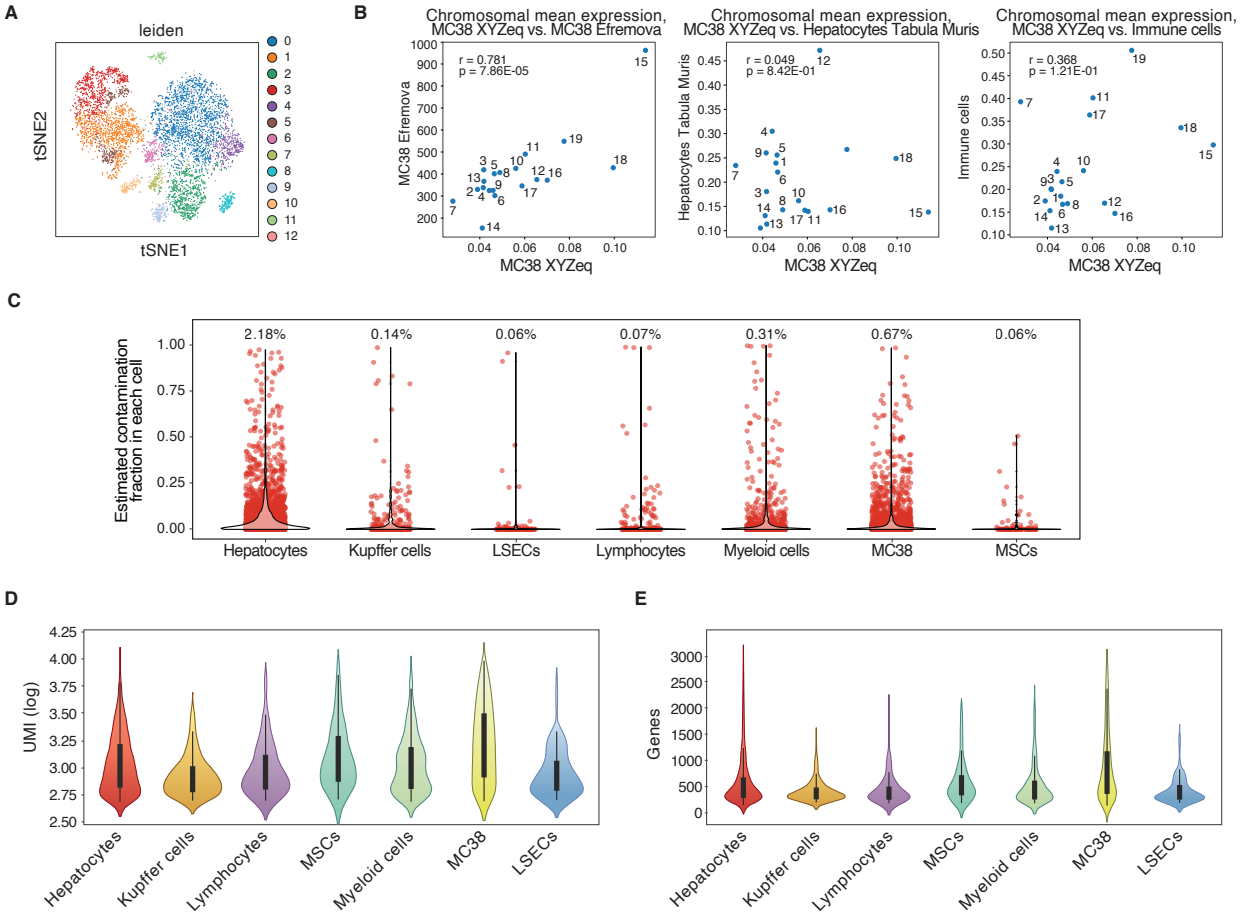


Figure A.4: Distinct cell types clusters identified from XYZeq of liver/tumor tissue. (A) t-distributed stochastic neighbor embedding (tSNE) visualization of Leiden cluster to annotated cell types. (B) Correlation of mean chromosomal expression of MC38 cells observed in XYZeq compared to MC38 cells from Efremova, et al., hepatocytes from Tabula Muris, and immune cells enriched from liver/tumor tissue from an independent internal experiment. Both the x-axis and y-axis denotes average expression of all genes on a given chromosome. (C) Violin plot representing estimated contamination fraction for each cell type from our liver/tumor XYZeq data. (D,E) Violin plot showing the number of detected UMIs and genes per cell cluster. Median UMI counts (log) and gene counts for each cell cluster: hepatocytes (3.04 and 552), Kupffer cells (2.92 and 420), lymphocytes (2.97 and 454), mesenchymal stem cells (MSCs) (3.08 and 594), myeloid cells (3.03 and 511), MC38 (3.22 and 851), and liver sinusoidal endothelial cells (LSECs) (2.94 and 431).

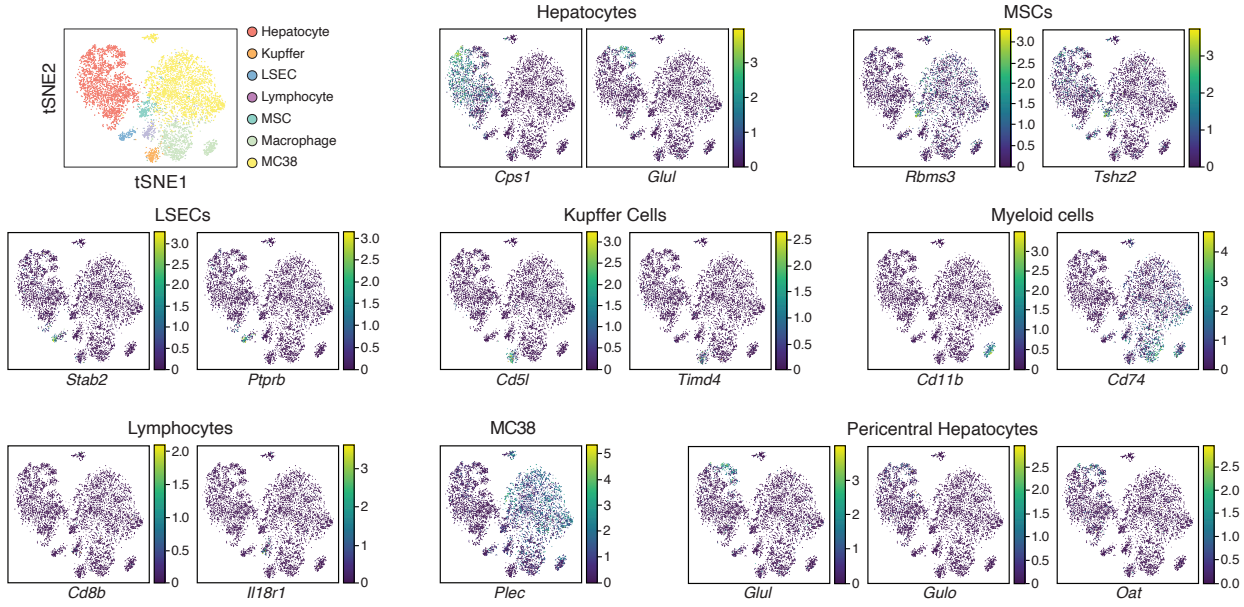


Figure A.5: Annotated cell-identity clusters. Feature plot of cells that are positive for each individual marker genes to identify hepatocytes (Cps1, Glul), MC38 (Plec), myeloid cells (Cd11b, Cd74), LSECs (Stab2, Ptprb), lymphocytes (Cd8b, Il18r1), Kupffer cells (Cd5l, Timd4), MSCs (Rbms3, Tshz2), pericentral hepatocytes (Glul, Gluo, Oat) from low expression (purple) to high expression (yellow).

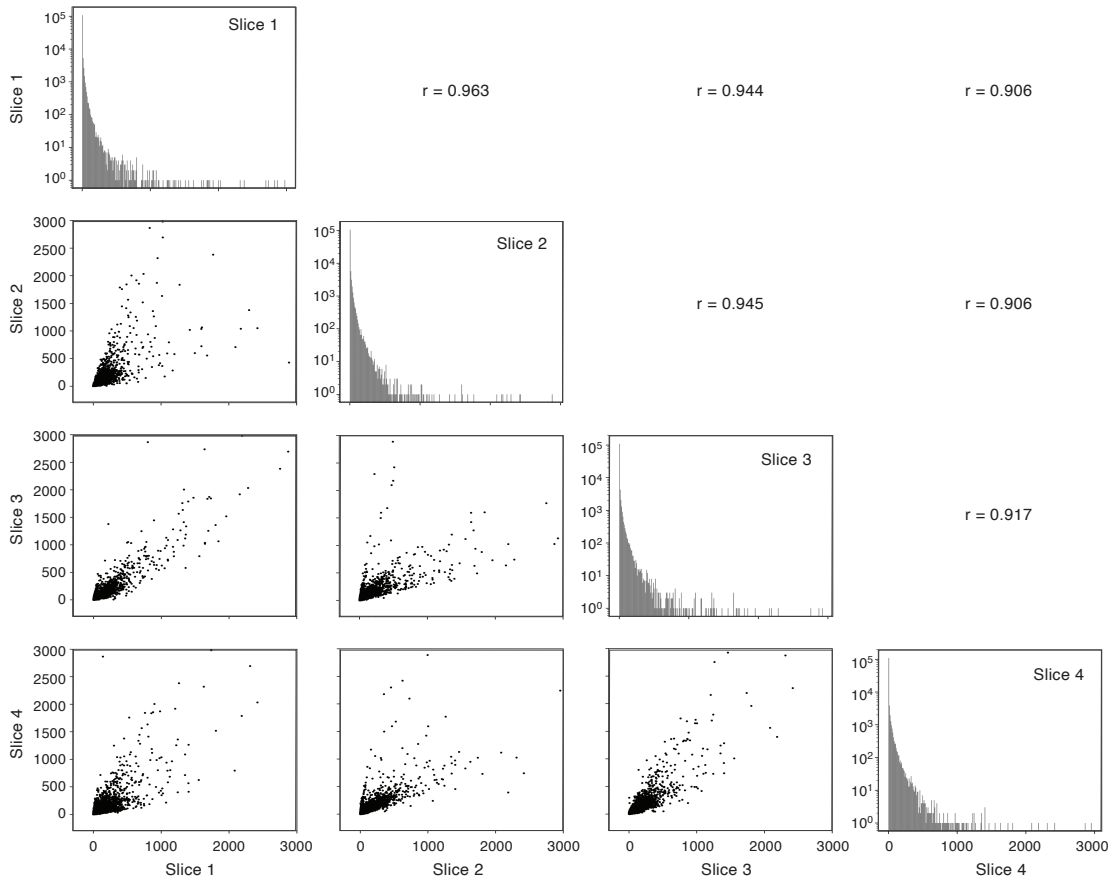


Figure A.6: Reproducibility of XYZeq across tissue slices. Four non-sequential z-layer slices of liver/tumor tissue processed with XYZeq (with HEK293T cells spiked-in as control). Pairplot showing the expression of common genes between different slices of liver/tumor. Scatter plots show the UMI counts for common expressed (UMIs > 0) genes. Histograms show the distribution of number of UMIs (x-axis) per gene (y-axis) for each slice.

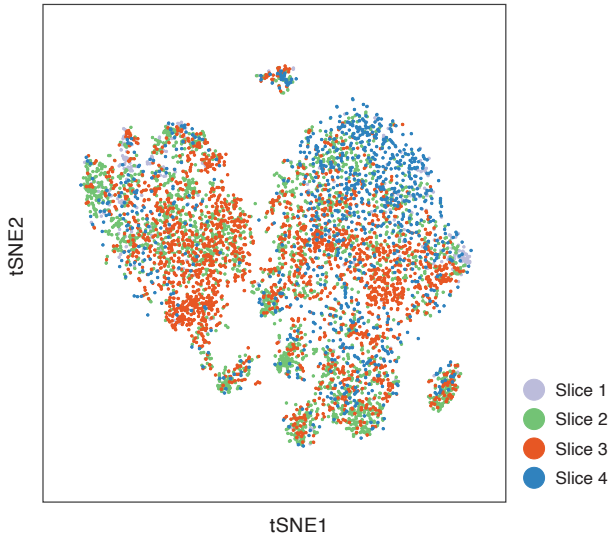


Figure A.7: tSNE visualization of Leiden clusters across four liver/tumor slices.

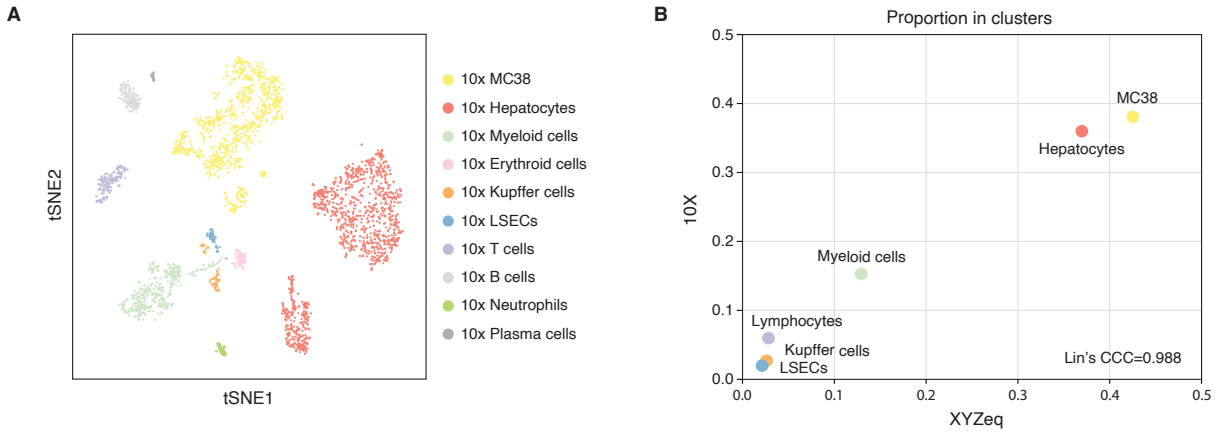


Figure A.8: Cell type clusters captured from XYZeq found comparable to 10X Genomics platform. (A) tSNE representation of liver/tumor tissue data generated with the 10X Chromium V3 kit. 2,703 total cells were plotted. (B) Scatter plot comparing the proportion of each cell type found in XYZeq and 10X Chromium V3. Lin's concordance coefficient = 0.988.

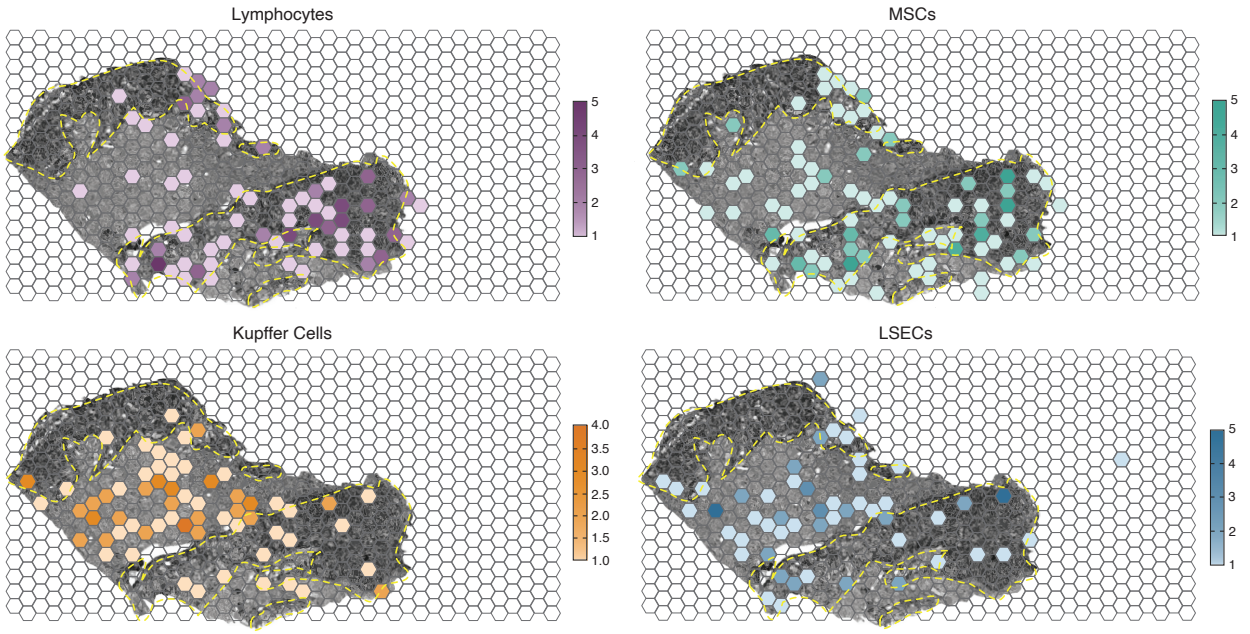


Figure A.9: Spatial density plot showing localization of lymphocytes (purple), MSCs (green), Kupffer cells (orange) and LSECs (blue) in the spatial array. Yellow dotted outline indicates tumor region.

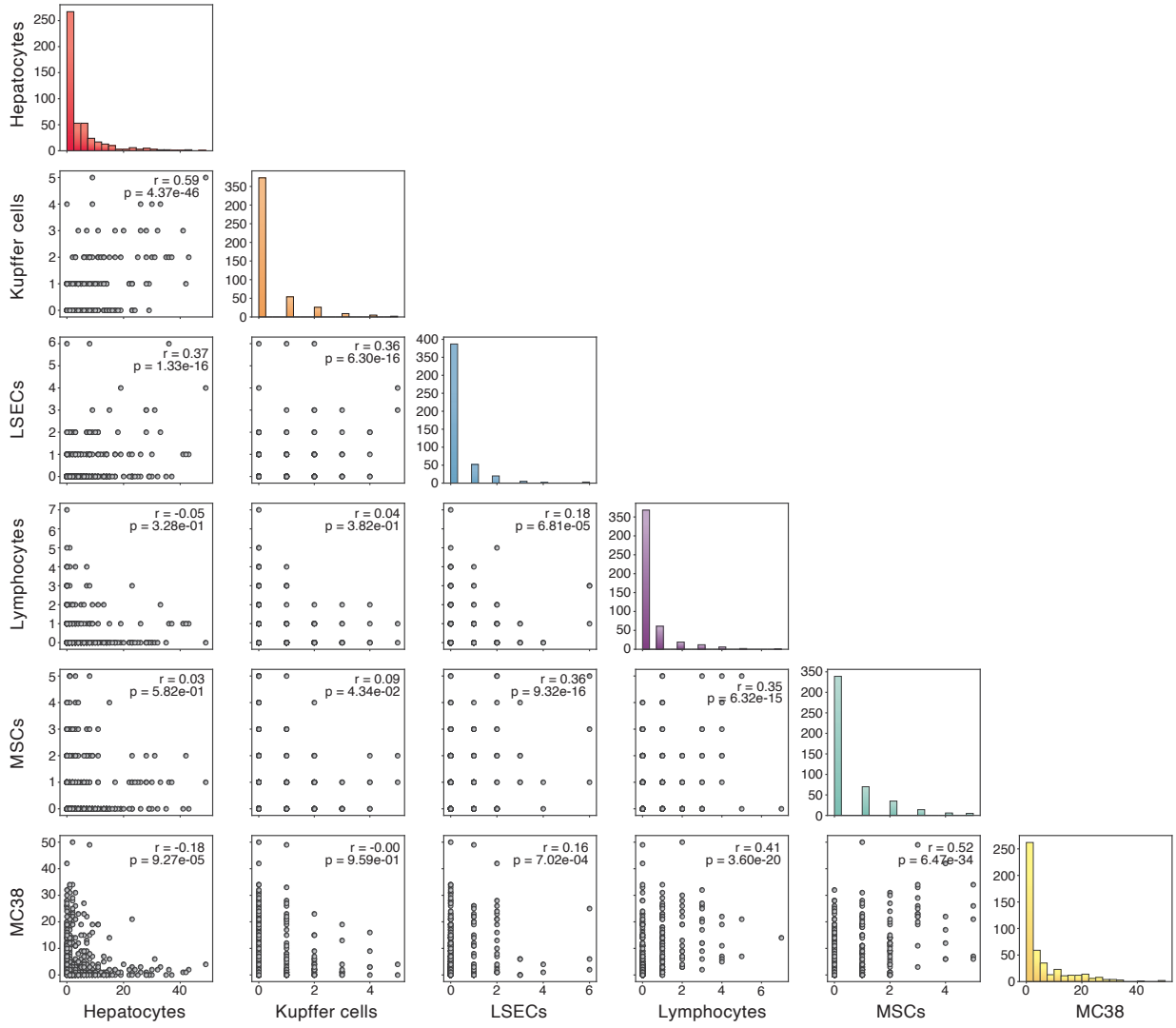


Figure A.10: Pairplot showing the frequency of cell types found in each well across the XYZeq array. Scatter plots show the co-localization of the cell types in each well. Histograms show the distribution for number of cells (x-axis) per well (y-axis) for each cell type. r and p values annotated.

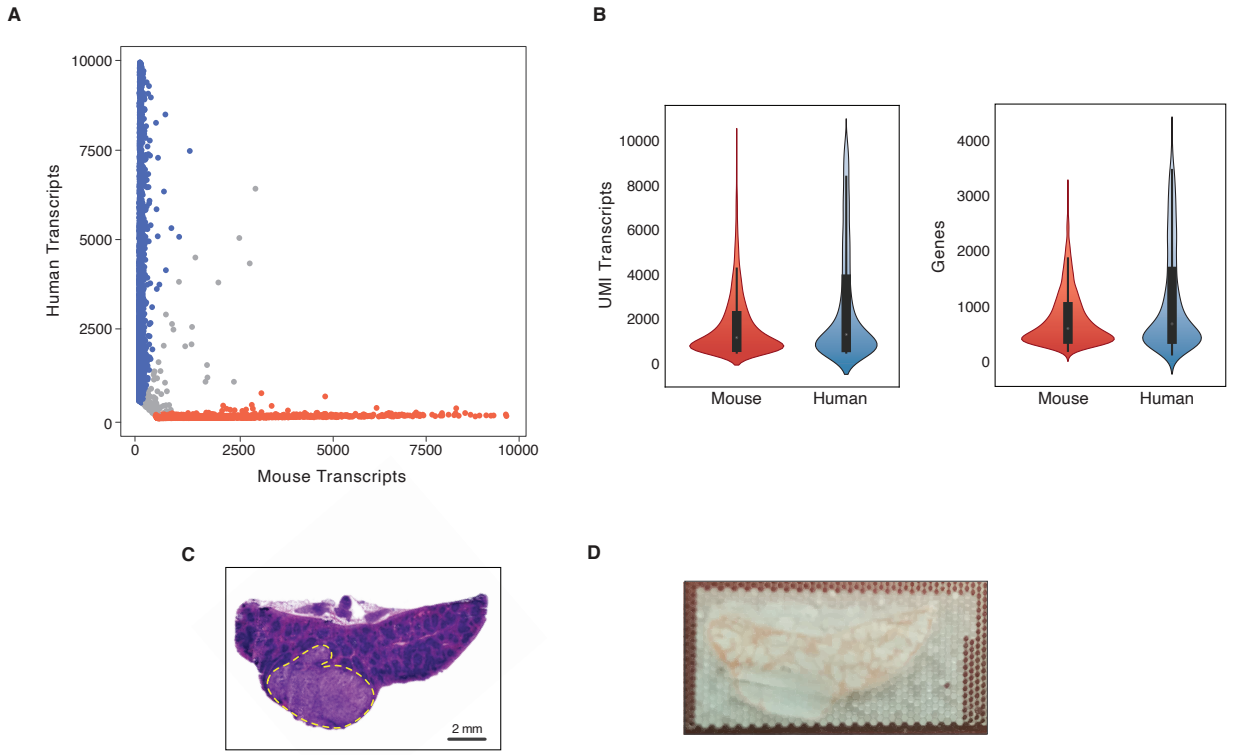


Figure A.11: XYZeq of spleen/tumor tissue reveals comparable data quality to liver/tumor tissue. (A) Scatter plot of mouse (x-axis) and human (y-axis) UMI counts detected from spleen/tumor tissues ($n = 4$). Blue refers to human cells ($n = 4,007$) and red refers to mouse cells ($n = 3,394$) and gray refers to collisions ($n = 104$). (B) Violin plot showing the number of detected UMIs and genes per cell. Median UMI counts for human cells ($1,312$) and mouse cells ($1,169$). Median gene counts for human cells (661) and mouse cells (577). (C) Hematoxylin and Eosin (H&E) stained image of the spleen/tumor tissue slice. Tumor region (light purple with yellow dotted outline); spleen region (dark pink with dark purple features). Scale showing 2mm. (D) Image of spleen/tumor tissue on frozen XYZeq microarray with reagents in wells (white). Photo credit: Youjin Lee (UCSF).

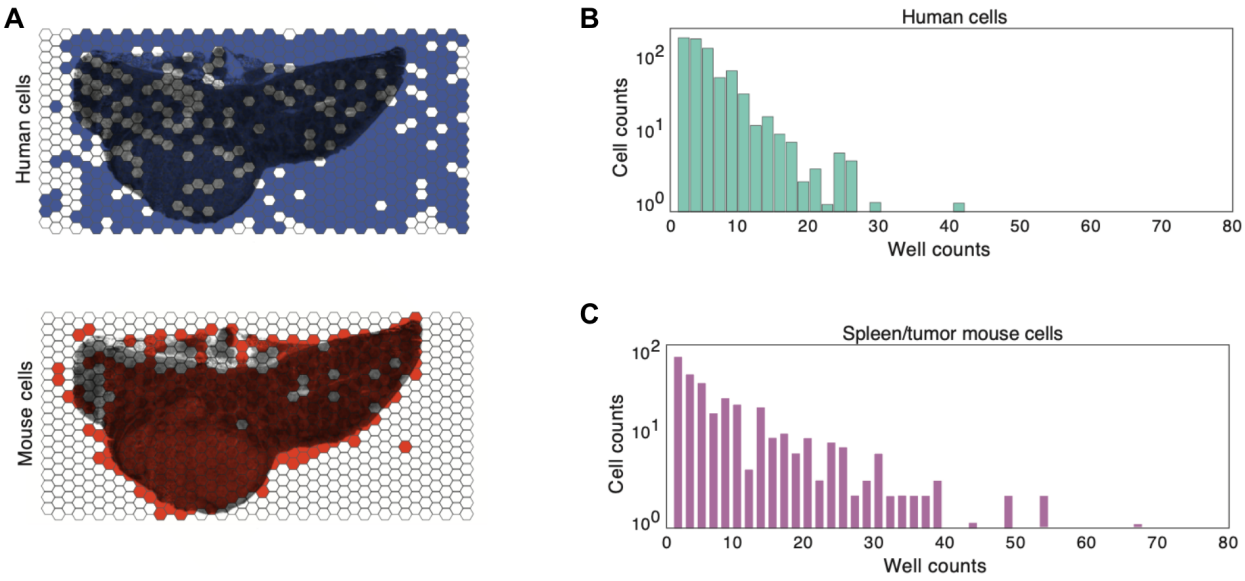


Figure A.12: (A) Visualization of human (blue) and mouse (red) cell distribution on the XYZeq array with 500 UMI cutoff overlaid on the image of H&E stained tissue slice. (B-C) Median cell number in wells across XYZeq array for HEK293T human (B) and spleen/tumor mouse (C) cells.

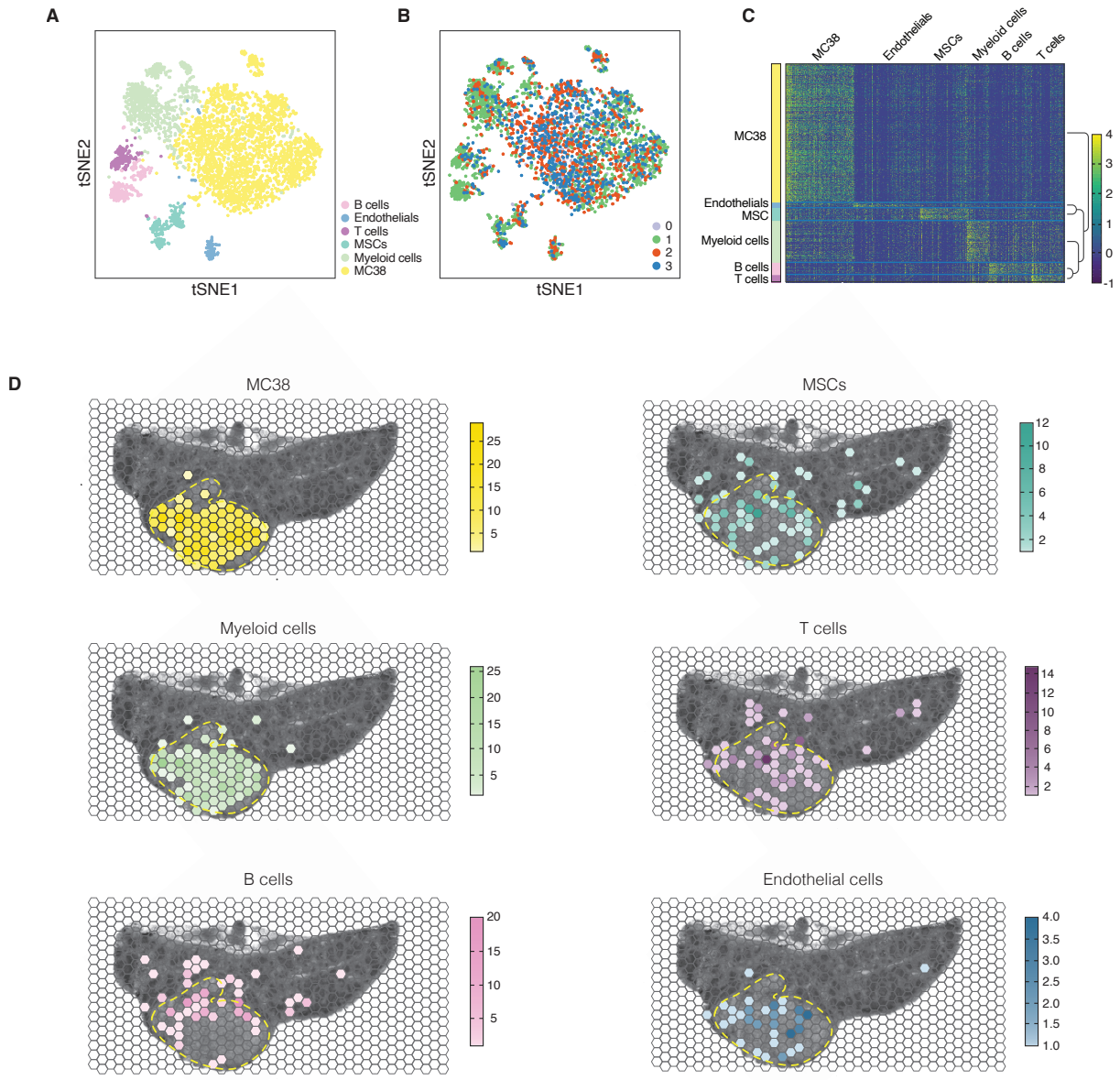


Figure A.13: Identification and spatial mapping of cell type clusters from spleen/tumor tissue. (A) tSNE projection of the spleen/tumor XYZeq data. 3,394 total cells were plotted. (B) tSNE visualization of Leiden cluster to annotate cell types for spleen/tumor. (C) Heat map of a scaled expression of marker genes and hierarchical clustering that define each cell type from XYZeq spleen/tumor tissue. (D) Image of spleen/tumor tissue overlaid with spatial plot of the XYZeq array showing localization of cell type clusters from (A) with 500 UMI cutoff. Yellow dotted outline indicates tumor region.

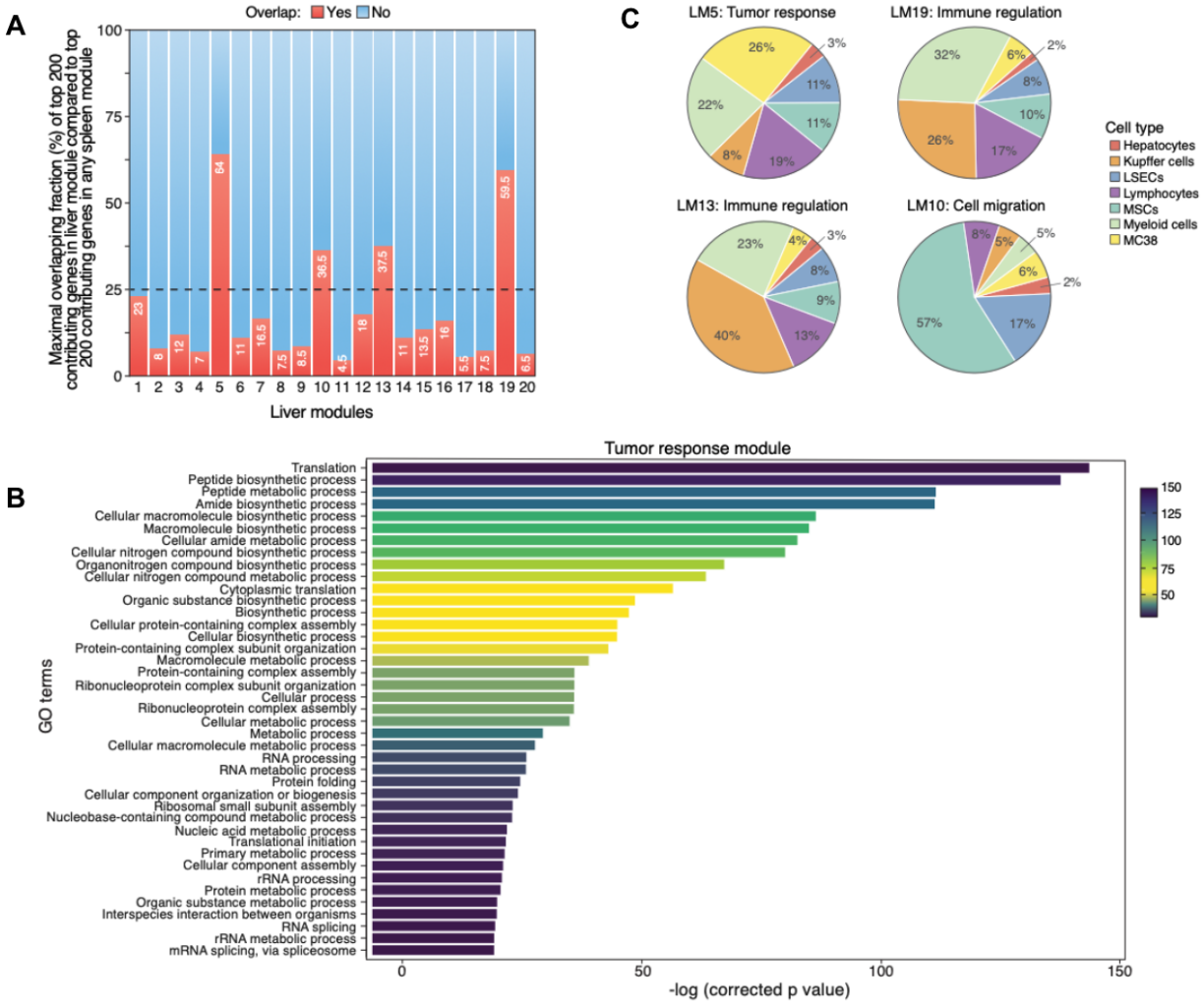


Figure A.14: Cell type contribution and functional annotation of gene modules. (A) Barplot showing percent fraction of overlapping genes in liver/tumor modules compared to corresponding spleen/tumor modules. Dotted line represents the threshold used to determine significant overlap between the modules. (B) GO annotations for tumor response module (LM5, SM2/12). P-values computed using GOrilla and adjusted by Benjamini-Hochberg correction. LM and SM denote liver/tumor modules and spleen/tumor modules respectively. (C) Pie chart representation of cell type fractions that make up each modules (See Methods).

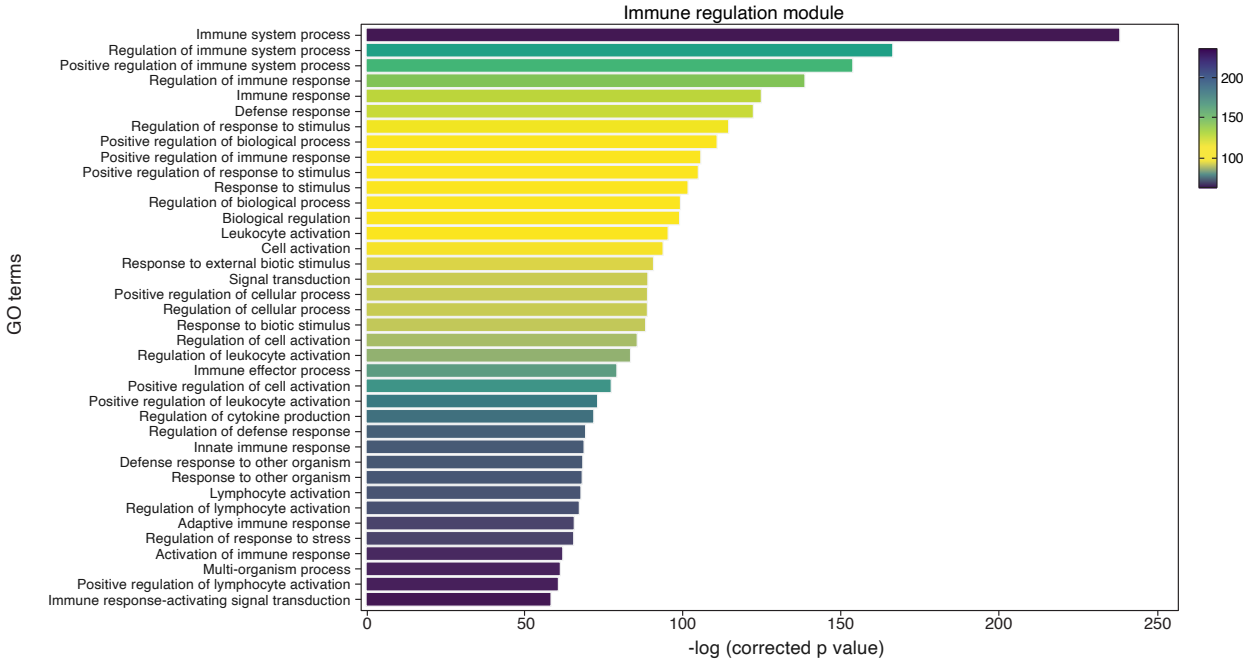


Figure A.15: GO annotations for immune regulation module (LM19, SM7/20). P-values computed using GOrilla and adjusted by Benjamini-Hochberg correction. LM and SM denote liver/tumor modules and spleen/tumor modules respectively.

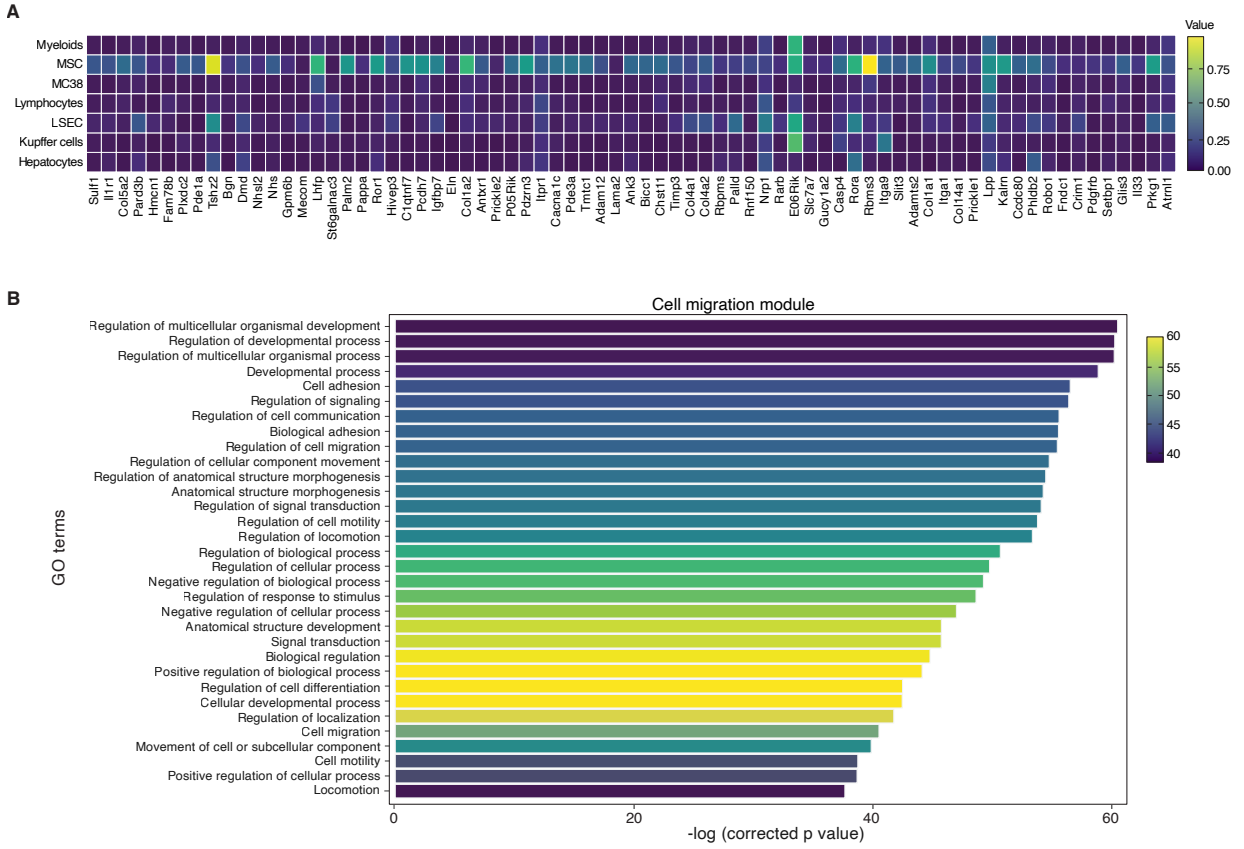
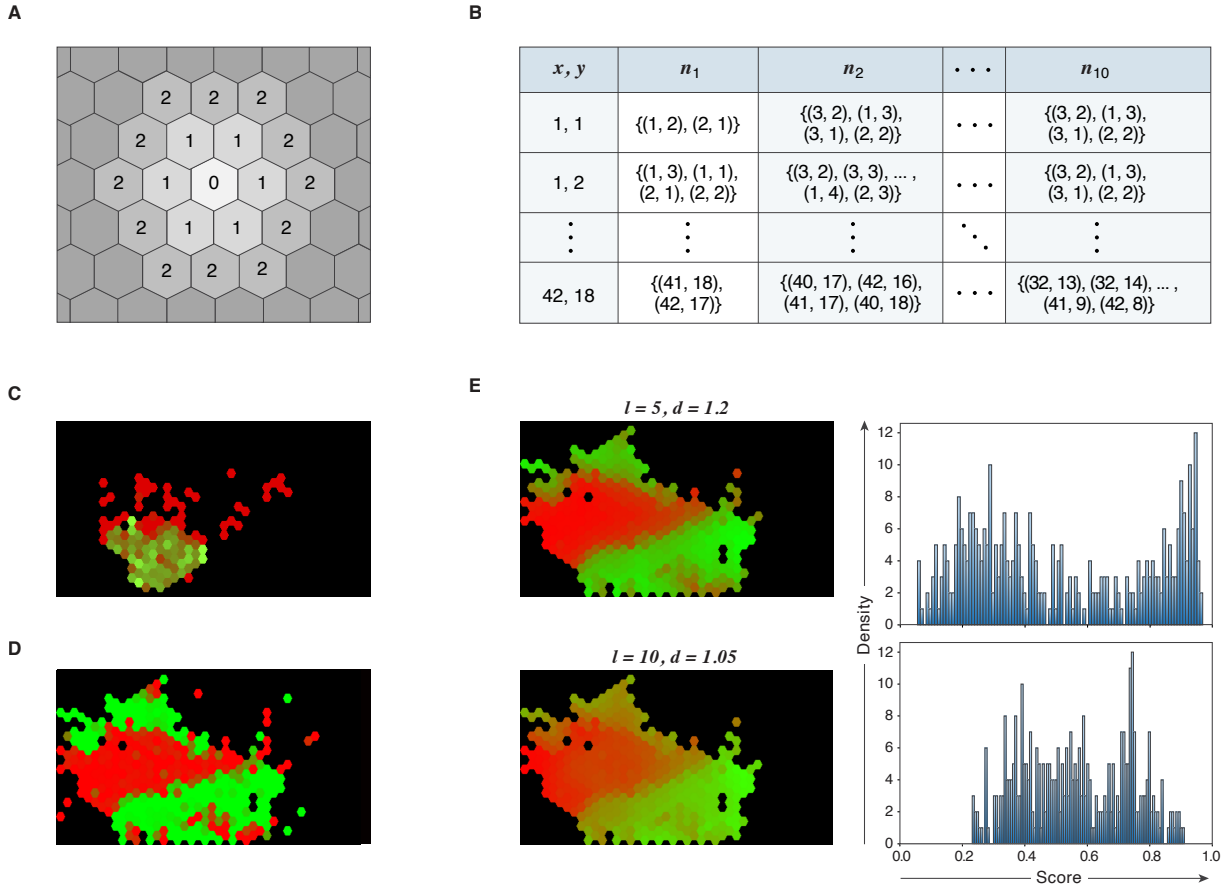


Figure A.16: Expression of the cell migration gene module enriched in MSCs. (A) Matrix plot of top overlapping genes in the cell migration module (LM10) across all cell types in liver/tumor. (B) GO annotation for cell migration module from LM10 and SM15/17. P-values computed using GOrilla and adjusted by Benjamini-Hochberg correction.



x, y	n_1	n_2	$\cdots$	n_{10}
1, 1	{(1, 2), (2, 1)}	{(3, 2), (1, 3), (3, 1), (2, 2)}	$\cdots$	{(3, 2), (1, 3), (3, 1), (2, 2)}
1, 2	{(1, 3), (1, 1), (2, 1), (2, 2)}	{(3, 2), (3, 3), ... , (1, 4), (2, 3)}	$\cdots$	{(3, 2), (1, 3), (3, 1), (2, 2)}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
42, 18	{(41, 18), (42, 17)}	{(40, 17), (42, 16), (41, 17), (40, 18)}	$\cdots$	{(32, 13), (32, 14), ... , (41, 9), (42, 8)}

Figure A.17: Tumor proximity score defined for both liver and spleen tissue. (A) The proximity scores for each tissue relied on the annotation of successive concentric layers of neighbors for a well in question. (B) The set of wells neighboring each well in the array were tabulated for up to 10 layers. (C) The cell-containing wells of representative spleen/tumor slice, where bright green indicates a higher proportion of tumor cells, and bright red indicates a higher proportion of non-tumor cells. (D) The wells containing cells of a representative liver/tumor slice. Bright green indicates higher proportion of tumor cells, bright red indicates higher proportion of hepatocytes. (E) The proximity score values annotated on each well (left top/bottom panels) from the liver tissue, where bright green is closer to the minimum value and bright red is closer to the maximum value. The scores are visualized for different values of l and d. The values of $l = 10$ and $d = 1.05$ were chosen as they rendered the distribution of scores (right top/bottom panels) more uniform across all wells.

A.3 Supplementary Tables

Modules	Overlapping genes
Tumor response modules (128 genes, LM5)	Lgals1 Rpl8 Rpsa Ly6e Rps12 S100a4 Rplp1 Tapbp S100a6 Rpl26 Ptma Rpl14 Rpl38 Rps24 Rps8 Vim Hsp90ab1 B2m Gnb2l1 Rpl21 Ptrf Rplp0 Col3a1 Rpl30 Rps20 Rps26 Rps19 Npm1 Rpl13 Rpl37 Rps5 Rps29 Ly6a Rpl32 Pfn1 Rps15 Rps15a Pfdn5 Ncl Tuba1c Rpl23a Ahnak Rpl13a Rps21 Arpc1b Eef1g Rpl41 Cct7 Rpl37a Rpl23 Calm1 Myh9 Psmc5 Hdgf Msn Pkm Rpl12 Fau Rps27 Rps14 Serpinh1 Cd63 Rpl9 Pdia3 Eef2 Rpl17 Arhgdia Cfl1 Rpl4 Rpn2 Edf1 Ccnd1 Ftl1 Sparc Rps25 Hspa8 Rplp2 Plec Rps10 Scd2 Cyb5r3 Serbp1 Rps17 Cct5 Rcn1 Rpl5 Rpl19 H1f0 Grn Rps27a Rpl11 Tuba1a P4hb Rpl36 Hnrnpab Rpl36a Pabpc1 Eif5a Eef1b2 Hspa9 Bsg Rpl27 Rps3 Rpl22 Ybx1 Psmb10 Tmsb10 Col6a1 Ecm1 Fth1 Set Cald1 Canx Tubb5 Cox4i1 Vcp Eef1a1 S100a10 Rpl34 Ldha Gabarap Rpl18a Hint1 Ppp1ca Kpnb1 Tcp1 Rps28 Pdap1
Immune regulation modules (75 genes, LM13)	Ptprj Fyb Mertk Aoah Adap2 Pou2f2 Adgre1 Myo9a Unc93b1 Dock10 Frmd4b Zdhhc14 Myo1f Prkcb Zeb2 Tmcc3 Zfp710 Slc9a9 Itga9 Lcp2 Hck Fcer1g Mrc1 Rreb1 Spi1 Ccr5 1700112E06Rik Elmo1 March1 Abcg3 Trpm2 Lyn Gm5150 Slc8a1 Nrros Ctsc Gab2 Sirpa Ly86 P2rx4 Pla2g4a Entpd1 Cd84 Fgr Acer3 Cd300lf Cadm1 Tmsb4x Msr1 Maf Mitf Lgmn Csf1r Ncf2 Ptprc Gmip Pik3ap1 Ctsb Lacc1 Arhgap15 Dock8 Plcl2 Shtn1 Prex1 Diaph2 Exoc6 Dock4 Mgat5 Slc43a2 Arrb2 Ppm1h Slc15a3 Pld4 Tbxas1 Vrk2

Table A.1: Lists of overlapping genes among top 200 contributing genes between liver and spleen, for tumor response modules (LM5) and immune regulation modules (LM13).

Modules	Overlapping genes
Immune regulation modules (119 genes, LM19)	Tgfbi Cd74 Arhgap15 Ccr5 Ctss Ly86 Myo1f 1700112E06Rik Dock10 Elmo1 Tgfbr1 Inpp5d Epsti1 Zeb2 Ccr2 H2-Ab1 Ptprc Slc8a1 Dock2 Dock8 Laptm5 Arhgap30 Hck Ms4a6c Cd84 Lyn Ctsc Cyth4 Fyb Itga4 Pld4 Gab2 Lcp1 Cybb Unc93b1 Tbxas1 Gpr141 Hmha1 C1qa Tmsb4x Plcg2 March1 Cx3cr1 Gm4955 Ctsb Pde7b Cd300lf Sirpa Dock4 Zfp710 Nlrp1b Pik3cd Fam49b Klra2 Ms4a6b Slfn5 H2-Eb1 Aoah Lcp2 Csf1r Arhgap24 Abr Cers6 Entpd1 H2-Aa Ptk2b Sp100 H2-DMb1 Slfn2 Mpeg1 Aim2 Apobec1 Rreb1 Etv6 Adap2 Irf7 Spi1 Ms4a4c Gm2a Lyz2 Lgmn Stat1 Ms4a7 Tnfrsf11a Apba1 Pik3ap1 Fam105a Nckap1l Psap Arid5b Mndal Slfn8 Trim30a Lair1 Picalm Il10ra Ncf1 Ccdc88b Coro1a C1qc Gatm Tgm2 Ctsh C1qb Pip4k2a Chd7 Mx1 Samhd1 Fcgr2b Diaph2 Mitf Mef2a Pou2f2 Pid1 Ifi203 Wdfy4 Fnbp1 Dennd1b Pik3r5
Cell migration modules (73 genes, LM10)	Tshz2 Col1a2 Rbms3 Igfbp7 Kalrn Prkg1 Ccdc80 Rora Palm2 Timp3 Pdzrn3 Pde3a Adamts2 Ror1 Sulf1 Palld Pcdh7 C1qtnf7 Tmtc1 Pde1a Slit3 Adam12 Col14a1 Col5a2 Casp4 Col1a1 Ank3 Bicc1 Col4a2 Plxdc2 Il1r1 1700112E06Rik Lhfp Rnf150 Pard3b Cacna1c Crim1 Itga1 Itga9 9530026P05Rik Gucy1a2 Col4a1 Pdgfrb Nhs Hivep3 Fndc1 Prickle1 Chst11 Rarb Robo1 Il33 Gpm6b Itpr1 Bgn Mecom Slc7a7 Atrnl1 Rbpms Hmcn1 Pappa Lpp Phldb2 Dmd Antxr1 Lama2 St6galnac3 Fam78b Nrp1 Setbp1 Eln Nhsl2 Prickle2 Glis3

Table A.2: Lists of overlapping genes among top 200 contributing genes between liver and spleen, for immune regulation modules (LM19) and cell migration modules (LM10).

Appendix B

Supplementary Material for Chapter 3

Authors contributions

Y.W. designed and implemented the ggPair model. Y.W. and F.Z. carried out data analysis and benchmarking against existing methods. F.Z. implemented the state-of-the-art methods. Y.W. implemented the variance decomposition analysis. Y.W. and F.Z. performed data processing. F.Z. implemented the data split strategy and compiled the interaction annotations. Y.W. wrote the initial draft of the manuscript, which was reviewed and approved by all authors. Y.S.S. provided overall supervision and guidance for the project. All authors were involved in project conceptualization, identifying key research questions related to GPCR ligand identification.

Acknowledgments

Y.W. extends gratitude to Scott Viteri for many fruitful discussions on image classification. Frank Liu is acknowledged for preliminary data analysis. We also thank members of the Song Lab for helpful discussions. This research is supported in part by an NIH grant R35-GM134922.

B.1 Supplementary Figures

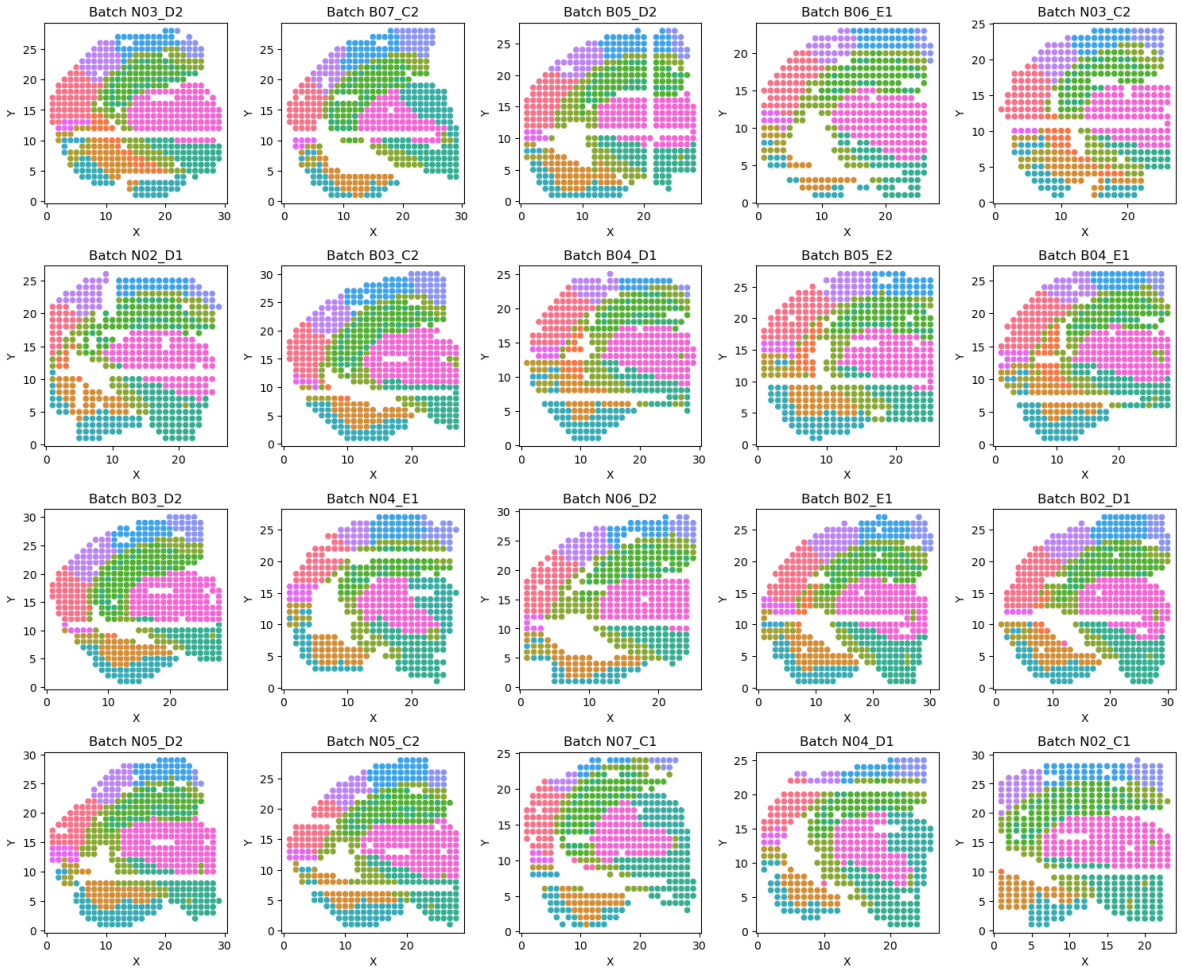


Figure B.1: The four spatial transcriptomic batches from mouse brain used in this study.

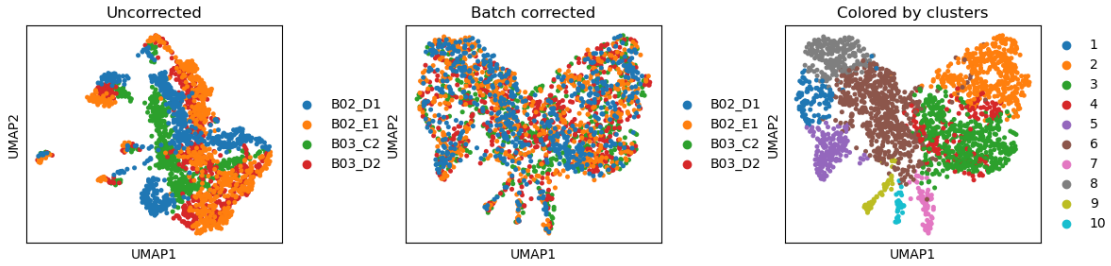


Figure B.2: Batch correction of the mouse brain data using the GraphST algorithm aligns the batches.

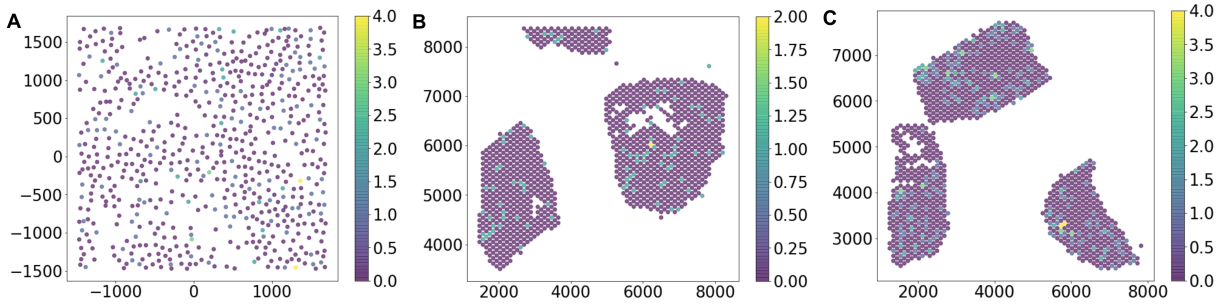


Figure B.3: Spatial gene expression of Itga9 gene in three human spatial transcriptomics data. A and B: human breast cancer 10X Visium data. C: human osteosarcoma MERFISH data. Each dot correponds to the normalized gene expression of Itga9 gene in the spatial spot. x and y axis correspond to the spatil image coordinate.

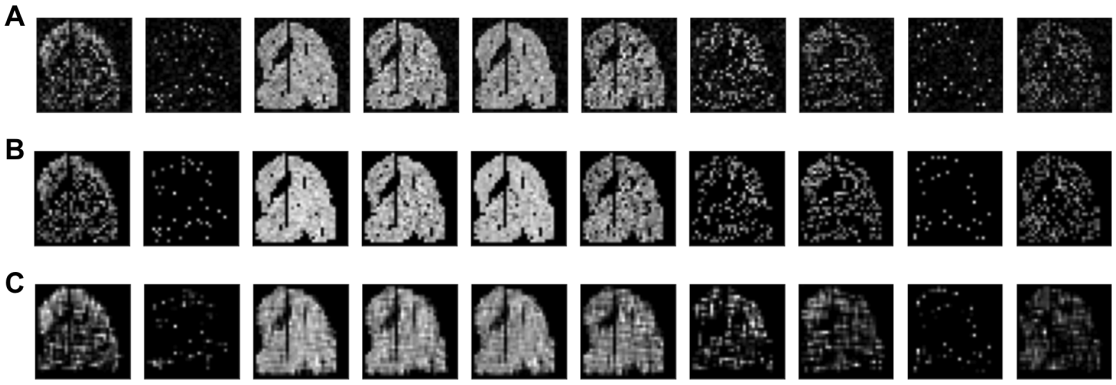


Figure B.4: Spatial gene expression of 10 randomly selected genes in the spatial batch B02-D1. Noisy images (A), original images (B), reconstructed images after CAE (C).

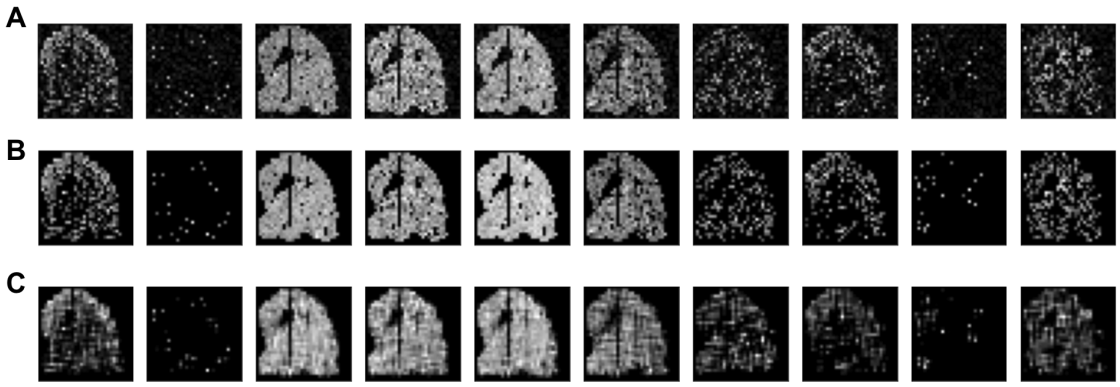


Figure B.5: Spatial gene expression of 10 randomly selected genes in the spatial batch B02-E1. Noisy images (A), original images (B), reconstructed images after CAE (C).

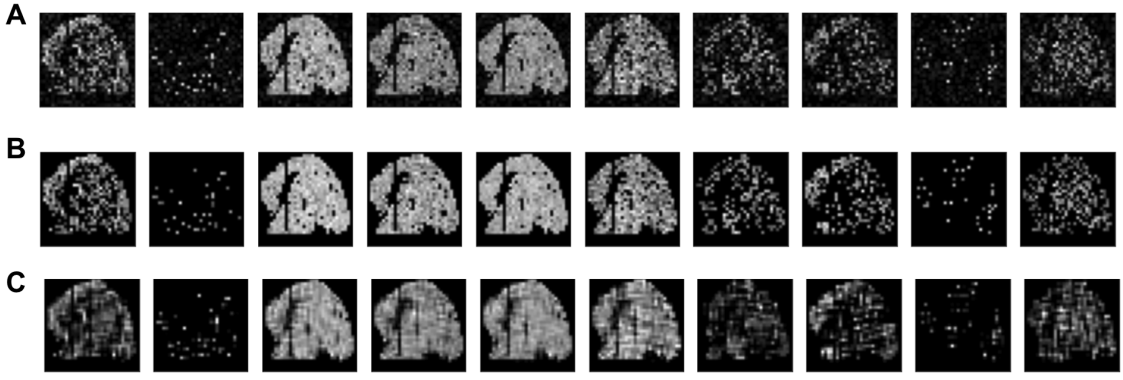


Figure B.6: Spatial gene expression of 10 randomly selected genes in the spatial batch B03-C2. Noisy images (A), original images (B), reconstructed images after CAE (C).

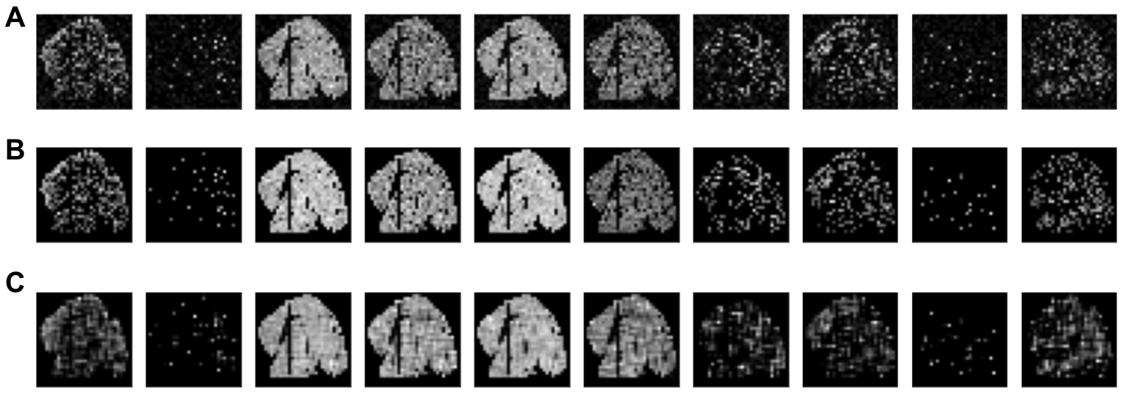


Figure B.7: Spatial gene expression of 10 randomly selected genes in the spatial batch B03-D2. Noisy images (A), original images (B), reconstructed images after CAE (C).

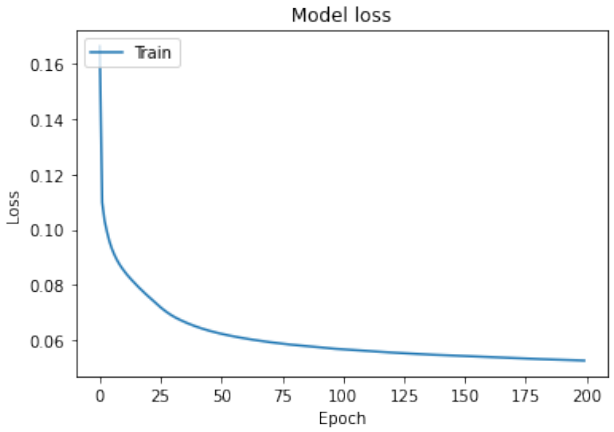


Figure B.8: CAE Training loss curve on mouse data.

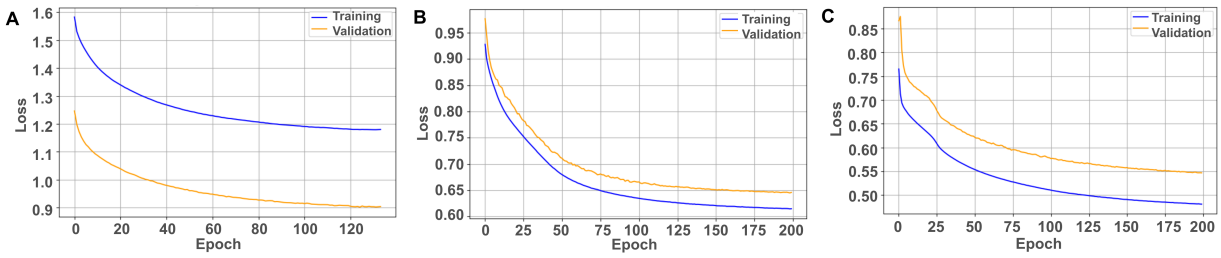


Figure B.9: CAE loss curves on training (blue) and validation (yellow) data. Human osteosarcoma MERFISH data (A), human breast cancer 10X Visium data (B-C).

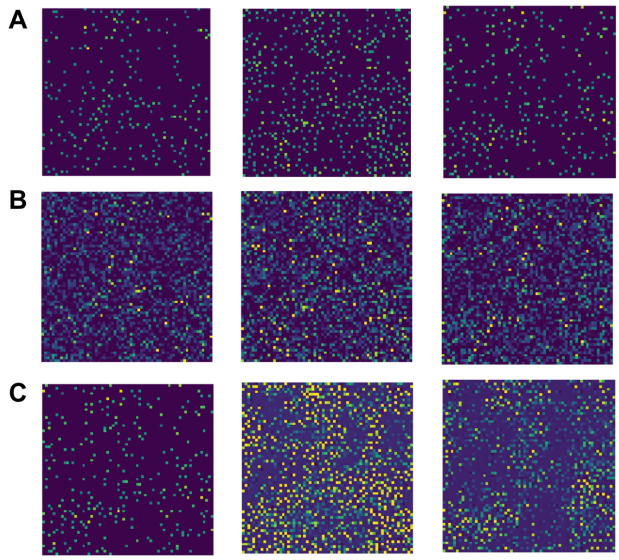


Figure B.10: Spatial gene expression for osteosarcoma MERFISH data. Original images (A), noisy images (B), and reconstructed images after CAE (C).

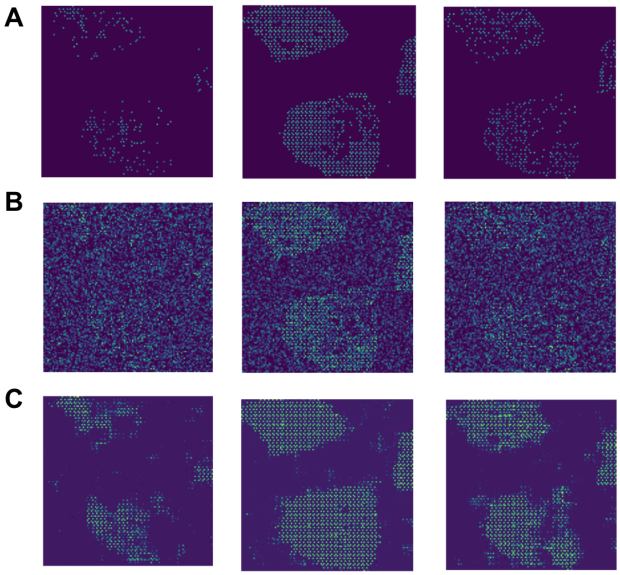


Figure B.11: Spatial gene expression of randomly selected genes in the first human breast cancer 10X Visium data. Original images (A), noisy images (B), and reconstructed images after CAE (C).

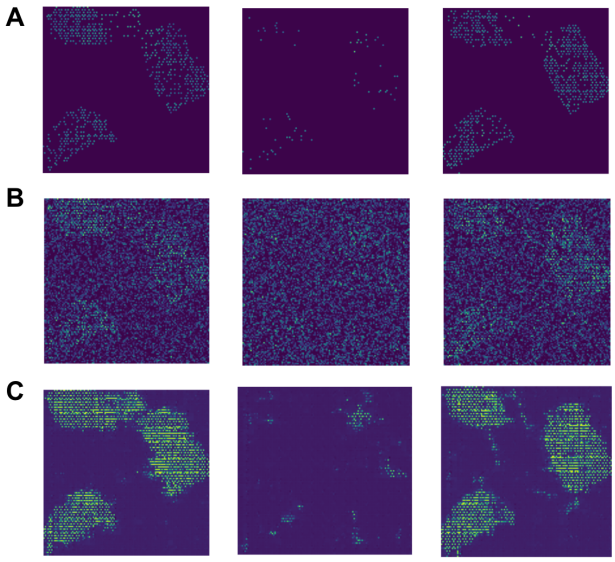


Figure B.12: Spatial gene expression of randomly selected genes in the second human breast cancer 10X Visium data. Original images (A), noisy images (B), and reconstructed images after CAE (C).

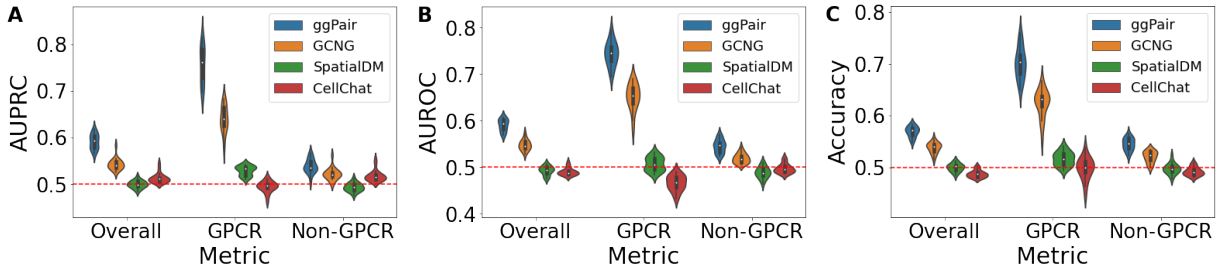


Figure B.13: Model performance comparison using 5-fold cross-validation on the mouse data, ggPair in blue, GCNG in orange, and SpatialDM in green. AUPRC (A), AUROC (B), Accuracy (C).

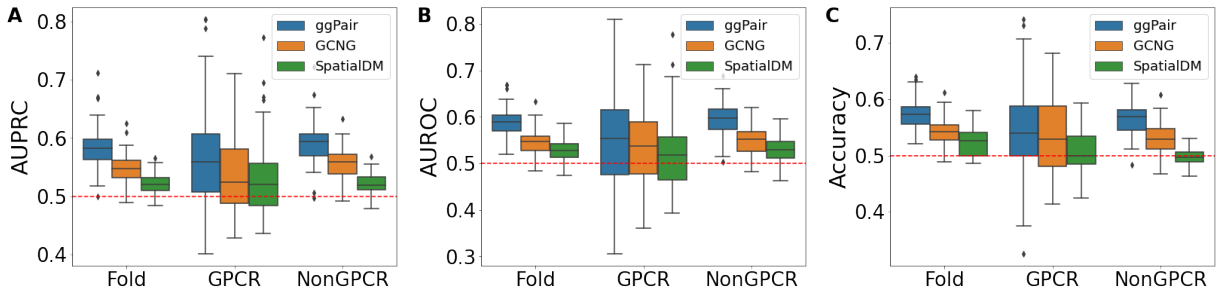


Figure B.14: Model performance comparison using 5-fold cross-validation on the second human breast cancer data, ggPair in blue, GCNG in orange, and SpatialDM in green. AUPRC (A), AUROC (B), Accuracy (C).

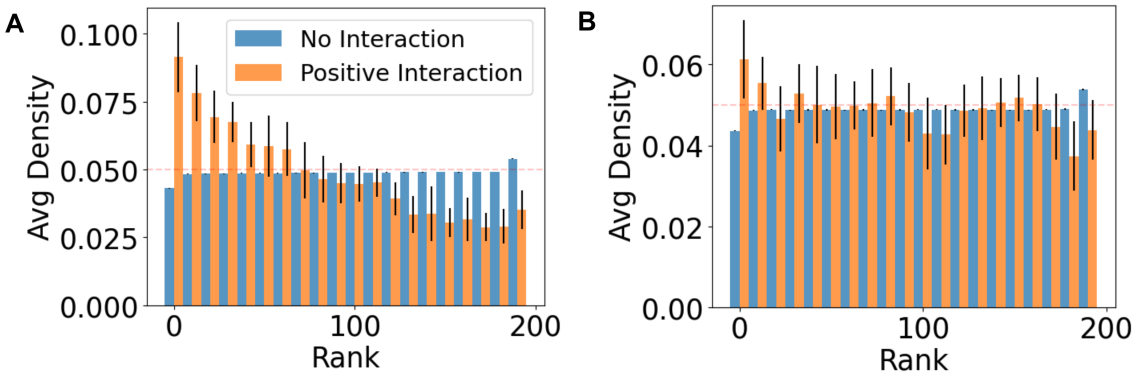


Figure B.15: The average density of the ligand candidates ranked by the predicted class probabilities, stratified by the true interaction label for each gene pair. In each LOOCV iteration, every nonGPCR receptor was held out for testing while all the other receptors were used for training. The hold-out nonGPCR receptor was then paired with all the possible ligand candidates for prediction. ggPair (A), GCNG (B).

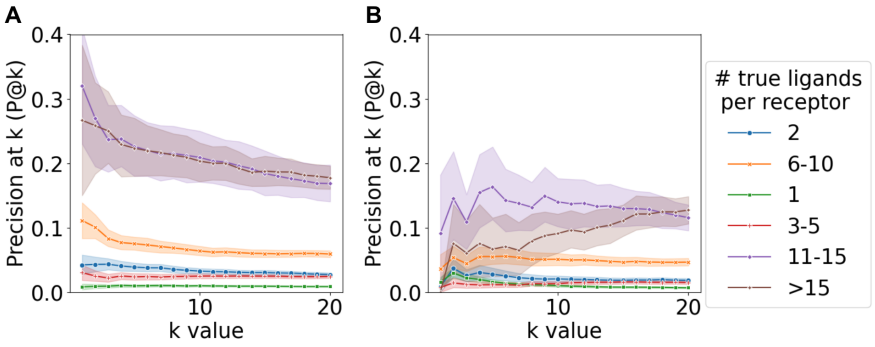


Figure B.16: Precision at the top k call(s) for nonGPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B).

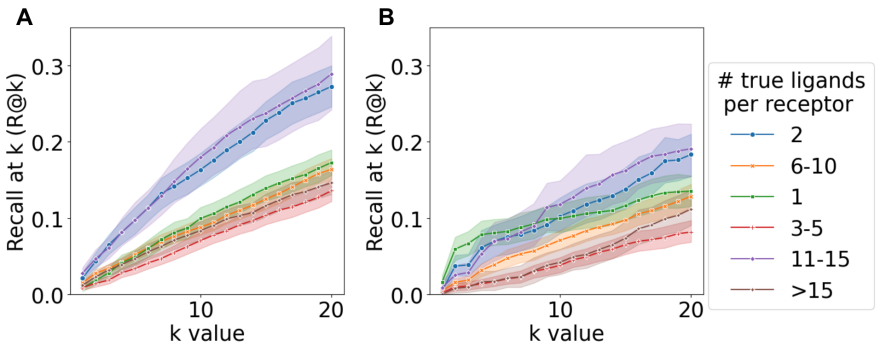


Figure B.17: Recall at the top k call(s) for nonGPCR receptors, stratified by the number of true interacting ligands. ggPair (A), GCNG (B).

B.2 Supplementary Tables

Table B.1: Comparison of methods based on AUPRC for GPCR, Non-GPCR, and Overall data. "CI lower" and "CI upper" denote the lower and upper bounds of the 95% confidence interval, respectively.

Method	**Mean**	**Median**	**Std**	**CI lower**	**CI upper**
GPCR					
ggPair	**0.7562**	**0.7614**	0.0401	**0.7386**	**0.7738**
GCNG	0.6404	0.6398	0.0285	0.6279	0.6529
SpatialDM	0.5297	0.5333	0.0104	0.5251	0.5342
CellChat	0.4953	0.4976	0.0120	0.4900	0.5006
Non-GPCR					
ggPair	**0.5394**	**0.5355**	0.0152	**0.5328**	**0.5461**
GCNG	0.5226	0.5214	0.0129	0.5169	0.5282
SpatialDM	0.4941	0.4937	0.0091	0.4902	0.4981
CellChat	0.5185	0.5155	0.0122	0.5131	0.5238
Overall					
ggPair	**0.5921**	**0.5946**	0.0161	**0.5851**	**0.5991**
GCNG	0.5435	0.5411	0.0123	0.5381	0.5489
SpatialDM	0.5003	0.4992	0.0081	0.4968	0.5038
CellChat	0.5136	0.5122	0.0106	0.5089	0.5182

Table B.2: Comparison of methods based on AUROC for GPCR, Non-GPCR, and Overall data. "CI lower" and "CI upper" denote the lower and upper bounds of the 95% confidence interval, respectively.

Method	**Mean**	**Median**	**Std**	**CI lower**	**CI upper**
GPCR					
ggPair	**0.7444**	**0.7449**	0.0260	**0.7330**	**0.7558**
GCNG	0.6515	0.6541	0.0284	0.6391	0.6640
SpatialDM	0.5072	0.5064	0.0150	0.5006	0.5138
CellChat	0.4664	0.4666	0.0167	0.4590	0.4737
Non-GPCR					
ggPair	**0.5456**	**0.5480**	0.0144	**0.5393**	**0.5520**
GCNG	0.5194	0.5176	0.0105	0.5149	0.5240
SpatialDM	0.4891	0.4874	0.0123	0.4837	0.4945
CellChat	0.4978	0.4948	0.0102	0.4933	0.5023
Overall					
ggPair	**0.5908**	**0.5938**	0.0129	**0.5851**	**0.5964**
GCNG	0.5480	0.5460	0.0113	0.5430	0.5530
SpatialDM	0.4932	0.4943	0.0095	0.4890	0.4974
CellChat	0.4906	0.4878	0.0082	0.4870	0.4942

Table B.3: Comparison of methods based on Accuracy for GPCR, Non-GPCR, and Overall data. "CI lower" and "CI upper" denote the lower and upper bounds of the 95% confidence interval, respectively.

Method	**Mean**	**Median**	**Std**	**CI lower**	**CI upper**
GPCR					
ggPair	**0.7134**	**0.7147**	0.0291	**0.7007**	**0.7262**
GCNG	0.6271	0.6328	0.0279	0.6149	0.6393
SpatialDM	0.5134	0.5141	0.0134	0.5075	0.5193
CellChat	0.4963	0.4958	0.0162	0.4892	0.5034
Non-GPCR					
ggPair	**0.5415**	**0.5414**	0.0112	**0.5366**	**0.5464**
GCNG	0.5184	0.5221	0.0111	0.5136	0.5233
SpatialDM	0.4973	0.4960	0.0098	0.4930	0.5017
CellChat	0.4932	0.4916	0.0082	0.4896	0.4968
Overall					
ggPair	**0.5696**	**0.5713**	0.0104	**0.5651**	**0.5742**
GCNG	0.5386	0.5404	0.0103	0.5341	0.5432
SpatialDM	0.5003	0.5022	0.0078	0.4969	0.5037
CellChat	0.4878	0.4872	0.0070	0.4848	0.4909

Appendix C

Supplementary Material for Chapter 4

C.1 Supplementary Figures

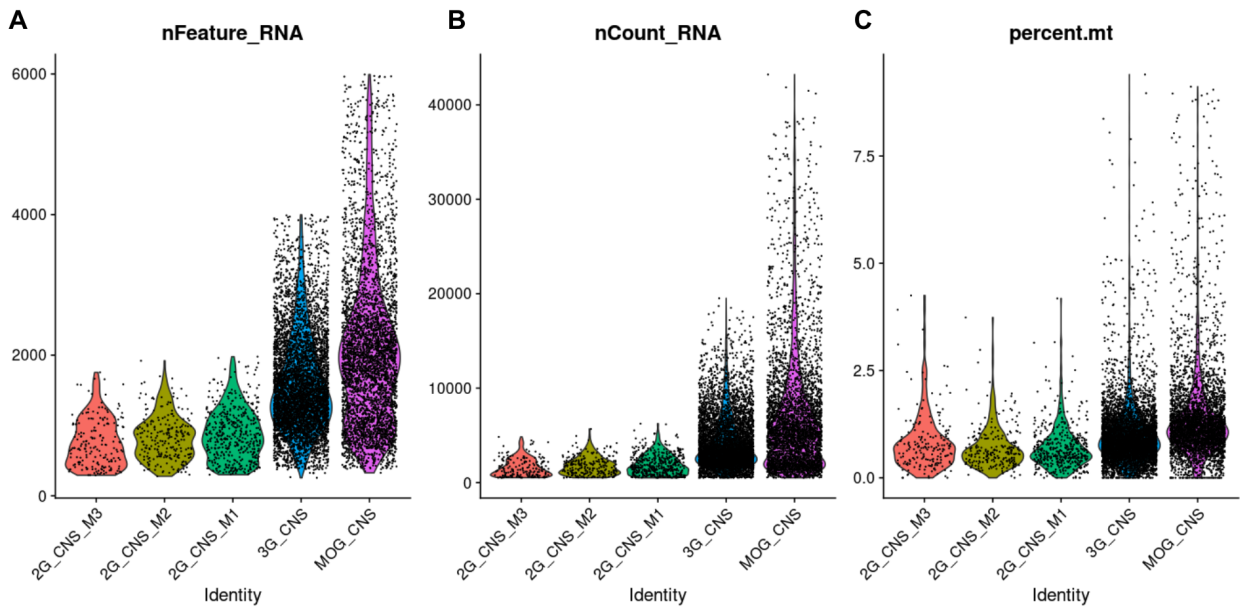


Figure C.1: Violin plots across different samples (batches) in the CNS, to visualize the distribution of the number of features per cell (A), the number of mRNA counts per cell (B), and the percentage of mitochondrial genes per cell (C). Each data point corresponds to one cell in the CNS sample.

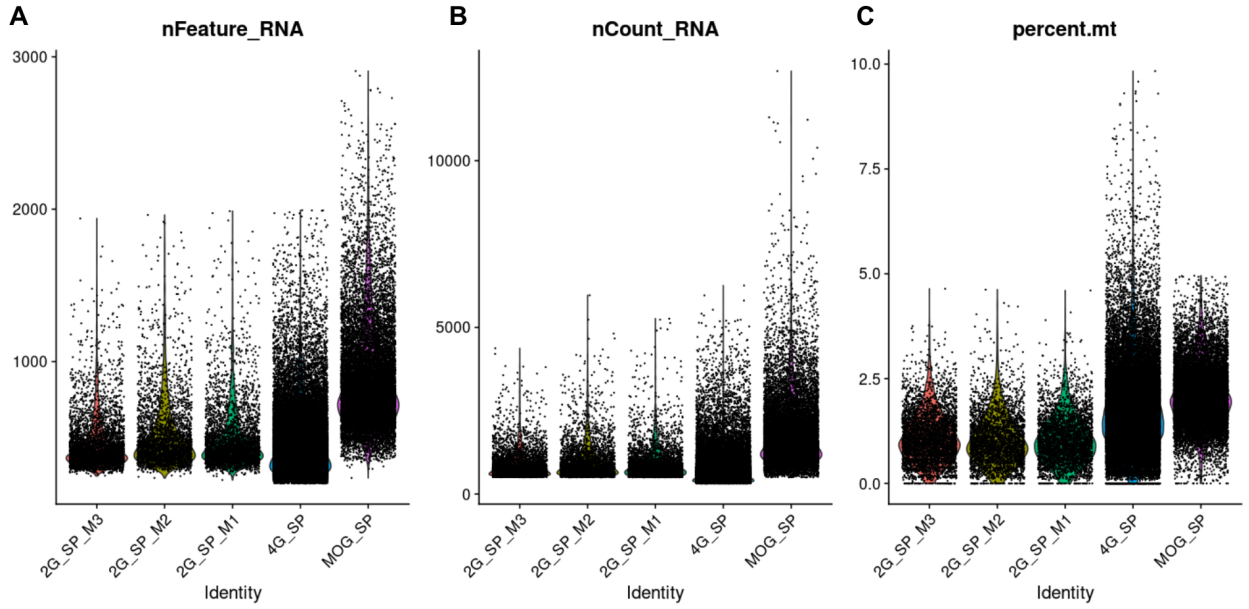


Figure C.2: Violin plots across different samples (batches) in the SP, to visualize the distribution of the number of features per cell (A), the number of mRNA counts per cell (B), and the percentage of mitochondrial genes per cell (C). Each data point corresponds to one cell in the SP sample.

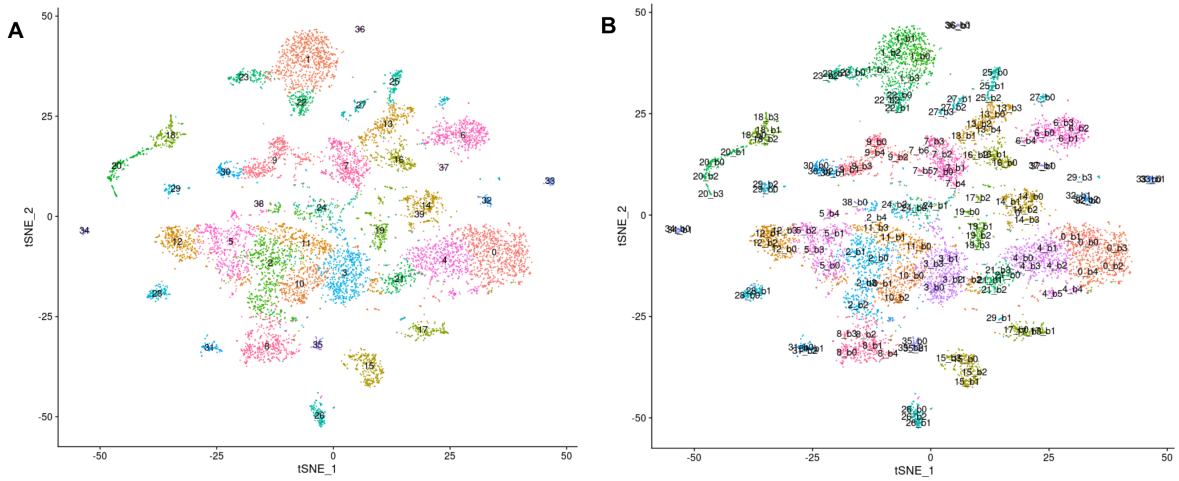


Figure C.3: Visualization of the global level 1 (A) and local level 2 (B) clusters for the scRNA-seq data in the CNS. Each data point corresponds to a cell, and the color corresponds to the cluster annotation in the transcriptional ambient space.

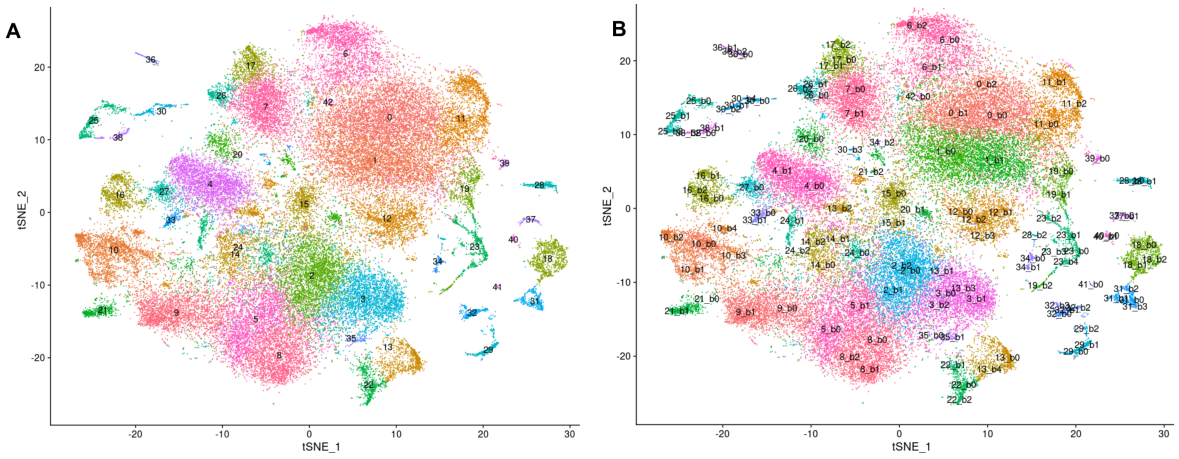


Figure C.4: Visualization of the global level 1 (A) and local level 2 (B) clusters for the scRNA-seq data in the SP. Each data point corresponds to a cell, and the color corresponds to the cluster annotation in the transcriptional ambient space.

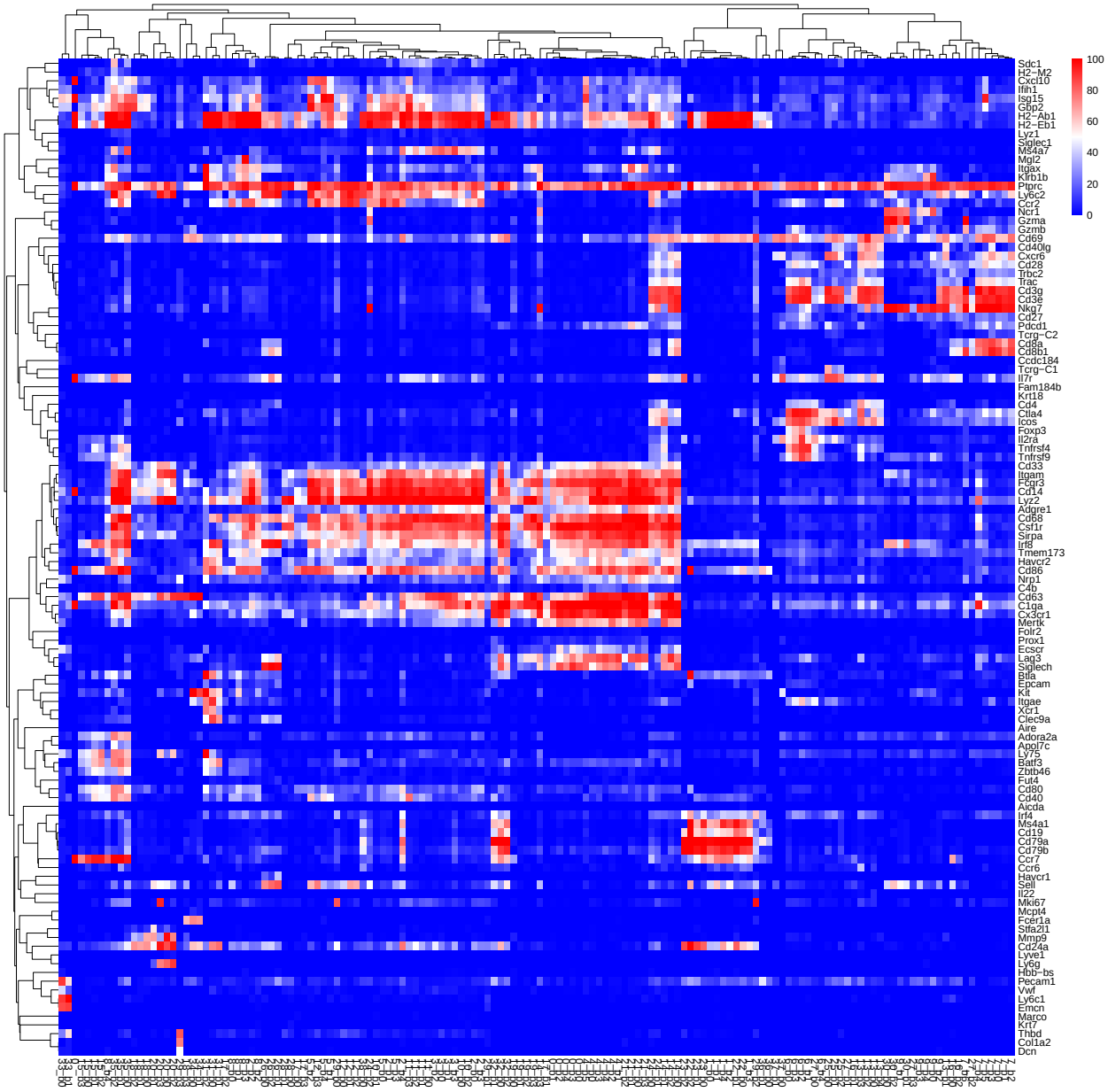


Figure C.5: CNS: each entry in the heatmap corresponds to the percentage of cells in a given level 2 cluster (cluster) that express a specific marker gene (row).

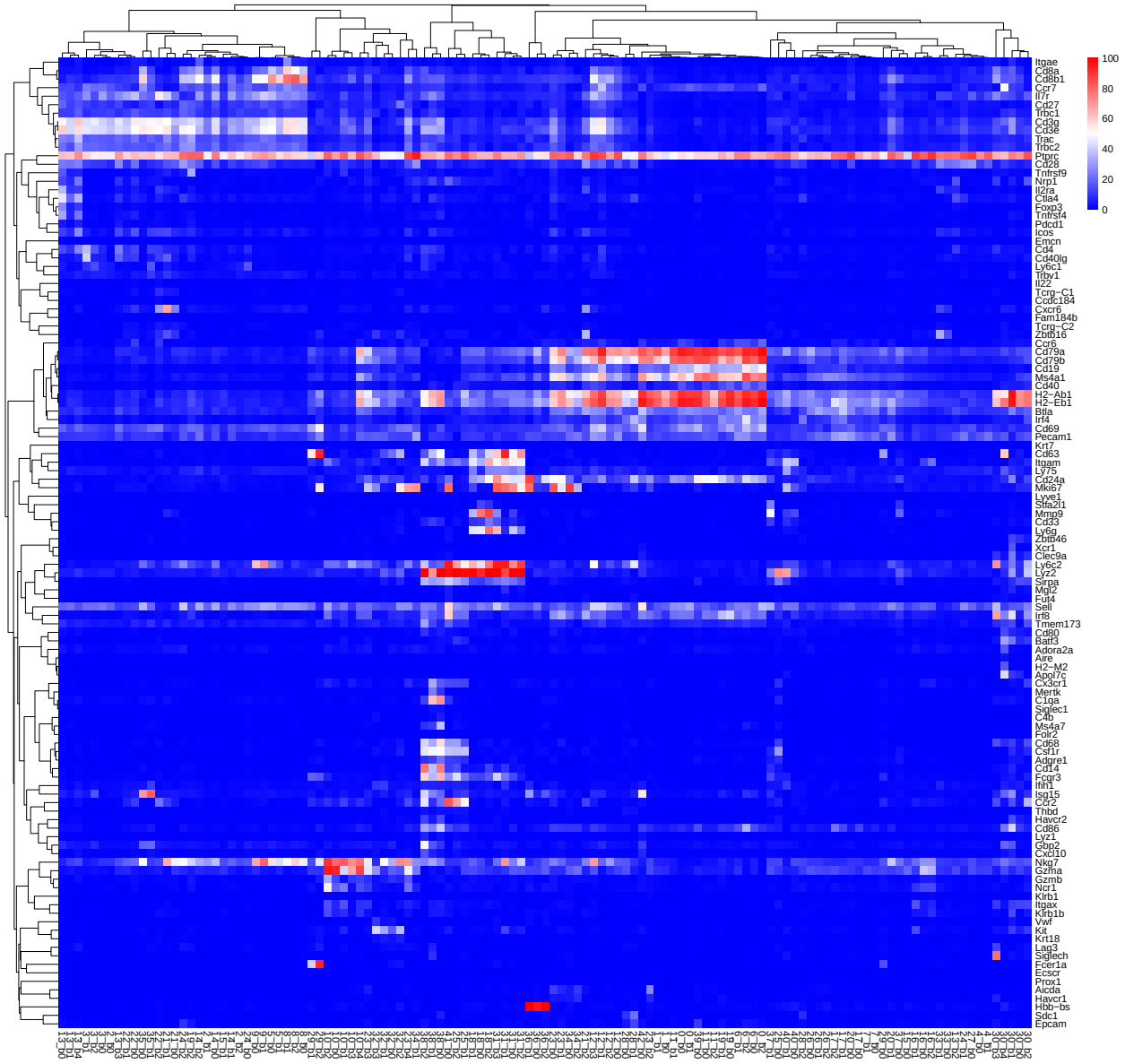


Figure C.6: SP: each entry in the heatmap corresponds to the percentage of cells in a given level 2 cluster (cluster) that express a specific marker gene (row).

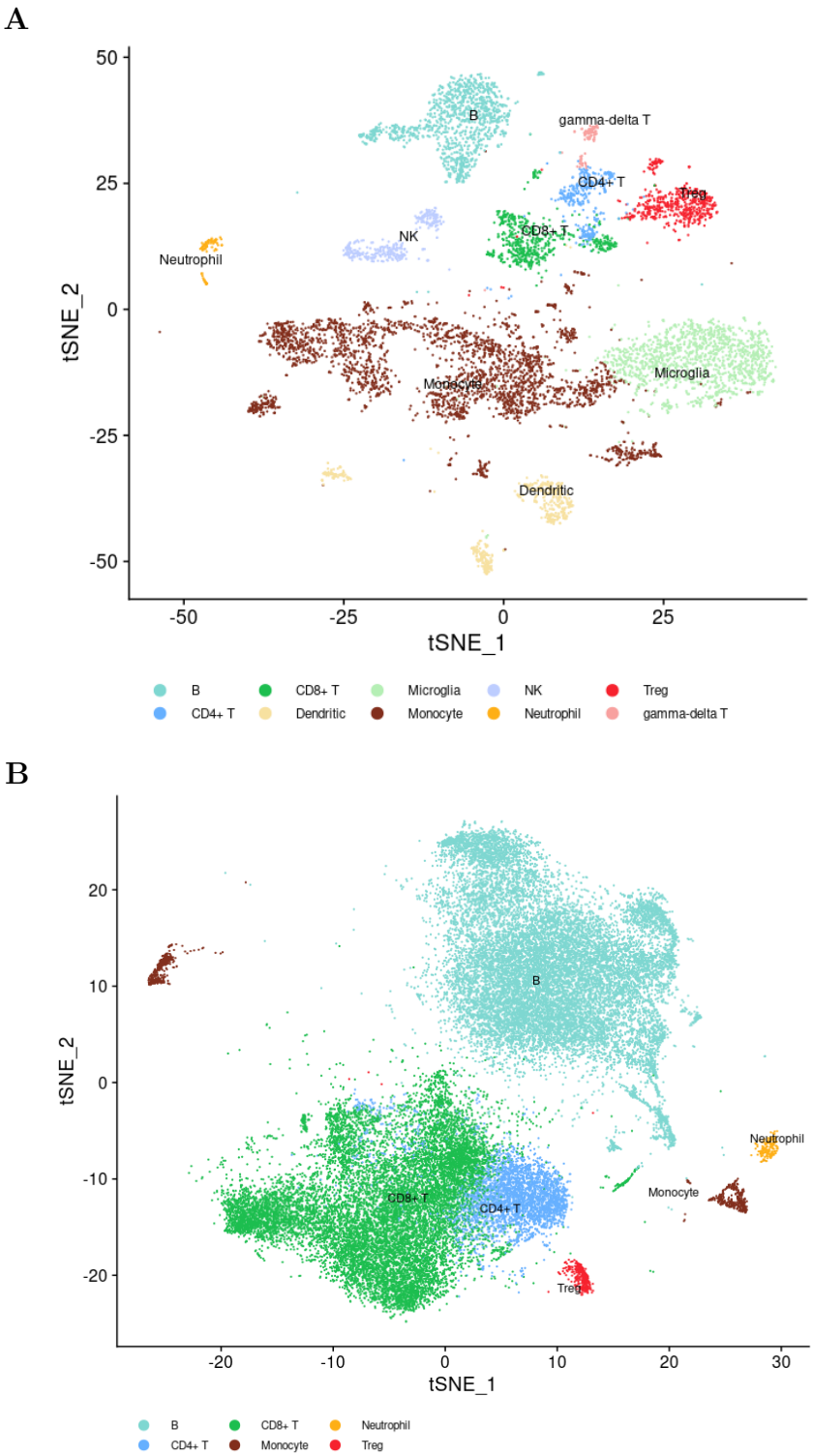


Figure C.7: Visualization of the cell type annotation for the scRNA-seq data in the CNS (A) and SP (B).

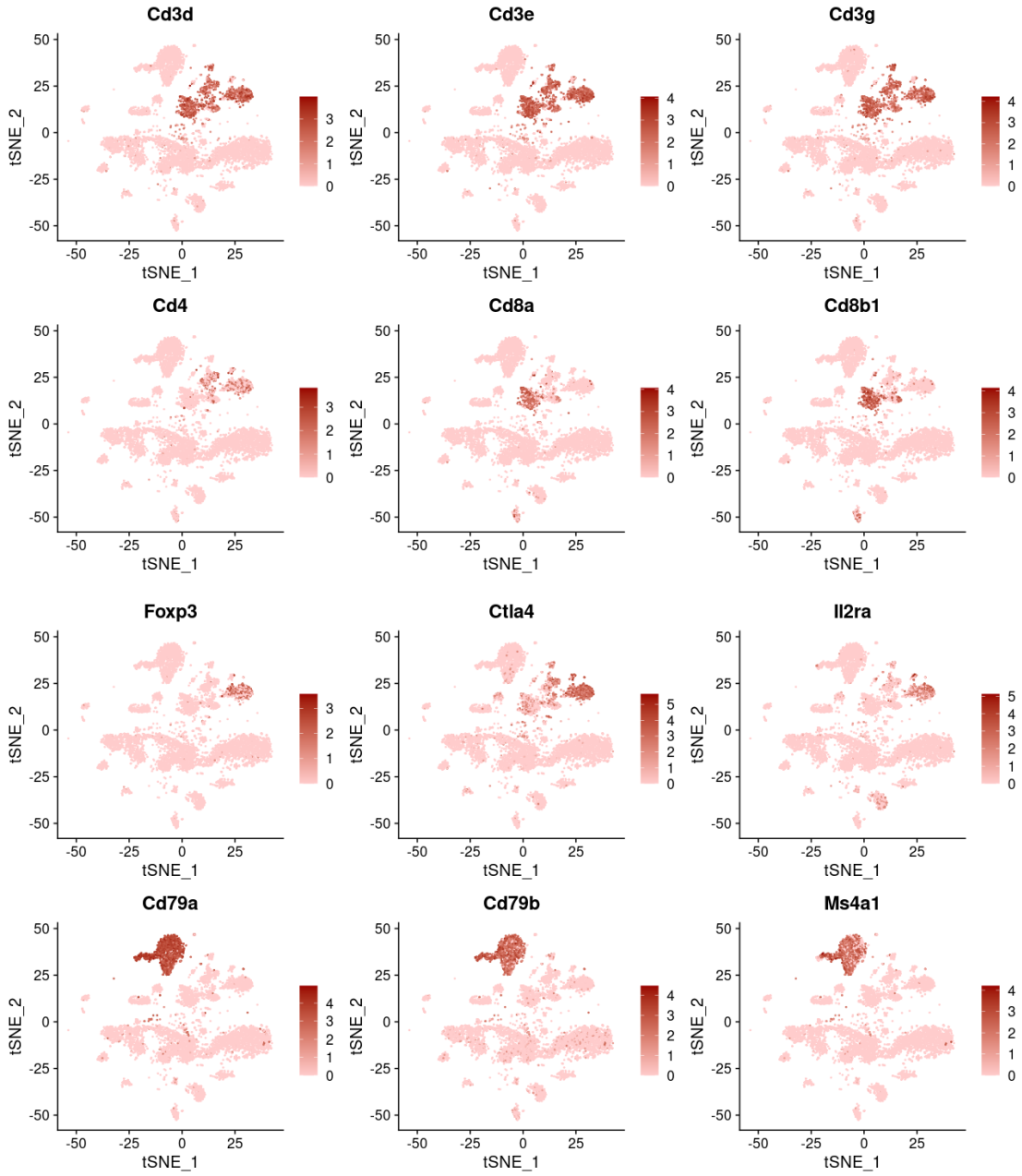


Figure C.8: Visualization of key marker genes in the CNS scRNA-seq data.

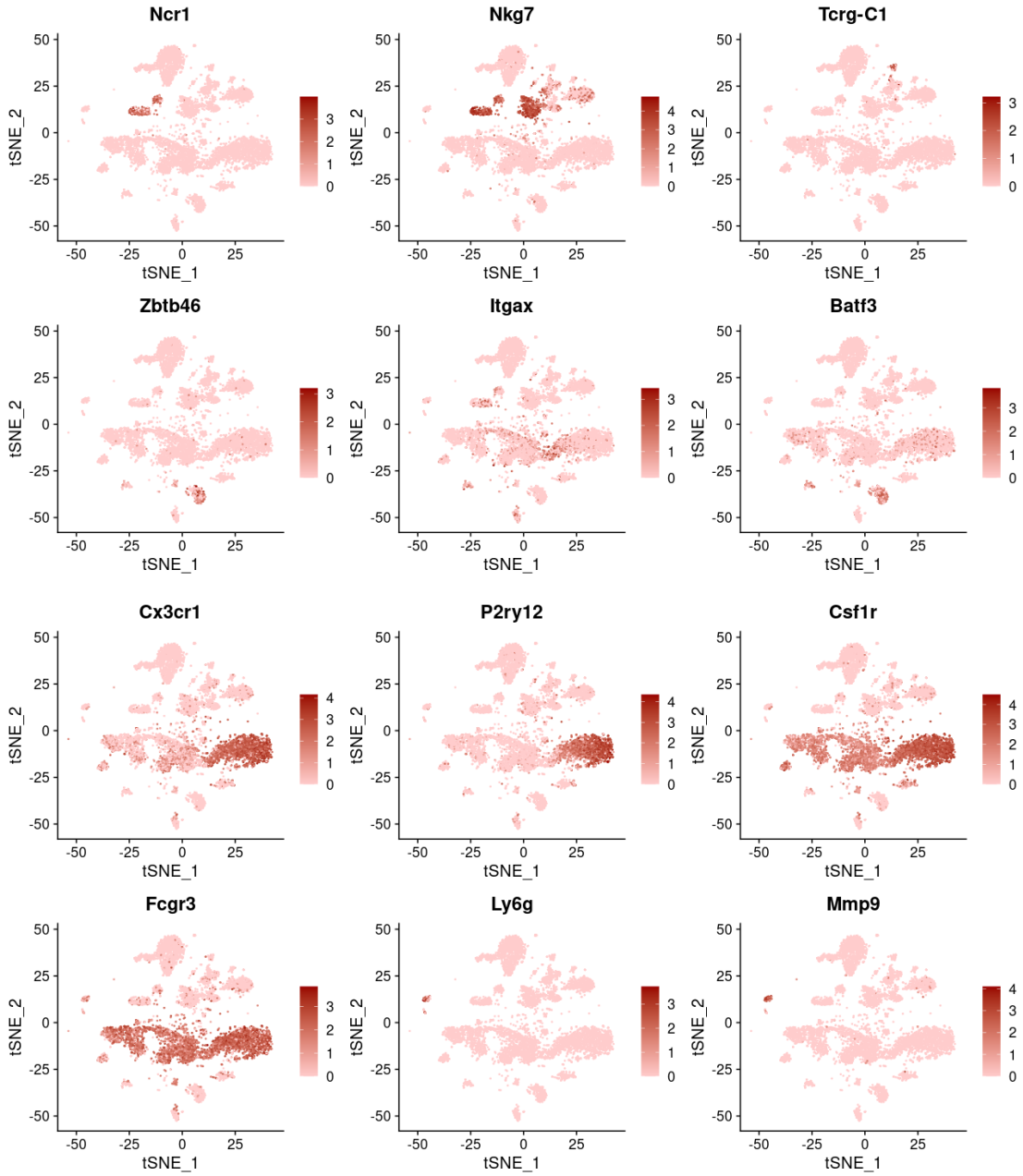


Figure C.9: Continued visualization of key marker genes in the CNS scRNA-seq data.

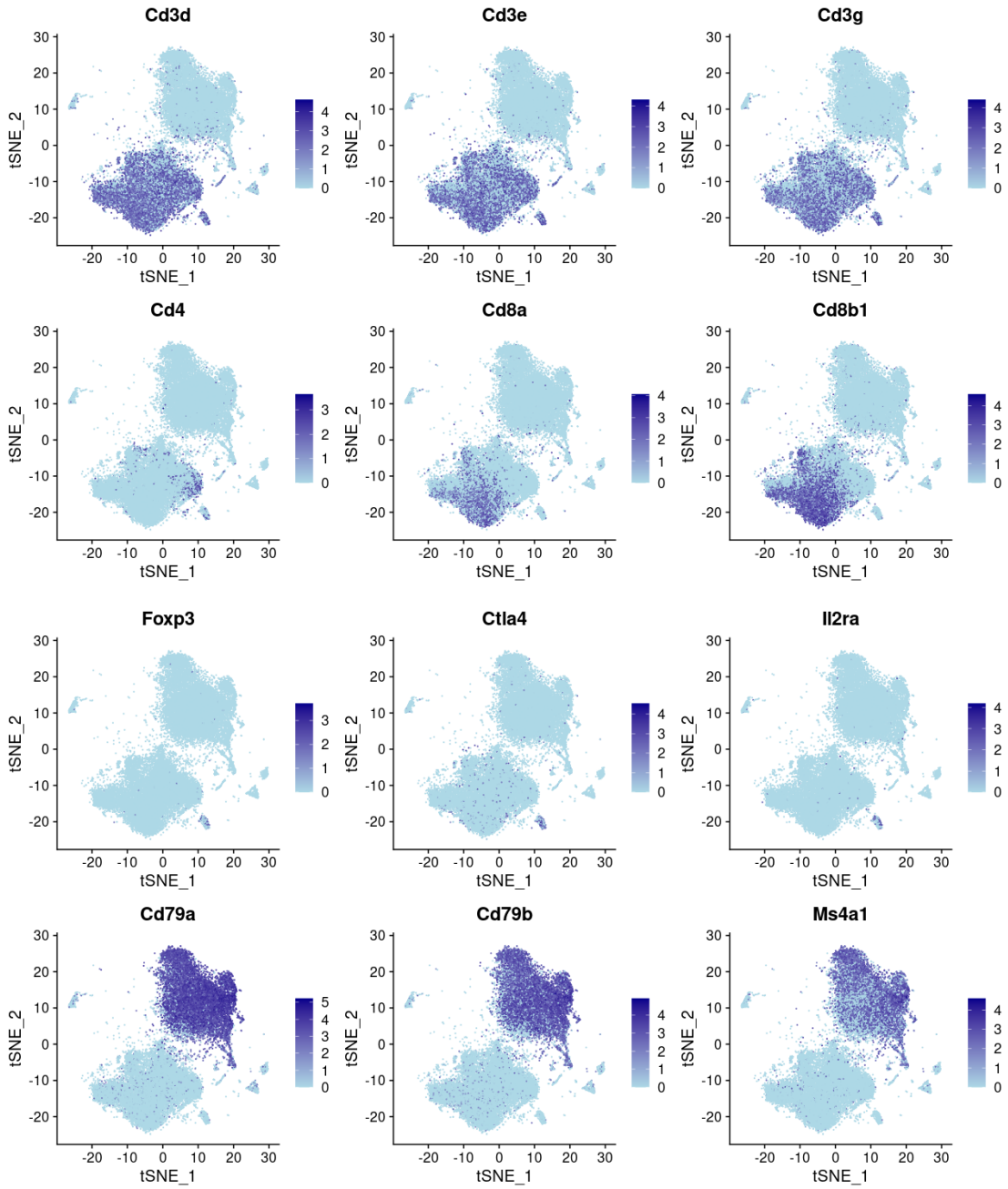


Figure C.10: Visualization of key marker genes in the SP scRNA-seq data.

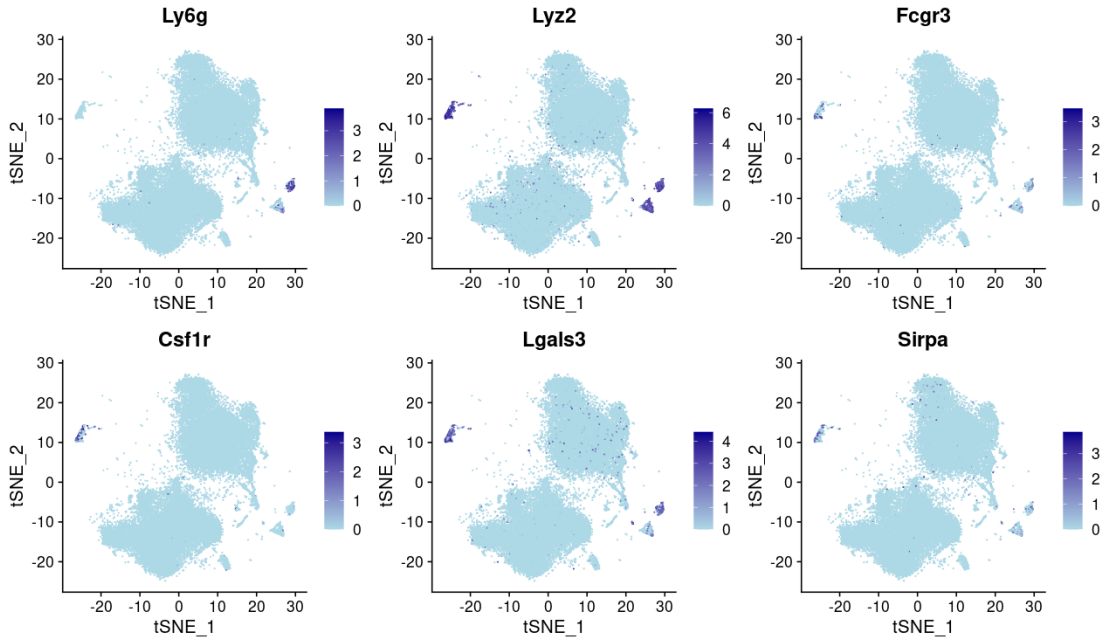


Figure C.11: Continued visualization of key marker genes in the SP scRNA-seq data.

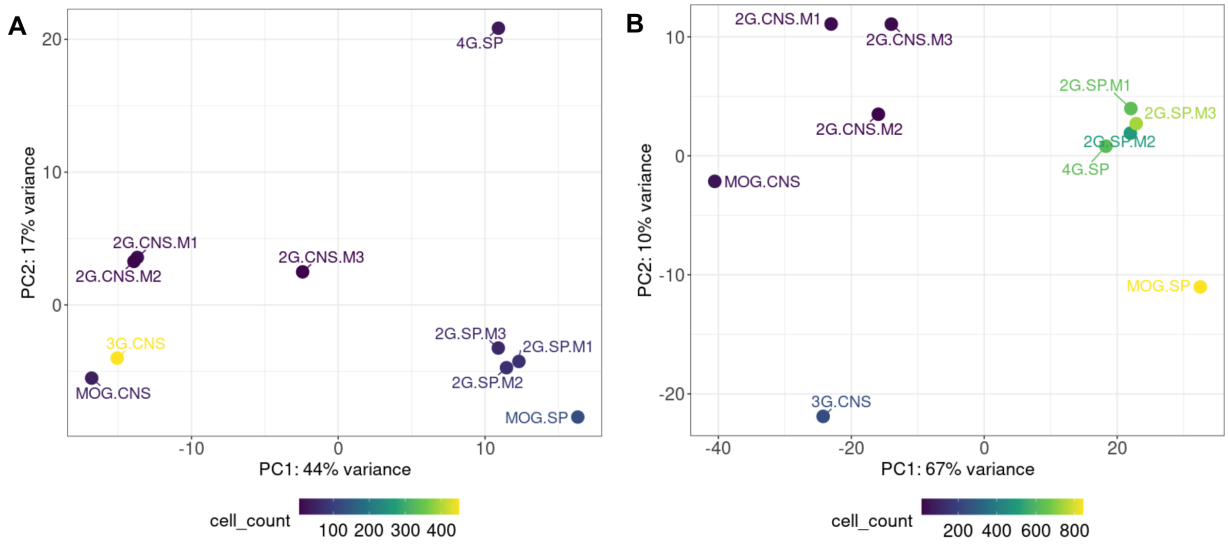


Figure C.12: Top two PCs of the samples in Tregs (A) and T effector cells (B), colored by the number of single-cell level UMI counts per sample.

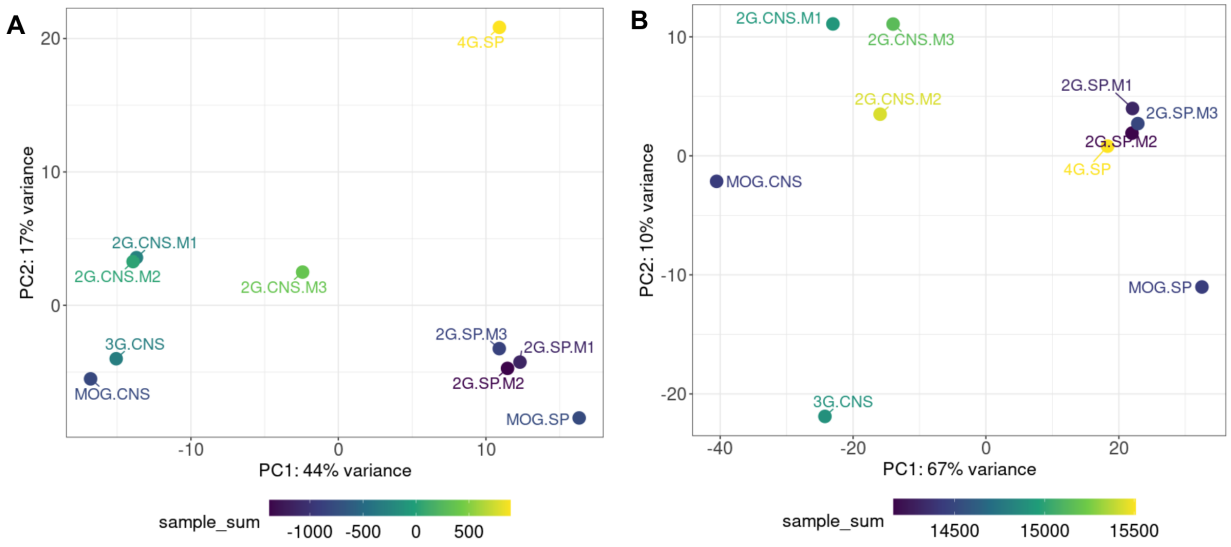


Figure C.13: Top two PCs of the samples in Tregs (A) and T effector cells (B), colored by the total sum of normalized data per sample.

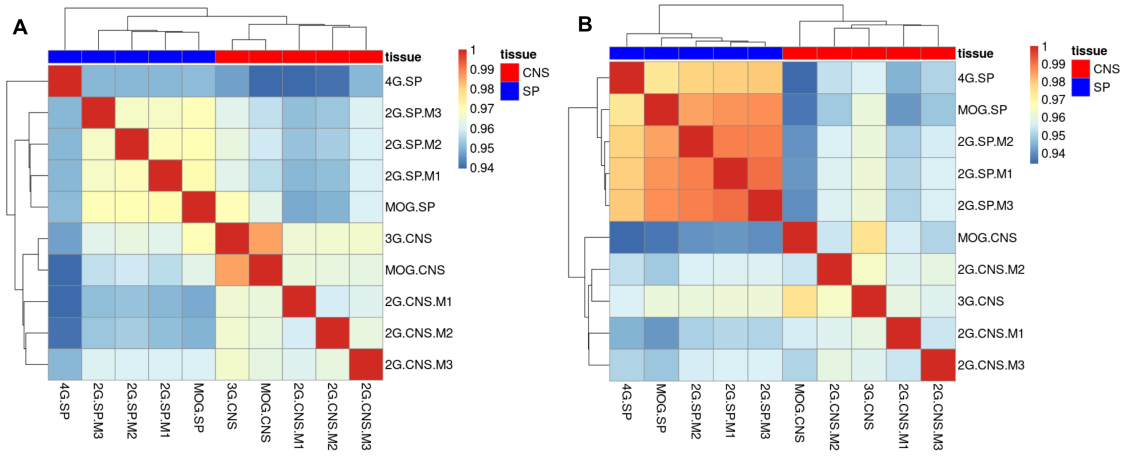


Figure C.14: Hierarchical clustering of the samples in Tregs (A) and T effector cells (B).

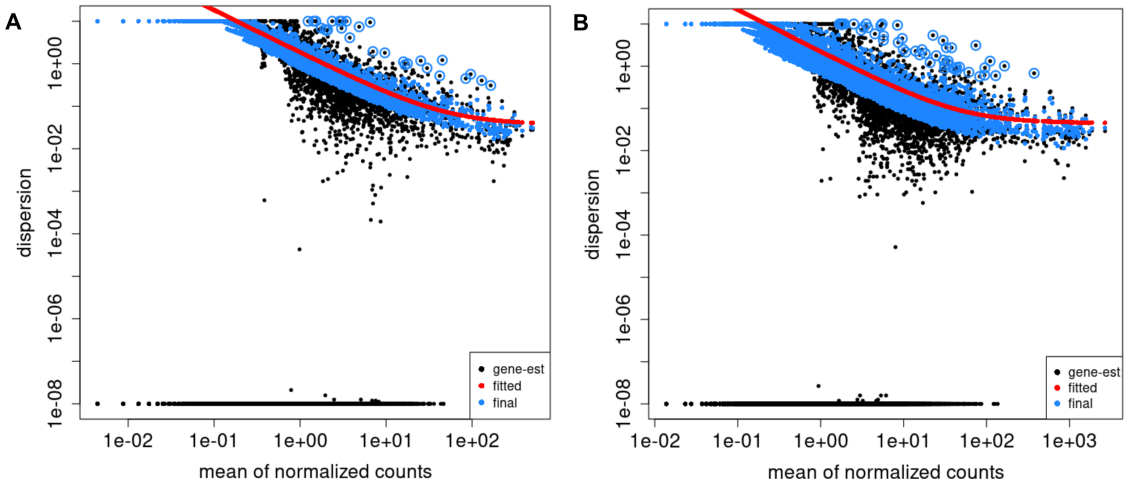


Figure C.15: Dispersion estimates of the samples in Tregs (A) and T effector cells (B).

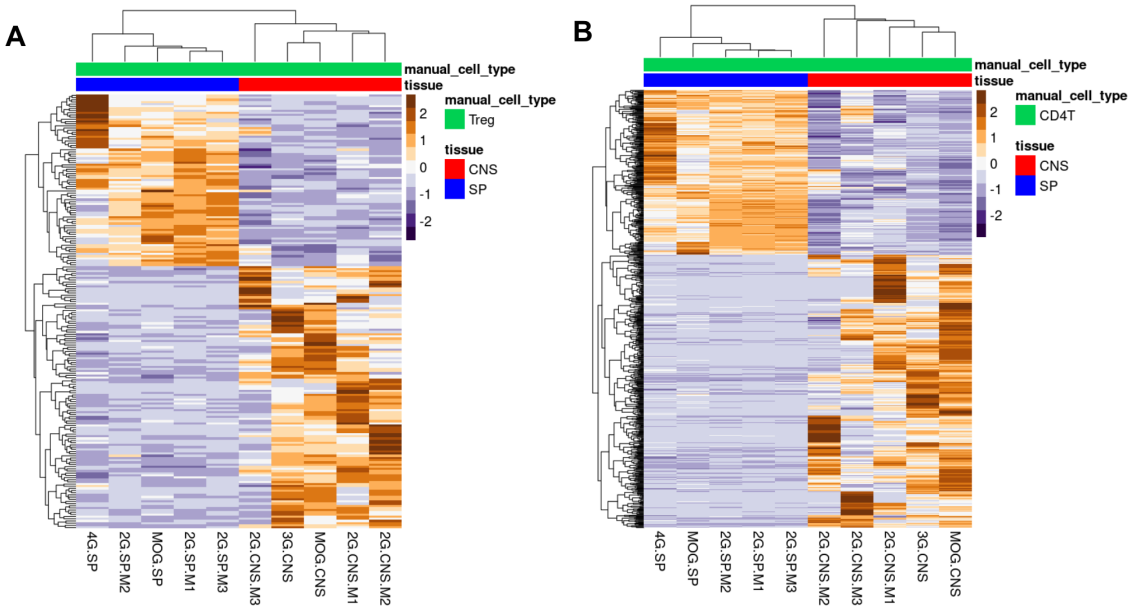


Figure C.16: Hierarchical clustering of the aggregated gene expressions in Treg (A) and T effector cell (B) samples. Each row corresponds to a gene that is statistically significantly differentially expressed between CNS and SP. Each column corresponds to a sample.

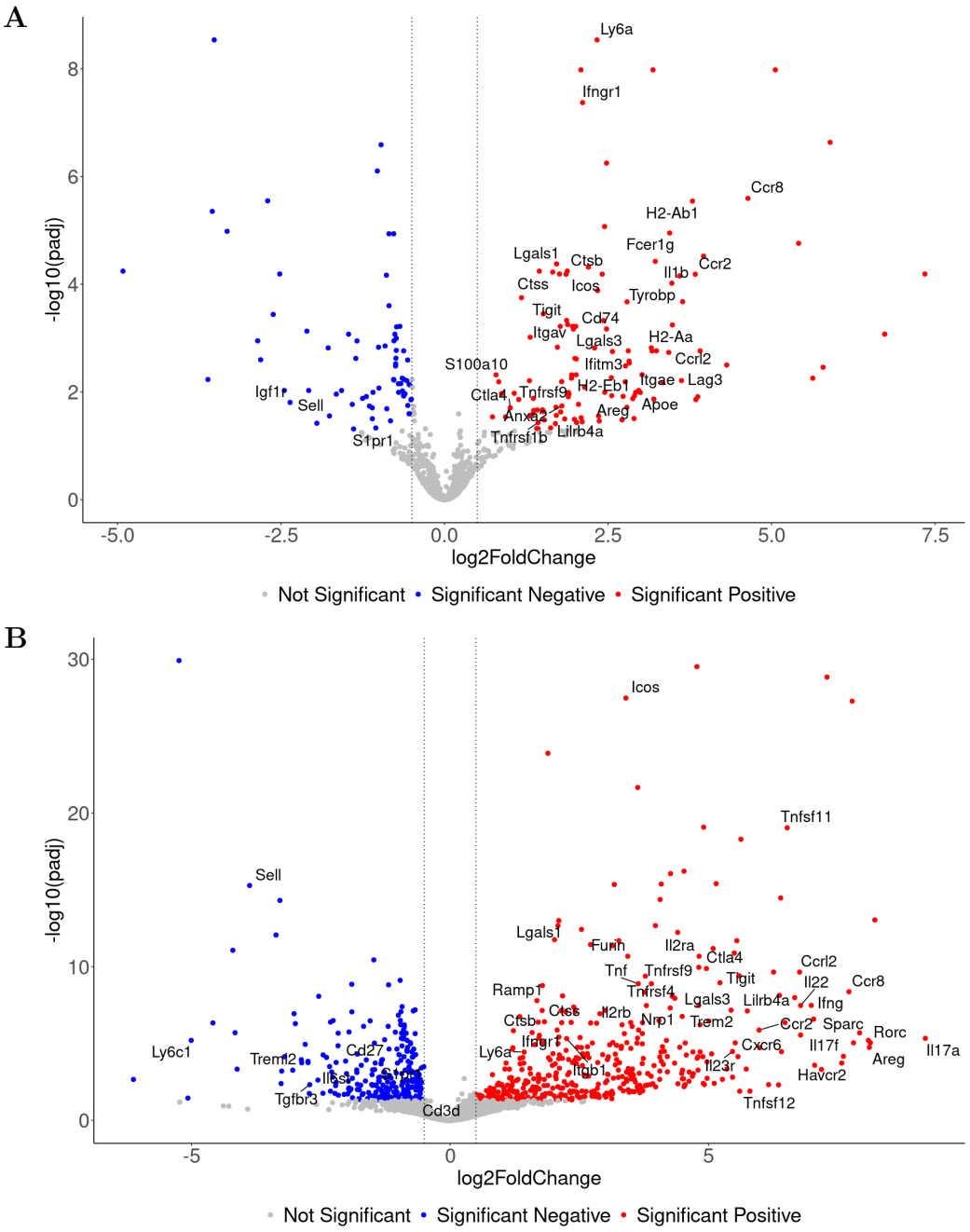


Figure C.17: Genes related to cytokines, cytokine receptors, and cell surface GO terms are annotated and highlighted among all the differentially expressed genes between CNS and SP, among Tregs (A) and T effector cells (B). Genes in red are more highly expressed in the CNS than SP. Genes in blue are more highly expressed in the SP than CNS. Genes in gray are not statistically significantly differentially expressed between the two tissue conditions.

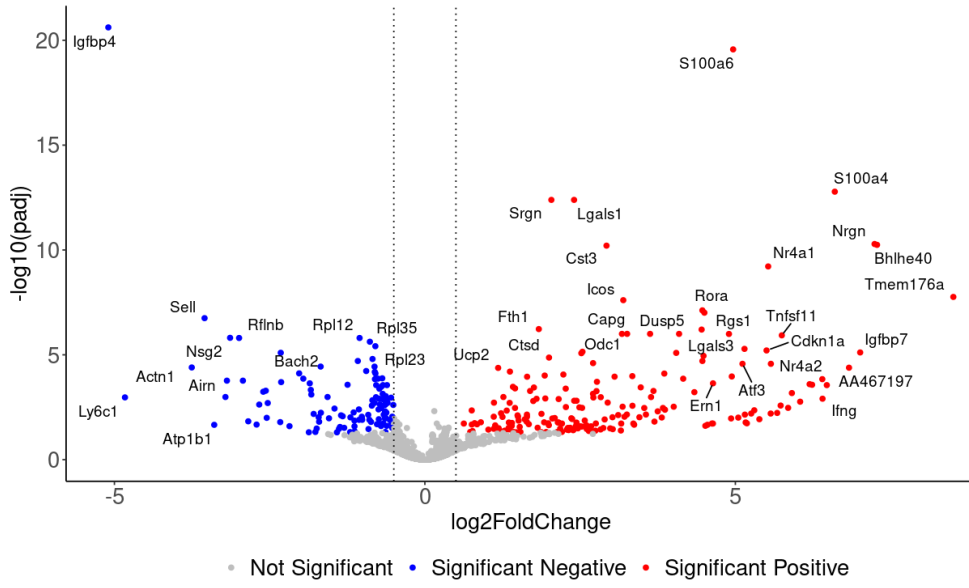


Figure C.18: Volcano plots highlighting the differentially expressed genes between CNS and SP, among a random subsample of T effector cells. Genes in red are more highly expressed in the CNS than SP. Genes in blue are more highly expressed in the SP than CNS. Genes in gray are not statistically significantly differentially expressed between the two tissue conditions.

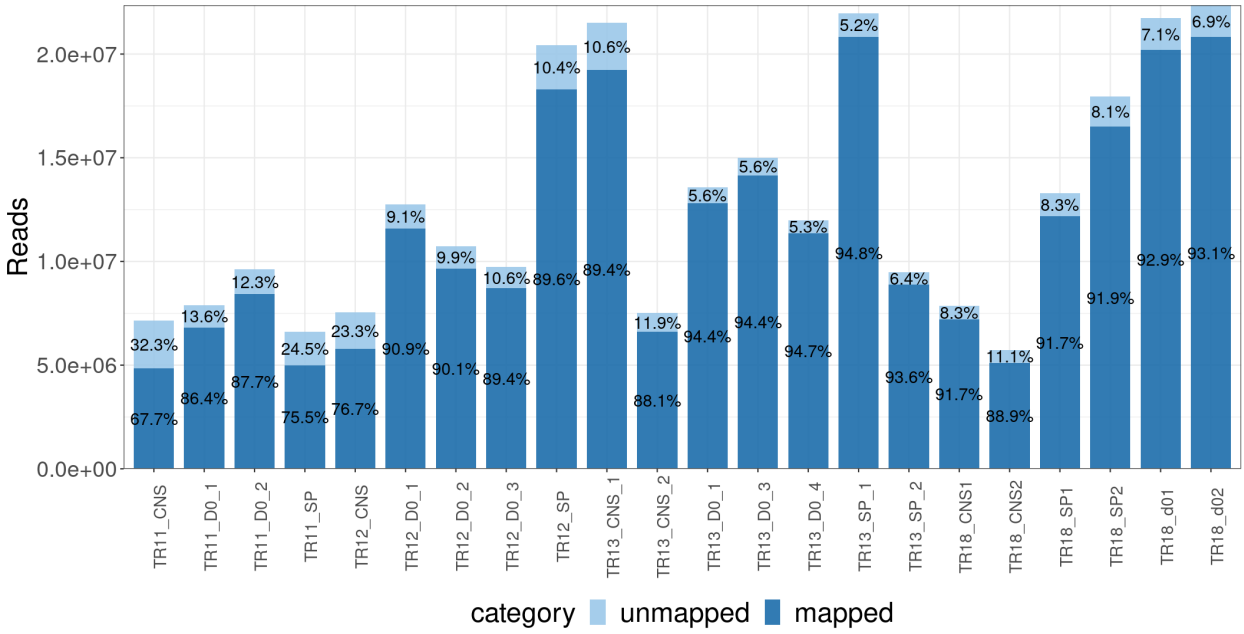


Figure C.19: The number of total reads, mapped and unmapped reads per CRISPR experiment.

C.2 Supplementary Table

Ptprc	Cd3g	Cd3e	Cd4	Cd8a	Cd8b1	Trac	Trbc2
Tcrg-C1	Tcrg-C2	Cd19	Ms4a1	Cd79a	Cd79b	Sdc1	Aicda
Itgam	Itgax	Itgae	Cd68	Csf1r	C1qa	Cx3cr1	Adgre1
Siglec1	Ly6c1	Ly6c2	Ccr2	H2-Ab1	H2-Eb1	Cd14	Fcgr3
Ms4a7	Thbd	Lyz2	Mgl2	Gzma	Gzmb	Ncr1	Nkg7
Nrp1	Cd33	Fut4	Foxp3	Ctla4	Pdcd1	Btla	Lag3
Havcr1	Havcr2	Adora2a	Il2ra	Cd69	Icos	Tnfrsf4	Tnfrsf9
Cd27	Cd28	Sell	Ccr7	H2-M2	Apol7c	Siglech	Irf8
Irf4	Cd207	Epcam	Cd80	Cd86	Cd40	Cd40lg	Mertk
Ly75	Cd24a	Ccr6	Il7r	Cxcr6	Il22	Sirpa	Xcr1
Clec9a	Batf3	Zbtb46	Tmem173	Ifih1	Isg15	Gbp2	Cxcl10
Mki67	Col1a2	Lyve1	Prox1	Vwf	Pecam1	Aire	Mcpt4
Fcer1a	Kit	Stfa2l1	C4b	Dcn	Emcn	Ecscr	Hbb-bs
Krt7	Krt18	Lyz1	Marco	Cd63	Folr2	Ly6g	Mmp9
Ccdc184	Klrb1b	Fam184b	Klrb1	Zbtb16	Trbv1	Trbc1	Cd1d

Table C.1: The list of 120 marker genes for cell type annotation in the CNS and SP.