UNIVERSITY OF CALIFORNIA SAN DIEGO

# Mapping human tissues with spatial transcriptomics

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bionengineering

by

Kian Kalhor

Committee in charge:

Professor Kun Zhang, Chair
Professor Prashant Mali, Co-Chair
Professor Francisco Contijoch
Professor Kevin King
Professor Bing Ren

2024

The Dissertation of Kian Kalhor is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

# DEDICATION

This text is dedicated to the fighting people in Iran.

For *Woman, Life, Freedom*

# EPIGRAPH

If you wanna make God laugh, tell Him about your plans.

*Woody Allen*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# LIST OF ABBREVIATIONS

DNA: Deoxyribonucleic acid
RNA: Ribonucleic acid
mRNA: Messenger RNA
FISH: Fluorescent in situ hybridization
DART-FISH: Decoding amplified targeted transcripts with fluorescence in situ hy- bridization
FOV: Field of view
RCA: Rolling circle amplification
Rolony: RCA colony rolony: RCA colony
NA: Numerical aperture
BF: Brightfield FISH: Fluorescence in situ hybridization
scRNA-seq: Single-cell RNA sequencing
snRNA-seq: Single-nucleus RNA sequencing
PA: Polyarcylamide
SpD: Sparse deconvolution
nt: Nucleotide
kb: kilobase
M1C: Primary motor cortex
UMI: Unique molecular identifier
AEA: Afferent efferent arterioles
VSMC: Vascular smooth muscle cells
TAL: Thick ascending limb
PC: Principal cells
FIB: Fibroblasts

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Dr. Kun Zhang for accepting me to his lab and kindly supporting me over the years. I am thankful for his patience, wisdom and mentorship. He created a friendly lab environment where I felt safe and welcomed everyday and was encouraged to try new things and fail. I am immensely grateful to Dr. Prashant Mali, my co-advisor, who helped me enter the Bioengineering program, and provided mentorship on my science and career all along this journey. I also want to thank my committee members, Drs. Bing Ren, Francisco Contijoch and Kevin King for giving me valuable advice on scientific rigor and effective communication of my work.

During my first rotation in the Mali lab, I met some of the most amazing people I know. I owe Dr. Amir Dailamy a debt of gratitude for being like a brother in San Diego, for teaching me about fashion and rap, for taking me to many restaurants and cafes, drinking many tequilas, and hanging out on many weekends at his place with his dog Chook-chook and our friend Robert Gutierrez. I want to thank Dr. Dhruva Katrekar for sharing many random joyful experiences around San Diego, from hiking five peaks in one day to joining a spin class on its last session. I also would like to thank Dr. Udit Parekh for being unbelievably knowledgeable and generous, and being a role model for me, from my first day at UCSD until now.

I started this journey with almost no wet lab skills, and I am grateful to all my lab mates for their openness and helpfulness. I would like to thank Dr. Matt Cai for creating the foundations on which I based my projects, Dr. Richard Que for patiently teaching me the basics, Dr. Noi Plongthongkum for teaching me the importance of checking for reverse complements, and Dr. Huy Lam for resourcefully reviewing all my experimental results and helping me troubleshoot my failed experiments, including my unsuccessful attempts at surfing.

I was fortunate to have wonderful collaborators for my human kidney projects. I was always amazed by Dr. Sanjay Jain's passion and drive in leading large efforts to profile the human kidney while being very accessible and helpful with my projects. I would like to thank Amanda Knoten for cutting many kidney sections for me, and my lab mate, Dr. Blue Lake for

being an effortless collaborator throughout.

My lab mates were vital to my well-being and day-to-day happiness and learning. I want to thank Kimberly Conklin for her genuine kindness and our early morning chats, Sarah Criner for ordering my first oligo, Dr. Daniel Jacobsen for teaching me the value of flattening the boxes, Dr. Jinghui Song for being so humble and astute, Joe Yuan for keeping me updated with the life of the new generation, Tony Cheng for our weekend ping pong training sessions, Sammi Lyu for introducing me to one of the best Chinese restaurants in town, Emilie Aghajani for sharing bits of her interesting life, Dr. Dinh Diep for coordinating lab BBQ potlucks, and Dr. Yan Wu for being the big brother in the lab. I would like to express my gratitude to Dr. Xuwen Li, Dr. Michael Wang and Dr. Harris Jeong for many heated games of ping pong and many hours of enjoyable conversations.

I would like to acknowledge my friends Mahsa Nafisi, Dr. Van Ninh, their partners Branden Cobb and Sid Swaim, and Dr. Chien-ju Chen. I cherish our collaborations on spatial transcriptomics, our many conversations about science and politics, and the many laughs, karaoke sessions, cocktails and parties we shared. I would like to extend my heartfelt gratitude to Dr. Blair Jia for his rejuvenating passion and curiosity for science, food and life and our many hours-long hikes and conversations on campus, Dr. Patrik Kasl for being a good friend whose dedication always surprises me, Sylvia Liang for all the trails we hiked, and Dr. Ayoob Shahmoradi for giving me bits of Philosophy and sharing my passion for Iranian music.

I finished the last year of my PhD at Altos Labs in San Diego. I was fortunate to befriend many brilliant scientists, including Drs. Valentine Svensson, Aditya Kumar, Ivan Rosshandler, Sergei Manakov, Marzia Savini, Farshad Babaeijandaghi, Xiaotong Hong, Mariana Yusupova, Francesco Della Valle and many others. I would like to thank them all for the interesting conversations and their support.

I would like to thank my college friends, Mohammad Tinati, Dr. Alireza Modirshanechi, and Mohammadreza Divsalar. Even though we are dispersed around the globe, we kept in touch and enjoyed many decompressing video calls. As for decompression, I would like to thank

Artpower for providing world-class yet affordable concerts on campus, Jordan Mojica for being such a good trainer that he made me look forward to going to the gym, and my aunt Khale Forouz and her children, their partners, and grandchildren for giving me love and support from the moment I entered the U.S. until now.

Last but not least, I feel most blessed for having the family that I have. They believe in me, and give me unconditional love. I was lucky to marry my oldest and best friend, Setayesh Radkani, in spite of Life's efforts to keep us apart. My brothers, Farid and Reza, have always been the guiding lights in my life. My sister-in-law, Alexis, is like a true sister to me. I discovered a new kind of love in my heart when my niece and nephew, Kimiya and Cyrus were born—the love of an uncle. And finally, to my parents, Farah and Morteza, I am grateful for their principles that I aspire to and their unwavering support.

Chapter 2 is, in part, a reprint of the material as it appears in Lake, B. B., Menon, R., Winfree, S., Hu, Q., Melo Ferreira, R., Kalhor, K., ... & Jain, S. (2023). An atlas of healthy and injured cell states and niches in the human kidney. *Nature*. The dissertation author was the primary investigator and co-first author of this paper.

Chapters 3 and 4, 5 are, in part, reprints of the material as it appears in Kalhor, K., Chen, C. J. ... & Zhang, K. (2024). Mapping human tissues with highly multiplexed RNA in situ hybridization. *Nature Communications*. The dissertation author was the primary investigator and co-first author of this paper.

## VITA

| | |
|---|---|
| 2018 | Bachelor of Science in Electrical Engineering,<br>Sharif University of Technology |
| 2024 | Doctor of Philosophy in Bioengineering,<br>University of California San Diego |

## PUBLICATIONS

**Kalhor, K.***, Chen, C. J.*, Lee, H. S., Cai, M., Nafisi, M., Que, R., ... & Zhang, K. (2024). Mapping human tissues with highly multiplexed RNA in situ hybridization. *Nature Communications, 15*(1), 2511.

Lake, B. B.*, Menon, R.*, Winfree, S.*, Hu, Q.*, Melo Ferreira, R.*, **Kalhor, K.***, Barwinska, D., ... & Jain, S. (2023). An atlas of healthy and injured cell states and niches in the human kidney. *Nature, 619*(7970), 585-594.

Leeper, K., **Kalhor, K.**, Vernet, A., Graveline, A., Church, G. M., Mali, P., & Kalhor, R. (2021). Lineage barcoding in mice with homing CRISPR. *Nature protocols, 16*(4), 2088-2108.

**Kalhor, K.**, & Church, G. M. (2019). Single-cell CRISPR-based lineage tracing in mice. *Biochemistry, 58*(48), 4775-4776.

Kalhor, R., **Kalhor, K.**, Mejia, L., Leeper, K., Graveline, A., Mali, P., & Church, G. M. (2018). Developmental barcoding of whole mouse via homing CRISPR. *Science, 361*(6405), eaat9804.

(* equal contribution)

ABSTRACT OF THE DISSERTATION

Mapping human tissues with spatial transcriptomics

by

Kian Kalhor

Doctor of Philosophy in Bionengineering

University of California San Diego, 2024

Professor Kun Zhang, Chair
Professor Prashant Mali, Co-Chair

The structure and function of tissues are highly intertwined, necessitating an understanding of their architecture to comprehend their normal and pathological states. While traditional histology and immunostaining offer high spatial resolution, they are limited in molecular resolution. Conversely, single-cell sequencing provides molecular resolution but lacks spatial context due to tissue dissociation. Spatial transcriptomics bridges this gap by combining microscopy and DNA technology, enabling high resolution molecular readouts with spatial information.

In the first part of this dissertation, we utilize a spatial RNA capture technology to map cell types in the human kidney at a near-cellular resolution, revealing cellular neighborhoods

across different anatomical regions of the kidney, identifying niches for novel populations, and uncovering strong mitochondrial gene expression patterns in a specific epithelial subtype.

While sequencing-based methods have a fixed spatial resolution, multiplexed in situ detection methods offer cellular resolution but face challenges in sensitivity, throughput, and cost which severely limit their application in human sections that are large and have low quality. In aim 2, I developed a novel multiplexed in situ RNA detection method, DART-FISH, capable of profiling hundreds of genes in large human tissue sections with ease of implementation. I further developed an accompanying suite of computational tools for designing and processing the output of in situ experiments.

Finally, in aim 3, the utility of DART-FISH is demonstrated by applying it to multiple tissue types. This includes the human brain where the layered organization of excitatory neurons was recapitulated and a rare subclass of inhibitory neurons uncovered. In the human kidney, I profiled healthy and pathological cell states in the cortex, and identified interactions between disease-altered epithelial cells and myofibroblasts. Finally, I showed how DART-FISH can be utilized for organ-scale measurements by imaging a cross section of a mouse kidney where all segments of the nephron can be visualized and pathological neighborhoods can be systematically analyzed with single-cell resolution.

In summary, this dissertation provides several experimental and computational solutions to advance the field of spatial transcriptomics and lower its cost. It also shows the application of spatial transcriptomics to map the architecture of human and mouse tissues.

# Chapter 1

# Introduction and background

## 1.1 Prelude

Understanding how our bodies work has been a central quest throughout the history. This is a problem that can be approached at different length scales, from the chemistry of different biomolecules and complexes to gross anatomical studies of different organs. Since the discovery of cells in the 17th century, cell research has been a major avenue in our quest. Studying how cells function and how their behavior affects the broader organism can lead not only to an understanding of our biology, but also can help us control the biological phenomena around us, as in medicine and bioengineering.

## 1.2 Visualizing cells with microscopes

Microscopy has been the main means for studying cells since the 17th century [1]. Through the lens of the light microscope, we can see how cells look like and how they change under different conditions. Combined with various stains that embellish particular cells or subcellular structures, we can measure cell properties and diagnose diseases. For example, by looking at slides stained with hematoxylin and eosin (H&E), pathologists can distinguish between a benign and a malignant tumor, or identify areas inflammation and fibrosis.

The development of fluorescent microscopy that started in early 20th century [2] along side with advances in sample preparation, genetics and molecular conjugation [3] significantly

expanded the scope and resolution of what could be captured by imaging and allowed us to visualize multiple analytes, be it proteins or nucleic acids, at a high resolution in tissues and even in live cells.

In spite of all the advances in the instrumentation, reagents and sample, there is a fundamental constraint as to how many different analytes one can image at a unit time with a light microscope. In order for fluorophores to be distinguishable, they need to emit light in different frequencies and the number of fluorophores with distinct emission spectra is limited to a handful. However, the number of molecules (proteins, RNAs, etc.) for a comprehensive analysis of cells in a tissue can easily exceed this limit. Hence, increasing the number of different molecules that can be imaged in the same experiment will be very valuable.

## 1.3 High throughput genomics

Genomics had already emerged as a field by the 1990s. At its heart, genomics owes its versatility to a "simple" rule that governs nucleic acids: predictable base pairing between molecules. Moreover, nature has produced a variety of tools to recreate DNA and RNA from an existing template (e.g., polymerases), and to cut (e.g., nucleases) and glue (e.g., ligases) pieces of nucleic acids in different ways. Advances in DNA sequencing technologies also contributed to the quick and widespread adoption of genomics. With all these tools at their disposal, scientists could measure different cell properties by converting them DNA libraries and sequencing them at high throughput.

By early 2010s, scientists could sequence whole genomes [4], exomes[5] and transcriptomes [6] from tissues, as well as various flavors of DNA modifications, and protein-DNA[7] and chromatin interactions [8] at bulk. These technologies have provided valuable insights into, for instance, the effect of DNA variants [9] and gene regulation [10] the first step in these workflows is to grind the samples and isolate nucleic acids the single-cell and cell type heterogeneity is lost which limits our understanding of individual cells' behaviors.

In the past decade, with many innovations in sample preparation, including droplet microfluidics [11] and combinatorial indexing [12], many of these sequencing methods have been extended to work with inputs as low as single cells. With high-throughput measurements being made at single-cell resolution, scientists have profiled human and animal tissues and discovered new cell types and cell states. High resolution molecular atlases shed light onto cell differentiation in development [13] and disease progression [14]. The missing dimension from all these technologies, however, is the architecture of the tissue architecture. Cells do not function in vacuum, rather, they work together and communicate to perform their homeostatic function, undergo repair or create pathology.

## 1.4 Genomics going to space

Spatial Omics is the cross between the two fields above: the spatially-enriched microscopy and the molecularly-enriched genomics[15]. It inherits spatial information from microscopy and multiplexing capacity from DNA technology so that many different analytes can be measured at the same time while preserving their spatial location. In my thesis, I explore different aspects of this marriage with a focus on transcriptomics. I describe the development of a new RNA detection technique, and show the utility of spatial transcriptomic technologies for studying human and mouse tissues.

The goal of spatial transcriptomics is to measure the expression of many genes in tissues while preserving information about their original transcripts. By measuring the gene expression in all cells, one could identify the different cell types and states present in our tissue samples. In addition to that, having the spatial information allows one to reconstruct the tissue environment and study the neighborhoods in which cells function. By comparing samples from different conditions such as disease or ageing, one could study how the cell states change and how cellular neighborhoods are affected by the condition. Thus, spatial transcriptomics can be a useful tool for studying the mechanism of disease by elucidating the cell states cell-cell interactions a at a

tissue scale.

In the past decade, there have been two major approaches for obtaining high-throughput spatial RNA information: 1) Direct identification of mRNA molecules in situ through multiplexed imaging, 2) spatial barcoding of RNA followed by extraction and sequencing. In the following, I briefly explain the first publications that laid the foundations of this thesis.

### 1.4.1 Multiplexed in situ RNA detection

In a typical fluorescent staining experiment, 3 or 4 molecules (e.g., mRNA) can be detected, each of which on a separate fluorescent channel. However, in a section of a complex tissue, as in the kidney, there maybe several distinct cell types present. Additionally, cells can be in different states depending on the experimental conditions and their interaction partners. Resolving this complexity is beyond what can be learned from a handful of molecules.

Fortunately, the reversible base-pairing of DNA allows us to introduce multiplexing through combinatorial barcoding. In other words, if one round of imaging can uniquely label 3 molecules, two rounds of imaging can label $3 * 3 = 9$ molecules, and $N$ rounds can label $3^N$. With this combinatorial growth, measuring the expression of hundreds and thousands of genes will be attainable with $N = 10$ yielding a higher labeling capacity than the number of human genes. This idea has been the core to most of the multiplexed in situ detection methods, but has been implemented in a variety of different ways.

The first demonstration of multiplexed in situ RNA detection was by Ke et al. [16] and Lee et al. [17]. Both methods shared some key characteristics: 1) fixed tissues underwent in situ reverse-transcription and cDNA anchoring, 2) in-tissue rolling circle amplification (RCA[18]) was used to create a cluster of DNA molecules 3) the identity of the DNA clusters (rolonies) was decoded by in situ sequencing. Ke et al. took a targeted approach and used barcoded padlock probes against the cDNA of selected genes (see section 3.1). In contrast, Lee et al. directly performed in situ sequencing on the amplified cDNA molecules, hence a whole-transcriptomic measurement.

Because of the numerous enzymatic reactions, both of these methods tend to have low sensitivity which severely reduces their utility on samples with low quality (e.g., human post-mortem tissues). Furthermore, the readout by in situ sequencing is both cumbersome and needs specialized equipment. However, the use of RCA for signal amplification obviates the need for complex optical systems and allows for the use of lower magnification objective lenses and hence, higher throughput.

There is another class of in situ methods that is not a focus of this thesis. These methods, pioneered by Chen et al. [19] and Lubeck et al. [20], do not use enzymatic signal amplification and are instead based on single-molecule FISH [21]. Briefly, mRNA molecules are tiled with short oligonucleotide probes that are either conjugated to fluorescent molecules or facilitate the binding of a secondary probe carrying a fluorophore. These workflows enjoy a high sensitivity in terms of the number of labeled mRNA molecules, because hybridization is an efficient process. However, they require mRNA molecules to be long to generate a detectable signal. Moreover, because of their low signal-to-noise ratio (SNR) they need very sensitive optical systems and long imaging times.

Chapters 3 describes my efforts in advancing the padlock probe-based methods by creating a new technology that is simple to implement, and robust to the low quality human samples. Chapter 4 summarizes my work on various computational challenges that arise in designing and processing

### 1.4.2 Spatial RNA capture and sequencing

The spatial barcoding approach was first demonstrated by [22] where the authors deposited barcoded poly-dT reverse-transcription primers onto a glass slide. A tissue section was then cut on this slide and the mRNA was captured by the primers followed by library preparation. Since the position of the barcoded primers was known a priori, sequencing the cDNA library revealed both the gene identity and the position of the cDNA molecules.

Even though an elegant strategy, the utility of this method was compromised by the

large size of the capture spots (100um) and the large spacing between them (200um). It was not until 2019 that two groups significantly increased the resolution of these assays. Rodriques et al. [23] and Vickovic et al. [24] randomly deposited small barcoded beads (10um and 2um respectively) onto a surface. Next, they performed in situ sequencing or hybridization to read out the barcode on each bead. With the location of the barcodes identified, they could spatially barcode the mRNA in tissue sections and map them back to their bead of origin. Later, slide-seq2 was introduced by Stickels et al. [25] in which the sensitivity was significantly increased and reached about 50% single-cell sequencing technologies.

Sequencing-based methods can capture transcriptome-wide information. They can even be extended to capture non-coding transcipts [26]. However, synthesizing high quality barcoded slides is very challenging for normal labs and would best used as commercial products.

Chapter 2 concerns the application of a sequencing-based spatial transcriptomics method (Slide-seq2) on human kidney sections. We leveraged the high sensitivity and fine spatial resolution of this assay to create maps of the kidney and study the cell type composition and cell-neighborhoods in different regions of healthy individuals.

## 1.5 Organization of this thesis

The organization of the later chapters of the this work are as follows. Chapter 2 describes a multimodal kidney atlas. In this work, we profiled kidney samples from multiple individuals using different technologies. The emphasis of the chapter is on the spatial mapping of the cell types and neighborhoods in healthy human samples using capture and sequencing-based technologies.

Chapter 3 describes a new in situ hybridization technology that I co-developed. This method, called DART-FISH, is capable of measuring the expression of hundreds of genes in their native tissue. DART-FISH has a built-in cytoplasmic stain that can be used for cell segmentation and provides more context for the tissue structure than the commonly used nuclear stains. I show

that the sensitivity of DART-FISH is high enough to provide new insights into the organization and interation of cells in human tissues.

Chapter 4 describes the pipeline that I developed for processing the output of multiplexed in situ hybridization experiments. Every DART-FISH experiment generates hundreds of thousands of images that require an efficient and automatic pipeline to process. This pipeline is modular and every module can be replaced by new ones that can better accommodate the needs of an experiment. As part of this pipeline, I developed a new algorithm for processing the DART-FISH images that can extract gene expression information even in presence of optical crowding. I also describe some efforts on barcode design, a key aspect yet an under-explored one in multiplexed in situ hybridization experiments.

Chapter 5 goes over the data generated by DART-FISH in human brain, human kidney and mouse kidney. The layer organization of excitatory neurons is resolved in the cortex of the brain. The neighborhoods of healthy and diseased cells in the human kidney are explored. And finally, in a full cross-section of an aged mouse kidney, all the major cells are uncovered. The data in this chapter demonstrates the application of the methods developed in this thesis on relevant tissue samples.

# Chapter 2

# Creating a spatially resolved map of the healthy human kidney

## 2.1 Introduction

In collaboration with multiple teams of scientists and clinicians, our lab set out to map the cells in the human kidney at the molecular level and reveal the different cell types and states present in the healthy kidney, as well as in patients with acute kidney injury (AKI) and chronic kidney disease (CKD) [27]. In summary, samples from tens of patients were processed with complementary sequencing technologies such as single-cell and single-nucleus RNA sequencing, as well as SNARE-seq [28, 29], a technology for joint profiling of mRNA and chromatin accessibility from single cells.

Samples were collected to cover the entire depth of the human kidney, from cortex to the papillary tip. Combined, we obtained more than 400,000 high quality profiles of single cells or nuclei. By clustering and careful inspection, these cells were in a hierarchical manner. Subclass level 1 includes the main epithelial, endothelial, stromal and immune cell types (10-15). In subclass level 2 and level 3, more than 70 subclasses were annotated that reflected the cell states as well as the regional variance of the cell types. About 50 of these populations are canonical to the human kidney, while more than 20 are associated with kidney disease and repair. Therefore, through profiling of mRNA and chromatin accessibility we created an atlas that defines the various cell types and states in healthy and diseased human kidney.

The single-cell atlas does not inherently provide information about spatial organization of cells or spatial niches of cells of healthy or diseased cell states. To create a comprehensive atlas that addresses this issue, we complemented our existing atlas with sequencing-based spatial transcriptomics technologies and integrated the single-cell and spatial modalitites. These techniques enable us to characterize healthy and disease neighborhoods at a resolution determined by their capture area. They further help us study cell-cell interactions that are associated with disease. In our publication [27], we performed 10x Visium (55um capture area) and slide-seq [23, 25] (10um capture area) on samples from 22 and 6 patients, respectively. For the rest of this chapter, I focus on the analysis of the slide-seq data which I lead.

## 2.2 A primer on slide-seq

Slide-seq is a technology first introduced by Rodriques et al. [23] and further improved by Stickels et al. [25], capable of spatially barcoding mRNA molecules in fresh-frozen tissue sections. Tissues are sectioned onto a surface (a puck) covered with barcoded beads. The beads have a diameter of 10um, and are coated by oligonucleotides that have a 3' poly-deoxythymine (dT) capture sequence that will anneal to the poly-A tail of the mRNA. Additionally, the beads are synthesized such that a region of their oligos shares the same bead-specific sequence, or barcode. As part of the manufacturing of the pucks, thet undergo in situ sequencing to reveal the barcode and location for all beads. Then, once a tissue section is cut directly on the puck, the mRNA is released and captured by the poly-dT sequences on the bead, and reverse-transcribed. Then the barcoded cDNA is sequenced using standard techniques in such a way that every read contains a portion that carries the spatial barcode, and a portion with mRNA sequence. Consequently, the position of each mRNA can be mapped back to its original spatial location with a resolution of about 10um.

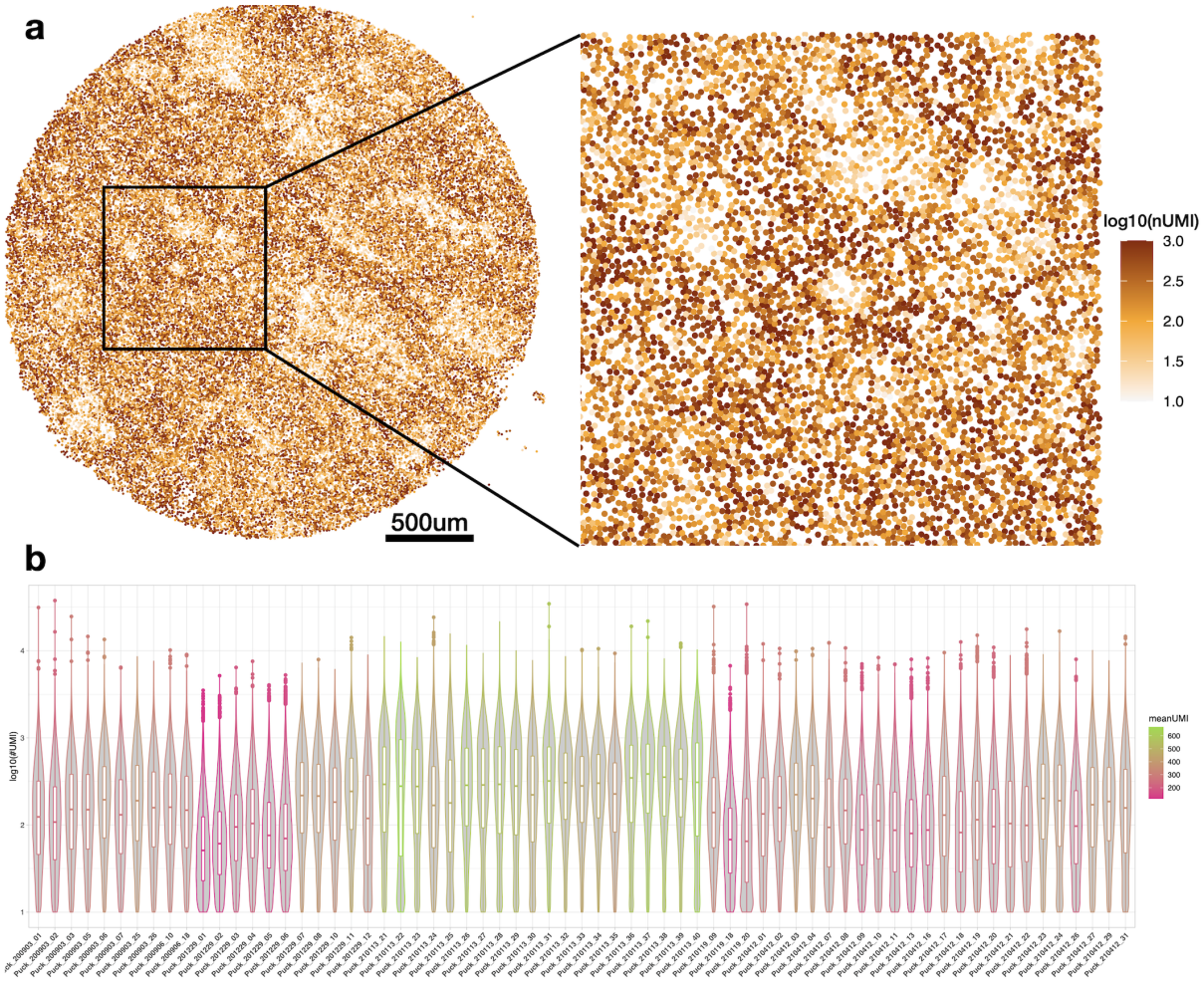



**Figure 2.1. a.** An example slide-seq puck. Colors show the number of captured UMI per bead on a logarithmic scale. The inset shows a zoomed-in area. **b.** Violin plots showing the distribution of UMI counts for all pucks in our slide-seq data set. The boxes indicate the quartiles. Puck_210113_21-40 are from medulla and have the highest UMI counts in the data set.

## 2.3 slide-seq on kidney

Our collaborators generated slide-seq data from 6 patients that encompassed the depth of the human kidney. In total we had 67 puck, of which 32 cover the cortex and 35 cover the medulla. Of the medullary pucks, 5 of them cover the inner medulla and the rest cover the outer medulla. Figure 2.1a shows an example of a puck, which a circular capture area that can fit it more than 60,000 beads.

Beads can capture up to a few thousand UMIs (figure 2.1b). On the lower end, the

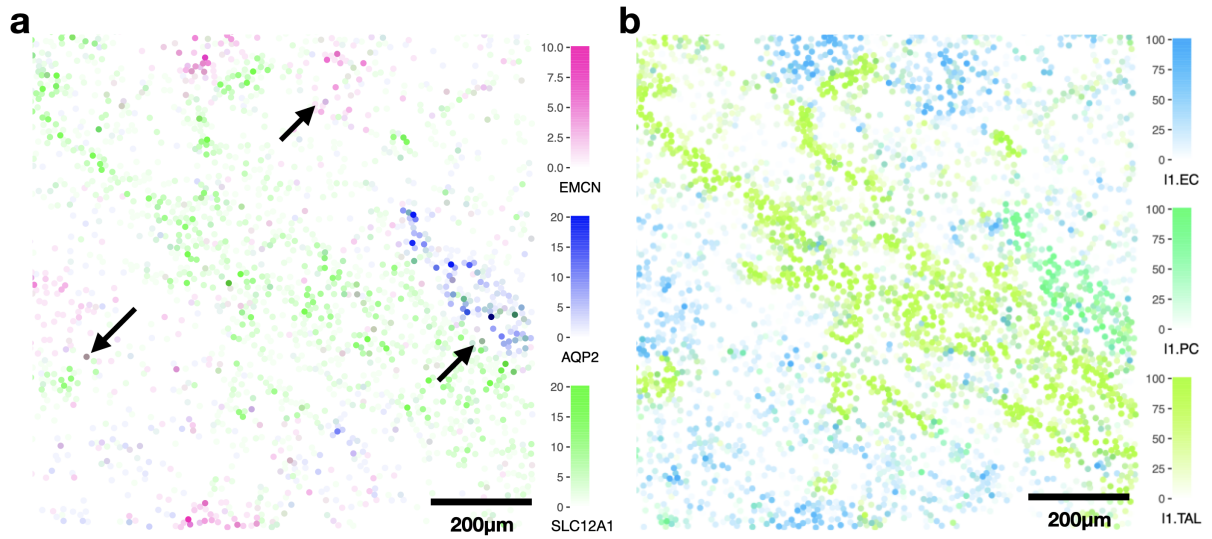



**Figure 2.2. a.** A zoomed-in view of a slide-seq puck rendered with expression of three marker genes. The lightness of colors is proportional to the count of the genes. Colors mix linearly upon gene mixtures. *EMCN* marks glomerular capillaries (EC-GC), *AQP2* marks principal cells (PC) and *SLC12A1* marks thick ascending limb cells (TAL). Arrows point to beads with mixed cell type signatures. **b.** RCTD deconvolution weights of three level one cell types (EC-GC, PC and TAL). The color mixing is same as in (a).

slide-seq processing pipeline discards beads with less than 10 UMIs. There are several factors affecting the distribution of UMI counts. Tissue type, quality and batch effects are the major global factors changing the mean UMI count per cell. For example, medullary pucks have considerably higher mean UMI count than the cortical pucks (figure 2.1b). Analyzing data sets that have lower overall UMI counts is more challening since many of the beads may not be correctly annotated. The other factor that increases UMI variance locally is tissue morphology. For example, kidney tissue is replete with tubules and lumens. Beads that border these areas will have lower UMI counts. Since slide-seq does not produce any output images that capture tissue morphology, accounting for this type of variance is not straightforward.

## 2.4 Cell type annotation of slide-seq data

The next step in analyzing a slide-seq data set is identifying the cell type and cell state that every bead has captured, or annotation. Since the size of slide-seq beads is 10um (comparable to

the size of a cell), and the beads are randomly positioned in the puck, each bead will likely cover more than one cell (Figure 2.2a). Hence, beads will likely show mixed signatures of different cell types if they cover two or more cells. As a result, naive scRNA-seq approaches that treat each slide-seq bead as a single-cell may fail to annotate cell types accurately. For example, de novo annotation of beads which involves clustering followed by manual annotation of the clusters will be challenging since the mixing fades the differences between clusters. Thus we need a tool that can account for this mixture. Additionally, our tool will ideally be able to perform label transfer, that is, automatically annotate the slide-seq beads using the annotations of a scRNA-seq reference.

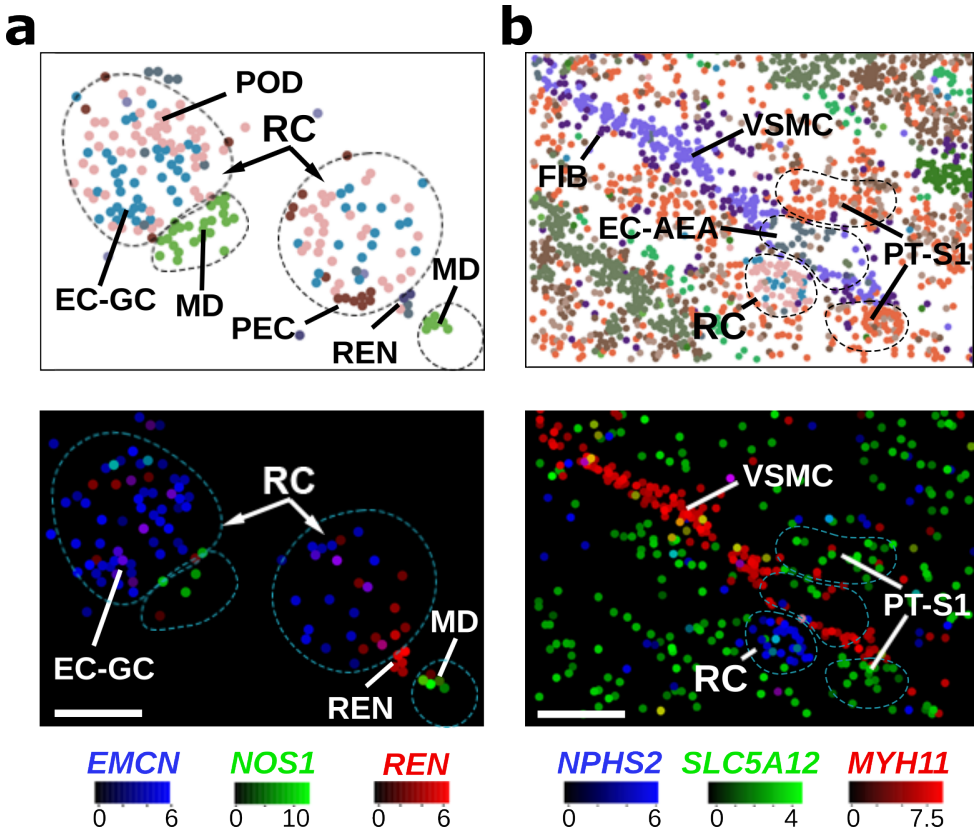



**Figure 2.3. a.** An area of the puck shown in figure S1 colored with (top) predicted l2 subclasses for renal corpuscles and (bottom) mapped expression values for corresponding marker genes (scaled). Scale bar is 100um. **b.** (top) Another area of the puck shown in figure S1 and predicted cell types for the AEAs and surrounding cell types. (Bottom) Mapped expression values for corresponding marker genes (scaled). Scale bar is 100um.

Since the birth of sequencing-based spatial transcriptomics methods, several algorithms have been developed that try to address the mixing of cell types in each capture entity [30]. For our purposes, we chose an algorithm by the same team that invented slide-seq, called Robust Cell Type Decomposition (RCTD) [31] for annotating the beads. RCTD uses a probabilistic model to find a mixture of cell types defined by scRNA-seq that best explains the UMI counts of a given bead. Furthermore, RCTD has the appealing feature of normalizing the effects between platforms, that is, modeling the biases that each capture technology may have (e.g., between snRNA-seq and slide-seq). This feature could further aid automatic label transfer. In a nutshell, given an scRNA-seq reference with annotated cell types and states, RCTD computes the contribution of every cell type/state to every bead and outputs a bead-by-cell type matrix. Figure 2.2b shows the cell type weights for the three populations marked by the genes in panel (a).

To annotate our slide-seq beads, I used RCTD along with our snRNA-seq atlas. To control the complexity of the analysis, I took a hierarchical approach in deconvolution (Methods). In short, I first performed the deconvolution with subclass level 1 (l1) groupings. Then, for cells with a strong belonging to a l1 subclass, I expanded the reference to include subclass level 2 (l2) groups for deconvolution. With this strategy, I created a metadata table for all beads of all pucks with beads in rows and the deconvolution weights for l1 and l2 subclasses in the columns. This table was my reference for any downstream analysis which used cell type annotations.

The results of the annotation at l1 and l2 subclasses are depicted in Figure S1 for one puck. To gain confidence into our annotations, we can look at how well slide-seq can resolve the prototypical structures in the kidney. For example, we can look at renal corpuscle, the main filtration unit of the kidney. The glomerulus is composed of glomerular capillaries, podocytes and mesangial cells and is lines with parietal epithelial cells. Adjacent with the glomerulus is the juxtaglomerular apparatus with renin-secreting cells and macula densa cells that constitute a feedback loops that regulates the renal glomerular filtration rate (Figure S6a). Figure 2.3a shows a representative example of the renal corpuscle with the above components. By comparing the top panel (annotation of the major subclasses) and the bottom panel (marker genes), we

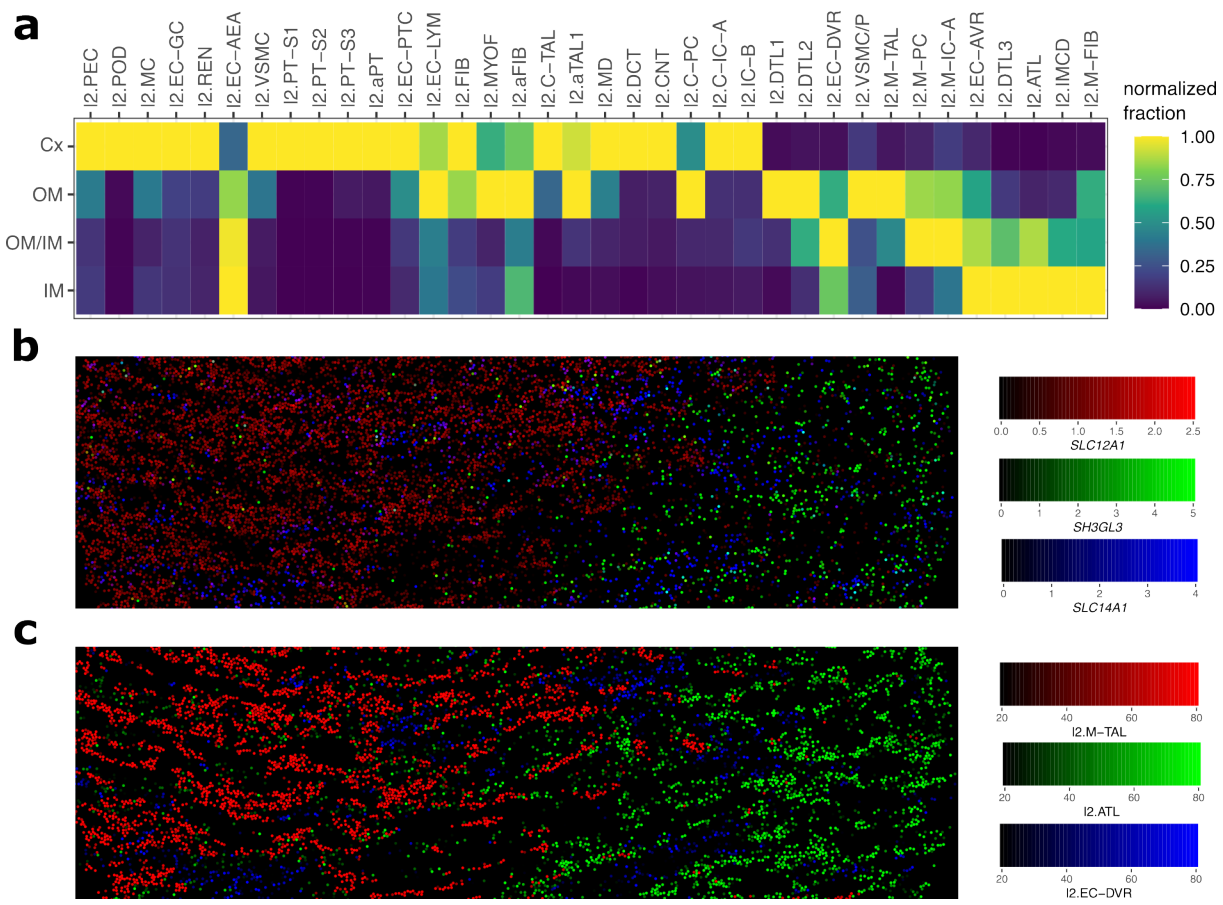



**Figure 2.4.** **a.** Heat map of Slide-seq cell type frequencies along the corticomedullary axis (three individuals). **b & c.** Representative tissue puck region showing the transition of M-TAL (outer medulla, left) to ATL (inner medulla, right) segments. Panel (b) shows the corresponding markers and panel (c) shows the deconvolution weights. Width of the plot corresponds to 3mm.

can visually confirm that RCTD is performing very well at annotating the beads in presence of mixing. The organization of these cell types, obtained by slide-seq largely consistent with what is known about them. Hence, we are able to recapitulate known biology about human kidney using slide-seq.

## 2.5 Distribution of cell types across the depth of the kidney

Some of the major cell types of in the kidney exist in more than one region, for instance TAL are present in both outer medulla and the cortex. While these subtypes are closely related, they present some level of heterogeneity [32]. In the snRNA-seq reference, such populations

were annotated mainly by utilizing the sample metadata (cortical vs. medullary vs. papillary). As such, TAL clusters majorly derived from medullary and cortical samples were annotated as M-TAL and C-TAL, respectively. However, in order to show that these are bona fide subtypes, they need to be validated using an orthogonal method. Since our slide-seq data is generated independently and annotated without using the regional information, it automatically bears the potential to be an orthogonal method to validate these populations.

To validate the regional annotations in the snRNA-seq atlas, we looked at the regional distribution of the cell type calls in the slide-seq data set. Figure 2.4a shows the normalized distribution of the cell types across the regions in the slide-seq data. The plot shows the transition of the populations along the corticomedullary axis, which proximal tubules and glomerular-specific cells mainly in the cortex, and thin limbs in the medulla. In particular, C-TALs are specific to cortex, while M-TAL are mainly present in the outer medulla. The difference between C-PC and M-PC seems more subtle, as some C-PC beads are detected in outer medulla, however, the converse does not seem to be true. An intriguing observation is the transition from outer medulla to inner medulla captured by two slide-seq pucks (figure 2.4b&c). Shown in figure 2.4c is the trasition of ATLs into M-TALs. Similar transitions can be observed between IMCD to M-PC (not shown). These observations are consistent with what is known about the distribution of cell types across kidney and hence, slide-seq is validating the regional annotations of the snRNA-seq atlas.

## 2.6 Cell-cell interactions in the kidney

In addition to validating of the annotations of single-cell data, slide-seq can give us insight into neighborhoods in which cell population interact and potentially identify the spatial localization of newly identified cell types. A straightforward strategy for attacking this problem is to find pairs of cell types that tend be physically closer to one another than expected by chance [33]. Then, one could create a network of cell types and look for communities of interacting

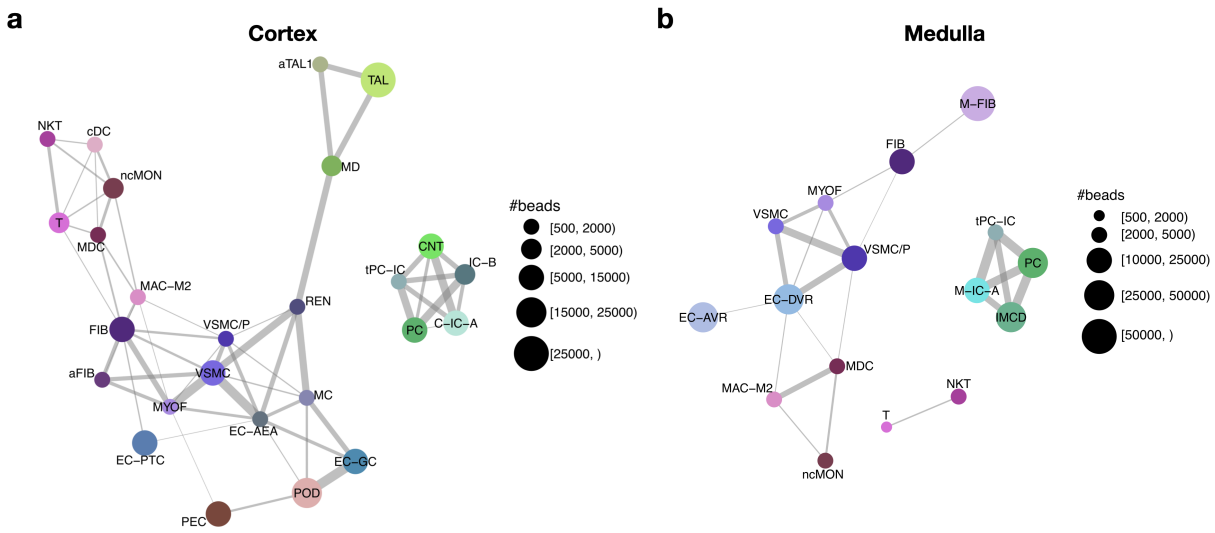



**Figure 2.5.** Proximity enrichment networks for cortex (**a**) and medulla (**b**). Each subclass level 2 cell type is node with size of the node related to the population size (# beads). The thickness of the edges increases with the enrichment. See Methods for details on the construction of the underlying neighborhood graph and enrichment values.

cell types. For this analysis, I separated cortical and medullary pucks because the cell types compositions are very different (Methods).

The interaction networks are depicted in Figure 2.5. In these plots, nodes are cell types and the thickness of the edges is proportional to the strength of the proximity enrichment (Methods). Figure 2.5a shows that, as expected, podocytes (POD), mesangial cells (MC) and glomerular capillaries (EC-GC) are strongly spatially bound. Furthermore, the glomerular cell types are physcially close to the juxtaglomerular apparatus consisting of afferent-efferent arterioles (EC-AEA), renin-secreting cells (REN) and macula densa cells (MD). This also validated the association between AEAs and a population of vascular smooth muscle cells (VSMC, figure 2.3b). Immune cells on the other hand, seem to exist in pockets containing both lymphocytes (T and NKT cells) and monocytes (e.g., macrophages). Immune cells seems to primarily interact with fibroblasts, perhaps in areas with undergoing injury. On the other hand in the medulla (figure 2.5b), immune cells seem to be present around blood vessels (EC-DVR and VSMC/P). In summary, this analysis uncovers the major hubs in which cells interact with each other, either as part of their homeostatic function (e.g., as in glomeruli), or pathology (e.g.,

fibrosis).

## 2.7 Mitochondrial transcripts in slide-seq

In the previous sections, we reviewed cases in which slide-seq can validate the findings of single-cell technologies. We then discussed the application of neighborhood analysis in uncovering the niches. In this section, I briefly want to note an overlooked application of slide-seq: detecting mitochondrial transcripts. In single-cell analysis, mitochondrial transcripts are usually seen as indicators of low-quality samples since increased mitochondrial activity is implicated in cell death [34] which can be driven by single-cell dissociation procedures. Mitochondrial transcripts are less of an issue with single-nucleus RNA sequencing of frozen samples, since the majority of mitochondrial transcripts are cytoplasmic.

Slide-seq can give us new insights about mitochondrial activity in tissues. Unlike single-nucleus RNA-seq, slide-seq does not distinguish between nuclei and cytoplasm, where the mitochondria reside. Additionally, in contrast to single-cell RNA-seq, slide-seq uses fresh frozen sections and is less prone to dissociation-induced artifacts, including cell damage and apoptosis. Hence, looking at mitochondrial transcripts may bring insights previously missed.

To look at the expression of mitochondrial genes, I used a subset of them with high expression (average >500 TPM, or 1 read per 2000 UMI per bead). Figure 2.6a shows a heatmap of normalized expression values for these mitochondrial genes. M-TAL is the cell type with the highest expression of mitochondrial genes. This is consistent with M-TAL having the highest mitochondrial density and oxygen consumption among the tubular segments (chapter 4 of [35]). Slide-seq can also help us create a cell type-specific map of mitochondrial transcripts. Figure 2.6b-c show the specific expression of the ribosomal gene MT-RNR2 in M-TAL in the outer medulla, while MT-CO1 (a part of cytochrome c complex) remains more or less constant across M-TAL and other epithelial cells of the inner medulla.

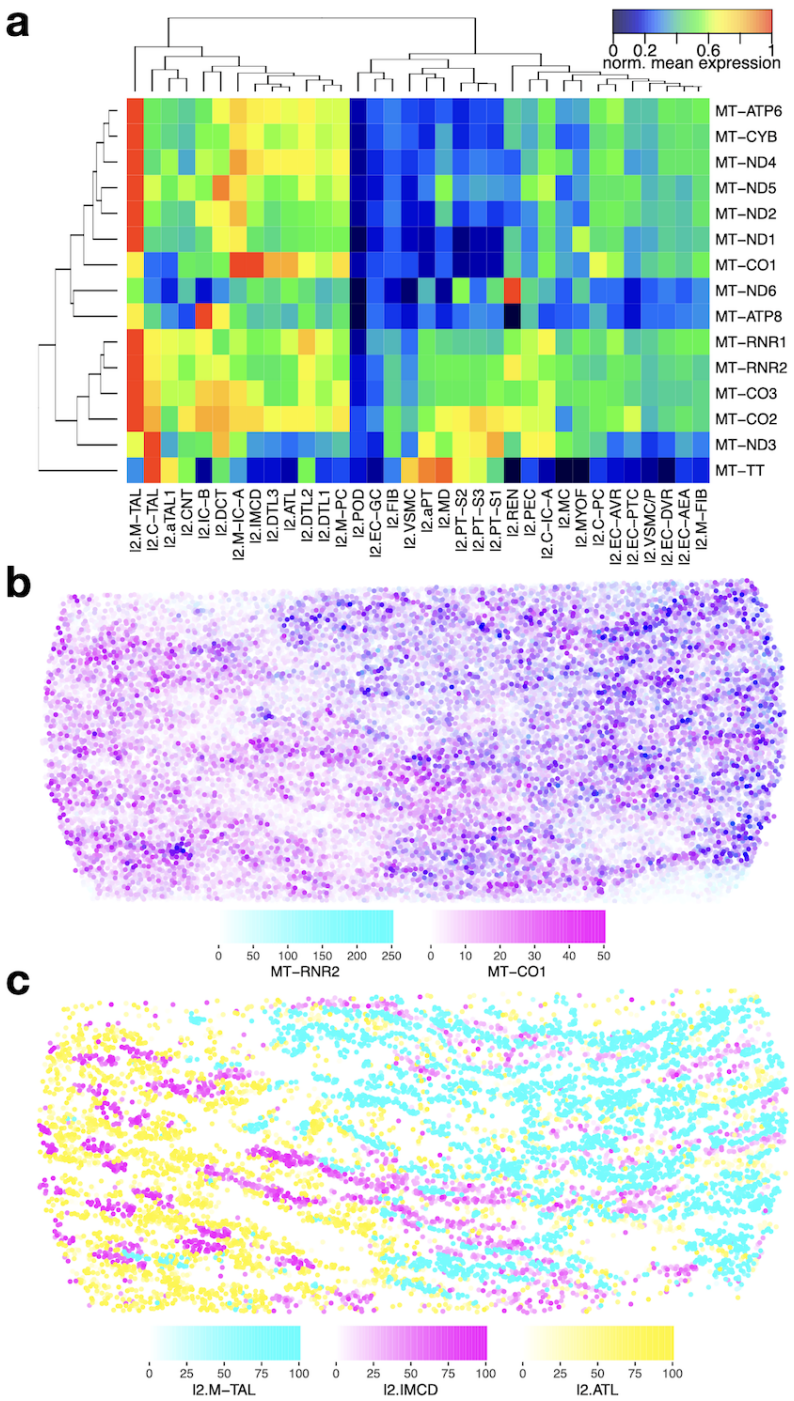



**Figure 2.6. a.** Heat map of average normalized expression values for mitochondrial genes. Bead were normalized by their total UMI count and then genes were normalized to have a maximum of 1 across cell types. **b.** Raw expression of MT-RNR2 (cyan) and MT-CO1 (magenta). Both genes are expressed in the outer medulla (right) and the beads look blue. In the inner medulla (left), MT-RNR2 is downregulated. Size of the field is 3mm across. **c.** Deconvolution weights for relevant cell types in panel (b).

## 2.8 Methods

### 2.8.1 Deconvolution

We used Giotto [33] (v.1.0.3) for handling the slide-seq data and RCTD[31] (v.1.2.0) for the cell type deconvolution. As only reference tissue was used for slide-seq, all degenerative states as well as PapE, NEU, B and N were removed from the snCv3 Seurat object prior to deconvolution. The Seurat object was randomly subsampled to have at most 3,000 cells from each level 2 (l2) subtype and the level 1 (l1) subclasses of ATL and DTL were merged. For each data source, that is, HuBMAP or KPMP (Supplementary Table 2 of Lake et al. [27]), the counts from all beads across all pucks were pooled and deconvolved hierarchically: first, beads with less than 100 UMIs and genes detected in less than 150 beads were removed. Then, the broad l1 subclass annotations in the Seurat object were used to deconvolve all beads (*gene_cutoff=0.0003, gene_cutoff_reg=0.00035, fc_cutoff=0.45, fc_cutoff_reg=0.6*, manually adding REN in the RCTD gene list and merging ATL and DTL subtypes as TL). The prediction weights were normalized to sum to 100 per bead. Beads for which one cell type had a relative weight of 40% or higher were classified as that l1 subclass. Then, for each l1 subclass, all classified beads were further deconvolved using the l2 annotation of that subclass, as well as the remaining subclass l1 annotations (same parameters as l1). Note that, for each l2 deconvolution, the bulk parameters in RCTD were fitted using all beads and then the RCTD object was subsetted to only contain the selected beads for the l2 deconvolution. Classification at subclass l2 was done similar to l1 with the maximum relative weight cut-off of 30% or 50% depending on the stringency needed for an analysis (50% for Figs. 2c,f and Extended Data Fig. 4b and 30% in other analyses). For plotting gene counts, the scaling was performed with the command *normalizeGiotto(gObj, scalefactor=10000, log_norm=T, scale_genes=T, scale_cells=F)*. The marker gene dot plots were plotted using the DotPlot function in Seurat (v.4.0.0).

### 2.8.2 Proximity enrichment

Coarse cell–cell interactions can be revealed by looking for cell types that tend to be in close proximity. For each puck, we created a neighbourhood graph based on Delaunay triangulation in which each bead is connected by an edge to at least one other neighbouring bead, provided that their distance is smaller than 50um. For each pair of cell types, we count the number of times they are connected by edges. Then, the cell type labels are randomly permuted 2,500 times to form the null distribution of the number of connections. The expected number of connections between pairs of cell types is calculated from this simulation and the proximity enrichment is defined as the ratio of the observed over the expected frequency of connections. The network construction and enrichment analysis were performed using Giotto's *createSpatialNetwork* and *cellProximityEnrichment* commands, respectively. Those beads with maximum level 2 weight less than 30% were removed. We further excluded spurious beads that were outside of the visual boundary of the tissue (only for the pucks of which the names start with 'Puck_210113') by manually specifying straight lines that follow the tissue boundary. For cortical pucks (Supplementary Table 2 of Lake et al. [27]), M-PC, C-PC and IMCD labels were relabelled as PC; M-TAL and C-TAL as TAL; and EC-DVR was merged into EC-AEA. Other medullary and cycling subtypes were removed. For medullary pucks, M-PC and C-PC were relabelled as PC; M-TAL and C-TAL as TAL; all DTL subtypes as DTL; and EC-AEA was merged into EC-DVR. Other cortical and cycling subtypes were removed.

To generate the proximity plots in Extended Data Fig. 4, the enrichment values for each cell type pair were averaged across all pucks from the same region and plots were generated using the R package ggGally (v.2.1.2). For the cortex and medulla, respectively, only the connections with mean enrichment values higher than 0.7 and 0.8 were plotted.

## 2.9 Acknowledgements

Chapter 2 is, in part, a reprint of the material as it appears in Lake, B. B., Menon, R., Winfree, S., Hu, Q., Melo Ferreira, R., Kalhor, K., ... & Jain, S. (2023). An atlas of healthy and injured cell states and niches in the human kidney. *Nature*. The dissertation author was the primary investigator and co-first author of this paper.

# Chapter 3

# Developing a highly multiplexed RNA in situ hybridization technique

## 3.1 Introduction

Analyzing single-cell expression of genes in their spatial context plays a critical role in deciphering the complex cellular organization in multicellular organisms. Gene expression in its spatial context is especially important in fields such as embryo development, neuroscience, and in histopathology. The emergence of single-molecule fluorescence in situ hybridization (smFISH) methods allowed us to simultaneously measure several RNA species in single cells by imaging fluorophore-tagged DNA oligos, or probes, that tile the RNA molecules [21]. Because of its high sensitivity, smFISH has become the gold standard assay to measure RNA expression in situ and has been used to show the importance of RNA localization in cell migration, neuron connectivity, and local protein synthesis [36]. However, since smFISH is limited by spectral overlap of the fluorophores, it has limited multiplexing capacity, and does not scale well for tasks such as resolving cellular heterogeneity in complex tissues, which require profiling hundreds of RNA species.

Recently, in situ hybridization techniques with combinatorial encoding have emerged in which the identity of hundreds or thousands of RNA species can be decoded with tens of FISH cycles[19, 37]. Although these methods have increased the multiplexity by 2-3 orders of magnitude compared to smFISH, they typically require longer target RNA transcripts (>1.5kb),

restricting the analysis from important molecules such as neuropeptides and interferons. Furthermore, because of the low signal-to-noise ratio (SNR) from detected transcripts, these methods need high magnification objectives with high numerical aperture (NA), making it difficult and time-consuming to image large regions of interest (ROIs). The low SNR also makes it challenging to apply these methods to human tissues which may have a high autofluorescence background caused by lipofuscin granules, proteins such as collagen and elastin, or mitochondria [38, 39, 40].

In parallel to smFISH-based methods, other strategies have been developed that use padlock probes [41]. Padlock probes are a special type of DNA oligos whose 5' and 3' ends are complementary to the target nucleic acid. For in situ applications, the probes are designed in such a way that in order to hybridize to the target, the 5' end anneals immediately downstream of the 3' end. The two ends can then be ligated to create a circular DNA. This is followed rolling circle amplification (RCA) [42] which creates hundreds of copies of the original probe in situ. This amplification step boost the SNR from individual transcripts. However, these methods are associated with high probe set expenses and complex decoding procedures. They further lack an efficient approach to stain the cell bodies for segmentation [43, 44, 16].

In this chapter, I describe our efforts to create a new multiplexed RNA in situ hybridization technology to overcome the above-mentioned limitations [45]. We named this technology DART-FISH, for Decoding Amplified taRgeted Transcripts with Fluorescent In Situ Hybridization. I first describe the rationale for DART-FISH. Then I describe the key experimental challenges that I solved. In the second half of this chapter, I detail the computational methods that I developed for designing a DART-FISH experiments, as well as the pipeline to process the outputs.

## 3.2 DART-FISH Framework

DART-FISH involves in situ feature generation by padlock probe capture of targeted transcripts and rolling circle amplification (RCA), followed by a highly robust decoding process of sequential isothermal hybridization [46, 47]. Below, I detail these two stages. The step-by-step

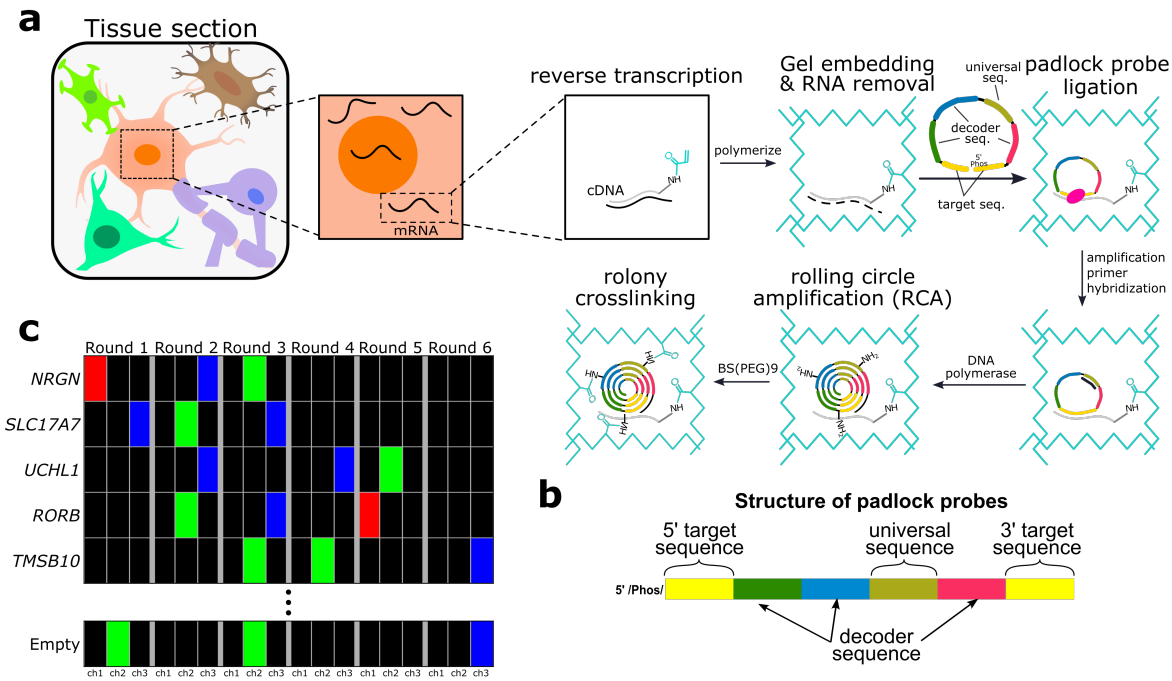



**Figure 3.1. a.** DART-FISH rolony generation workflow **b.** The structure of padlock probes. The 5' phosphate group is essential for ligation. The universal sequence is used for amplification. **c.** A part of an example codebook with $n = 6$ and $k = 3$

protocols will be in the methods section.

### 3.2.1 Rolony generation

The steps for rolony generation are depicted in figure 3.1a. Namely, RNA molecules in fresh-frozen tissue sections are fixed with paraformaldehyde (PFA), permeabilized, and then reverse-transcribed with a mixture of random and poly-deoxythymidine (dT) primers. The reverse-transcribed DNA (cDNA) molecules are then fixed in place, followed by RNA digestion. cDNA molecules are then hybridized with a library of padlock probes and circularized at a high temperature to ensure specificity [42]. The circularized padlock probes are then rolling-circle-amplified, generating RCA colonies in situ (rolonies) with hundreds of copies of barcode sequences concatenated in the form of a DNA nanoball. The rolonies are fixed in place, usually by crosslinking them to a polyacrylamide gel.

### 3.2.2 Decoding

To encode the identity of hundreds of genes, we label them with gene-specific DNA barcodes. These barcodes are placed on the backbone of the padlock probes(Figure 3.1b), thus upon RCA, hundreds of copies of the barcodes will be concentrated within the rolony. The task of decoding is to "read" the DNA barcodes from rolonies and assign them to the gene they are encoding. The "reading" process is done by sequential rounds of fluorescent imaging. Below, I explain the barcoding strategies that we use in DART-FISH.

#### 3.2.2.1 Barcoding scheme

To achieve high multiplexity, that is encoding hundreds of genes, within only a few rounds of imaging, we use combinatorial labeling to generate the gene-specific barcodes. The encoding and decoding scheme we used was inspired by a paper by Gunderson et al [48]. In our barcoding scheme, $n$ rounds of imaging are performed where every barcode is “on” in exactly $k$ rounds and “off” in other rounds. When “on”, the barcode signals in one of the three fluorescent channels; it emits no fluorescence when “off”. Figure 3.1c is an example of a codebook, a table that shows the expected signal pattern all genes. With the design rules mentioned, in $n$ rounds of imaging, a total of $\binom{n}{k}3^k$ unique barcodes can be generated, allowing us to encode hundreds of RNA species with limited rounds of decoding ($n = 6$ and $k = 3$ in figure 3.1c with 540 valid barcodes). This can be extended to 7 rounds of decoding for up to 945 genes ($k$=3), 8 rounds of decoding for 5670 genes ($k$=4), and so on. Hence, DART-FISH uses a barcoding strategy that can theoretically generate enough diversity to encode hundreds to thousands of genes within less than 10 rounds of imaging.

#### 3.2.2.2 Decoding by hybridization

The next step is to implement the decoding chemistry in such a way that it is fast, simple and robust. We based our chemistry on hybridization of short oligonucleotides because they can diffuse fast, and their annealing strength can be controlled by their length, base content

and temperature. Specifically, the gene-specific barcodes are created by the concatenation of $k$ 20-nucleotide-long decoder sequences placed on the backbone of padlock probes. Every column in the codebook (figure 3.1c) represents a unique decoder sequence and is conserved between all DART-FISH experiments. These sequences are derived from Illumina BeadArray technology and have limited cross-hybridization with similar bonding energy ($T_m \approx 55°C$) [48] (Supplementary Data 1). In each round of imaging, three fluorescent decoding probes corresponding to that round are hybridized and imaged. Rolonies will be "on" only if a decoding probe that corresponds to one of their decoder sequences is present. After imaging, the decoding probes are stripped and washed away at room temperature with a chemical solution containing formamide that denatures the decoding probes. This prepares the sample for the next round.

In summary, the decoding process in DART-FISH is based on hybridization of short fluorescent oligos. This strategy has several advantages compared to other chemistries used for decoding [16, 43, 49, 50]:

1. All steps are done at room temperature and hence no specialized equipment is needed for temperature control
2. It is based on passive diffusion and hybridization, free from enzymatic reactions. Hence, a strong signal can be obtained by short incubation times. Similarly, the signal can be stripped off rapidly.
3. It scales well to detect hundreds of genes within several rounds of imaging. The number of decoding probes needed for $n$ rounds of decoding is only $3n$ and is independent of the number of target genes.

### 3.2.3 Rooms for improvement

The number of transcripts detected per cell is directly related to how useful a spatial data set can get. More detected transcripts per cell can lead to increased granularity at resolving cell types and cell states. The methods based on padlock probes tend to have low sensitivity, that is, they can detect a small fraction of transcripts that a cell contains. For example, the early version

of ISS is thought to be around 5% (in cell culture) [16, 15]. Targeting human tissues becomes extra challenging since they tend to have compromised RNA content and quality. DART-FISH as described above, also suffered from low sensitivity. In this regard, our early attempts to generate useful data from human tissues, especially human kidney were unsuccessful. I tried several ways to increase the sensitivity of the assay and in the following sections I will detail two main ways that were successful.

## 3.3 The secret life of cDNA molecules

### 3.3.1 Motivation for a cDNA stain

Reverse-transcription (RT) is one of the key steps in the DART-FISH protocol. It is performed as an overnight reaction on the first day, and the RT product (i.e., the cDNA) should be retained in place until the RCA reaction on the 3rd day is performed. Between RT and RCA, there are two important steps: 1) RNA digestion, 2) padlock probe ligation. The cDNA is expected to form a duplex with RNA upon formation, however, if it is not kept in place by other means, it will float away once the RNA is digested. On the other hand, since the ligation step is performed at 55°C overnight, there is a chance for some crosslink reversal, protein denaturation and ultimately cDNA loss. Hence, a tool for visualizing the cDNA molecules can be highly valuable to track the cDNA levels in different steps of the protocol and troubleshoot in case there is a cDNA loss along the way.

Additionally, a cDNA stain can be useful in other ways. Since the cytoplasm contains RNA molecules, and we randomly prime all RNA for RT, then the cDNA stain could serve both as a cytoplasmic stain and an internal control for RNA amounts and quality. This necessitates the stain to be simple enough that it can be incorporated into all DART-FISH experiments.

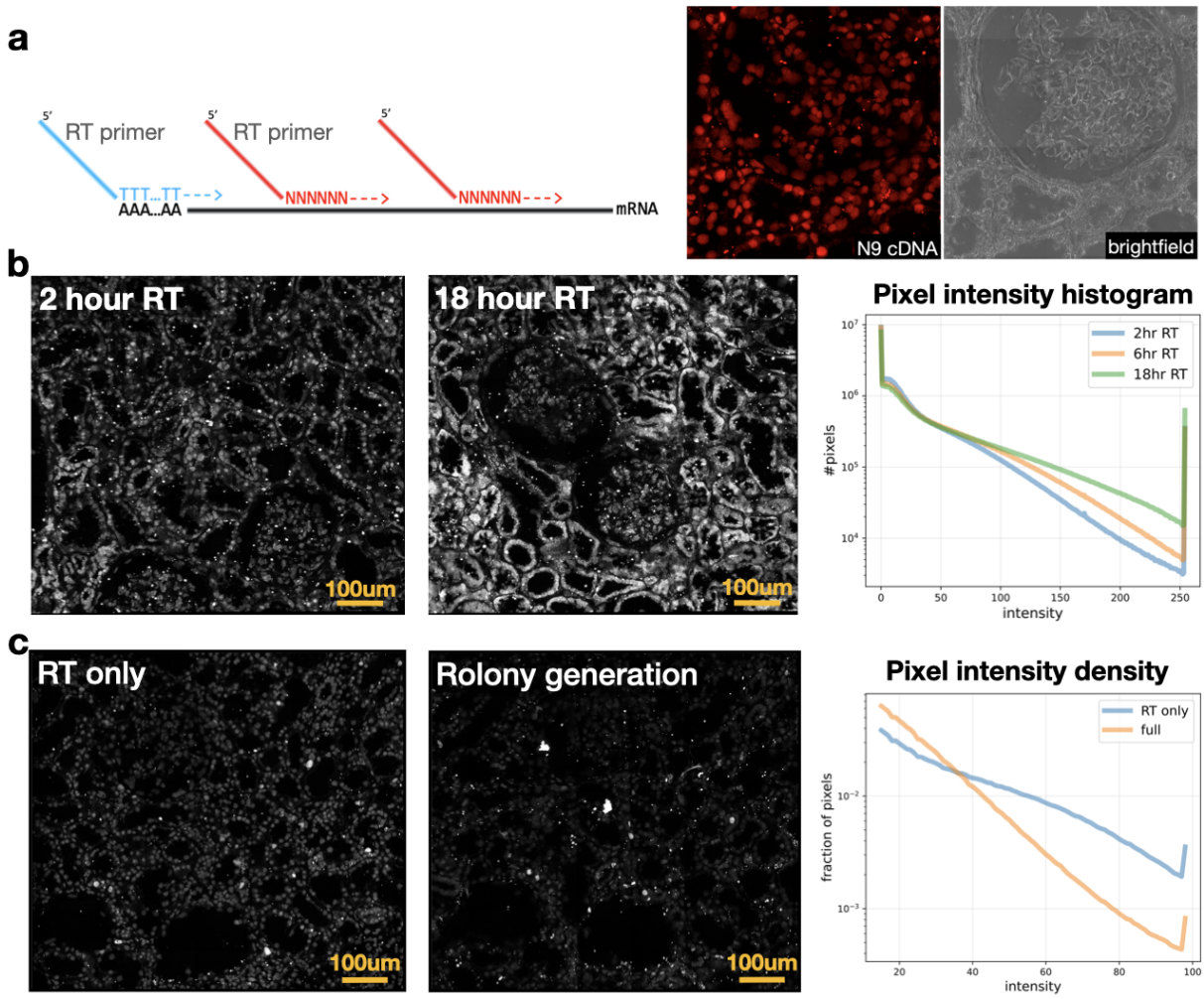



**Figure 3.2. a.** (left) The design of new RT primers that enable cDNA staining (RiboSoma). (right) example of N9 cDNA stain in reference to the brightfield image of the same field of view **b.** (left and middle) RiboSoma (N9 cDNA) after 2 hours and 18 hours of RT. (right) Quantitative comparison of pixel intensity values between conditions show higher cDNA intensity for longer RT reactions. **c.** (left) RiboSoma immediately after RT and (middle) after rolony generation. (right) Pixel intensities drop after rolony generation compared to after RT

### 3.3.2 A stain for total cDNA

To assess the RNA content in human tissues as well as the retention of the cDNA molecules in situ, I added a 5' handle to the reverse-transcription primers to enable the collective visualization of all cDNA molecules with fluorescent oligos (Figure 3.2a). We call this labeling method RiboSoma because the resulting signal labels the cell bodies. To test that RiboSoma is measuring the cDNA levels, I performed an RT time course on human kidney sections, where on

parallel sections I change the length of the RT reaction with the expectation that the levels of signal increase by longer incubation time. Figure 3.2b indeed shows the increase in levels of signal by longer reaction times.

### 3.3.3 Measuring loss of cDNA molecules

With our new tool for visualizing cDNA molecules, RiboSoma, I examined the cDNA levels after different steps of the protocol. Initially, by measuring RiboSoma after RT and after RCA, I observed that the RiboSoma levels are significantly lower after RCA (Figure 3.2c). The two leading hypotheses to explain this observation were: 1) cDNA loosening during one (or more) step of the protocol, 2) cDNA digestion by the 3' exonuclease activity of the Phi29 polymerase used for RCA. I showed that the second hypothesis was wrong and by further inspection found that it is during the incubation at 55 °C needed of hybridization/ligation that the cDNA was loosening. This data shows that cDNA molecules are being lost during a critical step of the protocol and suggests that by increasing the crosslinking, we may be able to save more cDNA molecules and increase the overall sensitivity of the assay.

### 3.3.4 Increasing cDNA crosslinks

To keep the cDNA in place, the initial DART-FISH protocol was inspired by FISSEQ [51] and used BS(PEG)9 to crosslink the cDNA. BS(PEG)9 is a bifunctional NHS ester that acts on primary amines. To enable crosslinking by this agent, a small amount of a modified dUTP, aminoallyl-dUTP, is added to the RT reaction. I hypothesized that extra primary amines on cDNA molecules would enhance the crosslinking efficiency. To achieve this, I proposed using RT primers with a 5' amine modifiers. This modification guarantees that even short cDNA molecules will have at least one primary amine on them. To test this idea, we performed side-by-side experiments with parallel tissue sections from the same tissue block using unmodified (standard) and amine-modified primers. Figure 3.3a shows the result on mouse brain sections. Not only did this modification intensify the RiboSoma signal, but also increased the number of rolonies per

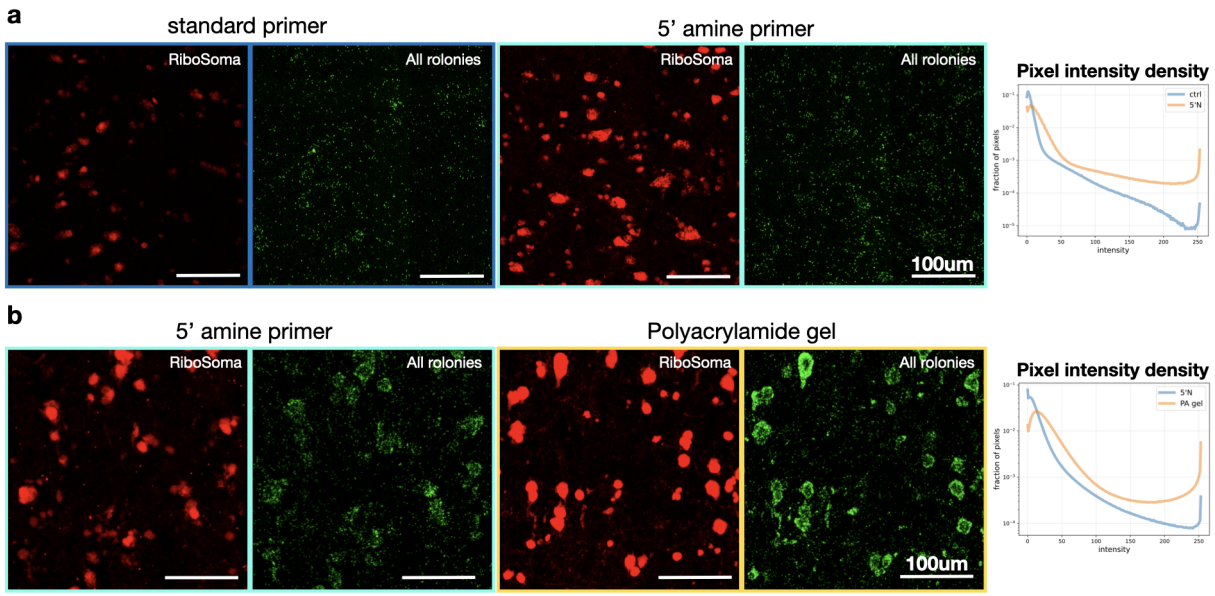



**Figure 3.3.** RiboSoma (N9) images are shown in red, and all rolonies are pseudocolored in green. **a.** Comparison of RiboSoma and rolony count between an experiment with standard primers or 5' amine-modified primers. **b.** Comparison of RiboSoma and rolony count between an experiment with 5' amine-modified primers and acrydite-modified primers. Data in both panels are from human brain.

unit area by more than 50%. In summary, increasing the level of cDNA crosslinking increased the sensitivity of DART-FISH.

Motivated by the improvements mentioned above, I reasoned that the ultimate way of keeping the cDNA in place would be to covalently crosslink it to a tissue-independent matrix. Inspired by the recent studies, I tried embeddeding the tissue in polyacrylamide (PA) gel under conditions that would result into covalent crosslinking of the cDNA into the gel matrix [52, 53]. This needed a few modifications to the protocol (Methods):

1. Using RT primers with 5' acrydite modification. This modification is automatically incorporated into the gel upon polymerization
2. Acryloyl-X treatment immediately after RT. This chemical reacts with primary amines to add an acryloyl group on them. Primary amines may be found on cDNA, as well as the glass and the remaining proteins in the tissue.
3. Incubation with gel monomer buffer and eventually gel formation using APS and TEMED

chemistry.

The gel embedding further increased the intensity of RiboSoma. Moreover, the rolony count per unit area also increased by more than 50% (Figure 3.3b). Taken together, the cDNA stain that I developed, RiboSoma, helped us identify a systematic problem with the original protocol in its cDNA retention capabilities. I demonstrated that embedding the cDNA molecules in a PA gel significantly enhances the retention of the cDNA throughout the rolony generation procedure and increases the sensitivity by more than 2 folds, a point not taken into account in previously published methods. Looking forward, cDNA embedding is compatible with gel expansion procedures ([54]) and can reduce optical crowding and aid with decoding.

## 3.4 Design and production of padlock probes

A key component of a DART-FISH experiment is the probe set u1sed. In every probe set, each gene is usually targeted by multiple padlocks. The reason is that the sensitivity of every single probe is quite low and has high variability [55] and thus pooling multiple probes targeting the same gene can increase the detection rate and lower the variability. In fact, it has been shown that the sensitivity keeps increasing by using up to 30 probes per gene without plateau [56]. Hence, in our bid to increase the sensitivity of DART-FISH (see section 3.2.3) we could increase the probe counts for higher detection rate. Note that, the gain from this strategy will be additive to any other improvements in the probe design, or the chemistry (section 3.3.4)

### 3.4.1 Cost of synthesis for padlock probes

The main barrier to increasing the number of padlock probes is cost. Padlock probes tend to be longer (>80) than what standard methods can synthesize accurately. For example, ordering oligonucleotides from IDT, it is customary to order them as either Ultramers or with PAGE purification. An ultramer oligo of length 100 costs about $50. If 200 genes are targeted each with 10 probes, the cost will be about $100,000. Ordered as standard oligos, the same number of

probes will cost around $50,000. Such a high cost maybe manageable for large project in data production phase, but is prohibitively expensive for more exploratory phases and for smaller labs.

A cheaper alternative to direct synthesis of padlock probes is ordering them as oligo pools. Large number of oligos can be synthesized at a miniaturized scale on silicon substrates for a small fraction of the cost of direct synthesis. For example, the same 2000 oligos will cost only $2000. Moreover, the cost of synthesis per probe reduces by ordering larger pools. However, the challenge introduced by oligo pool is their small amount. Hence, post-synthesis amplification needs to be performed to obtain large amounts of probes.

### 3.4.2 Synthesizing padlock probes from oligo pools

Our lab developed a protocol for synthesizing padlock probes from oligo pools [57]. The steps of the protocol are depicted in Figure 3.4 a. In this workflow, the oligo pool must be designed with specific PCR handles on the two ends which are subsequently used for amplification (figure 3.4a top). This effectively creates an unlimited source of probes that can be reamplified whenever needed. Next steps include digestion of the complementary strand, and then removal of the two amplification handles. First, the complementary strand is removed by Lambda Exonuclease which preferentially digests DNA strands with 5' phosphate. The 5' phosphate is placed on the reverse primer of the PCR reaction. The 5' amplification handle is removed by USER, which excises the uracil that is incorporated by the PCR's forward primer. Finally, the 3' amplification handle is removed through a restriction digestion with the enzyme DpnII that recognizes *GATC* motifs. Thus, this probe production protocol offers a solution to synthesizing thousands of padlock probes in a pooled format with low cost.

While useful, the protocol above tends to be tedious and low-yield. The final product is usually very messy: when visualized on a denaturing polyacrylamide gel, a large fraction of the product is under-digested that runs above the desired length. A significant amount of the end product is over digested which creates a strong smear shorter than the desired length (Figure

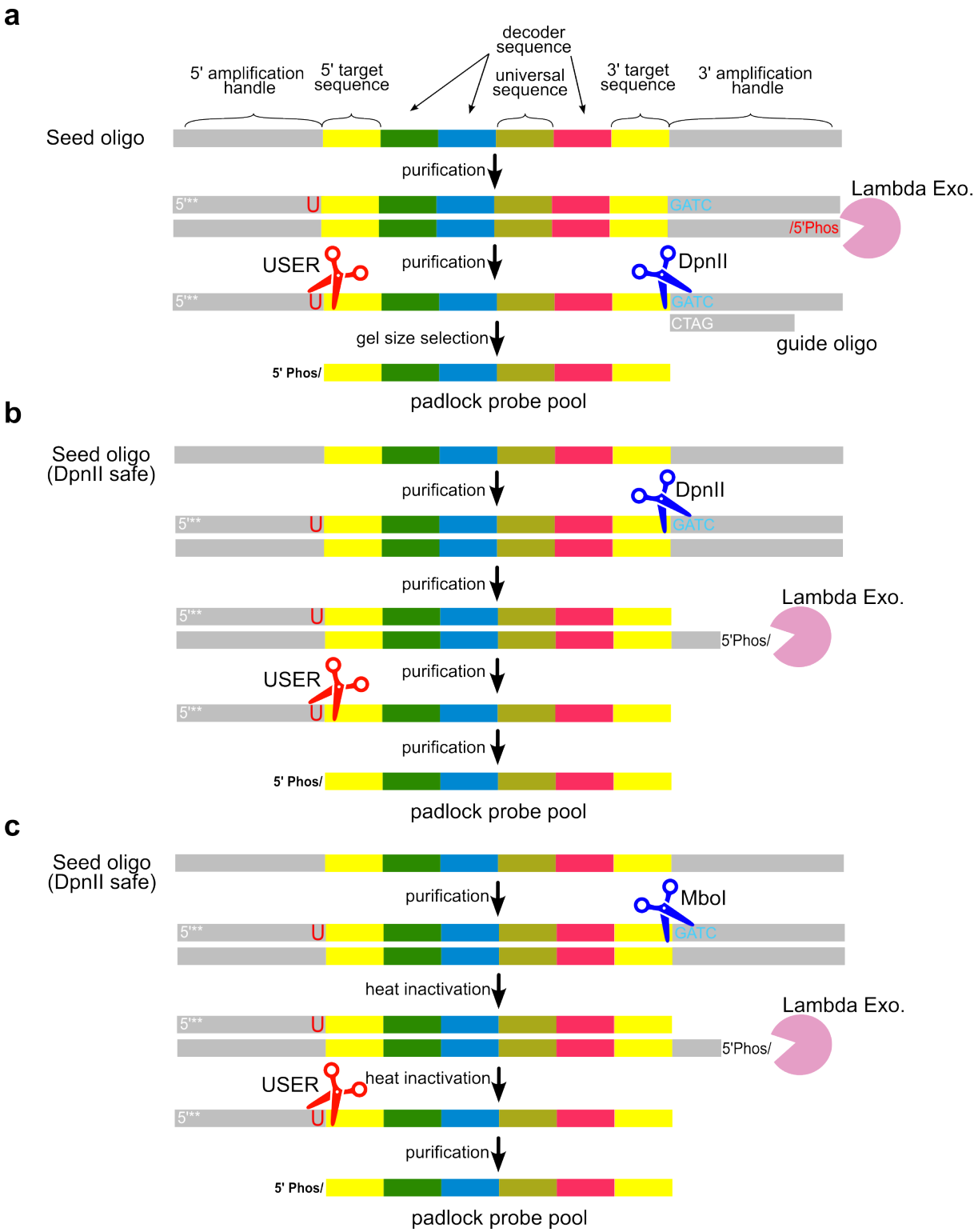



**Figure 3.4.** **a.** Initial method from Diep et al (2009) **b.** Improved method free of gel size selection **c.** Enhanced one-pot digestion for padlock probes

3.5a). Consequently, to get a clean probe set, one would need to physically cut the band with full-length probes (depicted by a star in figure 3.5a) out of the denaturing gel. This size-selection step creates a major bottleneck, in efficiency, reproducibility and the convenience of the protocol. All in all, the low efficiency and high complexity of this workflow limit its utility and preclude it from being regularly used.

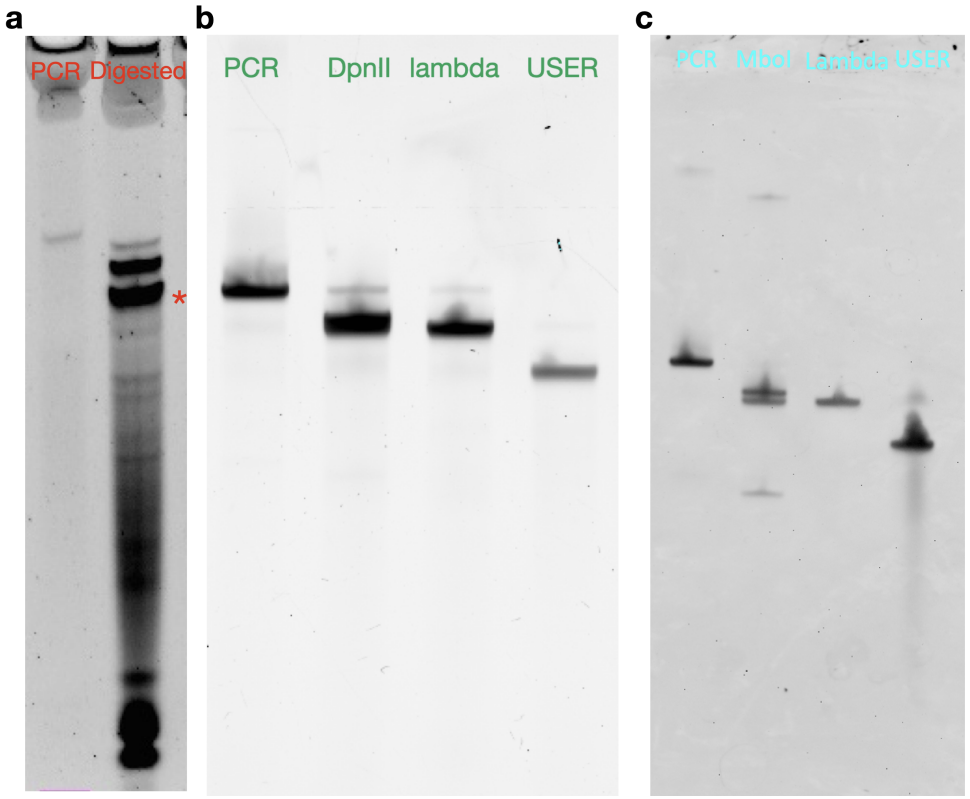



**Figure 3.5.** Denaturing PA gels comparing different enzymatic probe production methods in figure 3.4. **a.** The digested product size distribution from Diep et al (2009) (figure 3.4a) that needs gel size-selection. The band with a red star (*) shows the desired product. **b.** Size distribution for the improved method depicted in figure 3.4b **c.** Size distribution for the one-pot digestion method in figure 3.4c.

#### 3.4.2.1 A size-selection-free workflow for padlock probe synthesis

To create a more streamlined probe production workflow with higher yield, I set out to identify the problems with Diep et al's protocol. I recognized that while the underdigestion (upper bands in figure 3.5a) could be the result of compounding inefficiencies in USER and

DpnII digestions, the overdigested products (running below the desired product in figure 3.5a) can only be caused by non-target digestions by DpnII and not USER. I hypothesized that the overdigestion by DpnII is caused by non-target GATC motifs that are present in the probe sequence. Restriction enzymes are expected to act only on double-stranded DNA (dsDNA), hence the guide oligo is designed to create dsDNA exactly where the DNA needs to be cut (the 3' amplification handle). However, different single-stranded DNA (ssDNA) molecules can transiently interact with each other and form dsDNA structures that include GATC motifs in non-target locations. These ectopic double-stranded GATC regions can be cut by DpnII and create the over digested products seen on the PA gel. Statistically, for a random 40bp stretch, one expects a $40/4^4 = 0.15$ chance of having a GATC motif, meaning that around 15% of the products are prone to be non-specifically cut by DpnII. Furthermore, I realized that some decoder sequences that were historically used in the lab had a GATC motif.

The solution that I proposed was as follows:

1. Removing probes with target sequences with GATC
2. Replacing decoder sequences that have GATC with ones that do not
3. Careful assembly of the padlock probe sequences to avoid GATC motifs formed at the junctions of different components

With no internal GATC motifs other than the one to cut the 3' amplification handle, we could safely perform the restriction digestion on the dsDNA after PCR with maximal efficiency. Thus, I introduced the modified probe production protocol that is depicted in Figure 3.4b (Methods). In this protocol, DpnII digestion is performed on purified PCR products. This is followed by Lambda Exonuclease digestion to remove the complementary strand. Finally, USER reaction is done on ssDNA. Note that, the product of every digestion needs to be purified. Figure 3.5b shows the product after every step of the protocol. The end result (under USER) is now clean enough to obviate the need for gel size-selection: the underdigested band is much dimmer than the desired product, and the overdigested smear is also very faint.

#### 3.4.2.2 A one-pot digestion recipe for padlock probes

The protocol in the previous section improved the padlock probe production workflow in two important ways: 1) higher yield for the same amount of PCR product, 2) more straightforward and less cumbersome without gel size-selection. However, as shown in figure 3.4b, there is one purification step after every digestion. Every purification step, which in this protocol is with silica-based spin columns, have an efficiency of less than 80%. Moreover, every purification step add 1 hour of hands-on time to the workflow. Hence, an alternative method with fewer purification steps will be highly desired if it has higher yield and shorter hands-on time.

To create a more efficient and simple protocol for padlock probe production, I set out to create a one-pot digestion recipe with no purifications in between. There are to challenges that need to be addressed for this:

1. The reaction buffers have to be compatible for all enzymes: This is not the case for the current set of enzymes as DpnII's buffer is of pH 6 and Lambda Exonuclease's buffer has pH 9.4.
2. Lambda Exonuclease needs to be completely inactivated before USER is added to the reaction. The smallest activity of Lambda Exo will remove the USER digestion products.

We created a new protocol that addresses these points (figure 3.4). To address the first point, we chose MboI, a restriction enzyme with the same motif as DpnII. MboI is 100% active in CutSmart buffer. Lambda Exo's buffer is also compatible with CutSmart and USER does not have a specific buffer requirements. To address the second point, we first tested heat-inactivation for MboI and Lambda Exo following manufacturer's protocol (75°C for 10 minutes), however, we obtained little product suggesting the incomplete inactivation of Lambda Exo. We eventually obtained the best inactivation results by using higher temperature and longer duration (95°C for 20 minutes) in combination with EDTA. Figure 3.5c shows the step-by-step results of this protocol. Even though this workflow has more than 50% yield compared to the previous section with 2-3 hours shorter hands-on time, the quality of their end products is comparable.

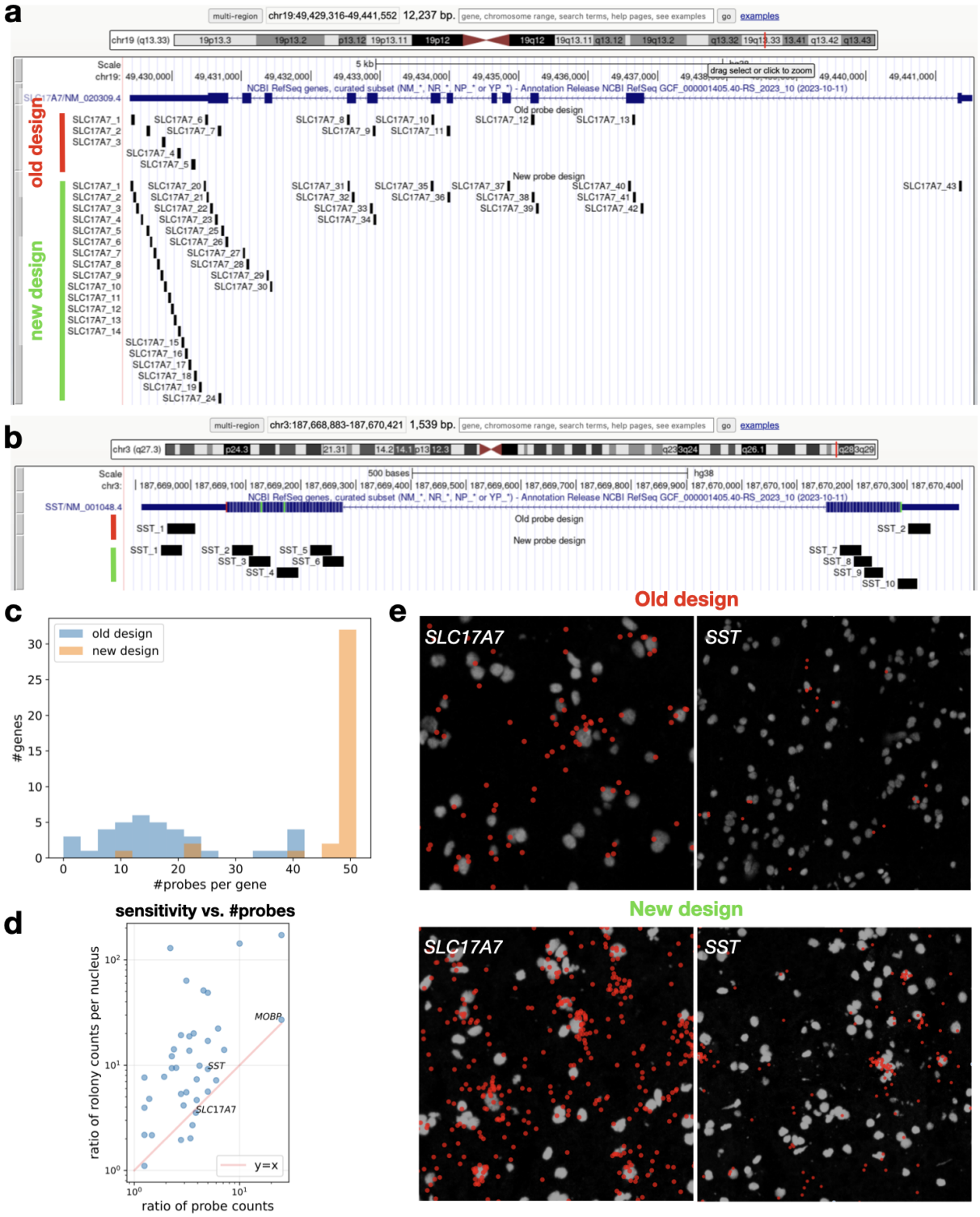
In summary, this section summarized iterations of my efforts to create a simple and efficient protocol for production of padlock probes. Now with these protocols, we can efficiently use oligo pools to create probe sets that consist of tens of thousands of padlock probes at a cost that is accessible to many labs.

### 3.4.3 More probes for higher sensitivity

As discussed in sections 3.2.3 and 3.4, our large-scale padlock probe production workflow allows us to increase the number of probes per gene to improve the detection efficiency of DART-FISH. In this regard, I optimized our original target selection pipeline (ppDesigner from [57]) to generate more candidate target sites by more rigorously tiling the constitutive exons (Figure 3.6a). Additionally, since shorter genes have fewer candidate probes, to compensate for their length, I further modified the pipeline to allow for overlapping targets (Figure 3.6b). Figure 3.6c shows the distribution of probes in our old design (in blue) versus the new design in orange. With the new pipeline we could target genes with more than 3-times more probes. I also chose different parameters for target selection; for example, the target length was reduced from 50 nucleotides to 40.

To evaluate the changes in our probe design strategy, we applied both designs to parallel tissue sections from human motor cortex. The change in the probe design resulted in a substantial increase in the sensitivity. Figure 3.6e provides a visual comparison between the old and new designs. For the short gene *SST*, the old design would lead to low-confidence calls of $SST^{+}$ cells, while the new design can mainly alleviate this issue. For a more quantitative comparison, figure 3.6d shows that all genes witnessed an increase in their absolute counts. The increase in sensitivity for each gene correlated with the increase in the number of probes used to target that gene. Overall, the sensitivity boost was roughly 6x, out of which I think about 3x would be attributed to the probe counts and the rest to other design choices and batch effects.

**Figure 3.6.** **a.** Genome browser view of the target sites for padlock probes for gene *SLC17A7*. Each black bar is a probe. The section in red labels the probes that were design with the lab's old pipeline while the green section depicts the output of my new pipeline. **b.** Same as panel a for the short gene *SST*. Note the overlapping targets in the new design. The new design outputs 6x more probes. **c.** Histogram of number of probes per gene in the new and old probe selection pipelines. **d.** Relation between sensitivity gain and increase in probe counts per gene. Sensitivity in y-axis is defined as the fold-change in the average number of transcripts per nucleus between the two pipelines. X-axis shows the fold-change of the number of probes in the two designs (#probes in new pipeline/#probes in old pipeline) **e.** Each plot shows the decoded transcripts of a single gene (*SLC17A7* or *SST*) overlaid on a nuclear stain. Top panels are produced using a probe set with the old pipeline, bottom panels with the new pipeline.


a
multi-region
chr19:49,429,316-49,441,552
12,237 bp.
gene, chromosome range, search terms, help pages, see examples
go
examples
chr19 (q13.33)
Scale
chr19:
5 kb
drag select or click to zoom
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI RefSeq GCF_000001405.40-RS_2023_10 (2023-10-11)
SLC17A7/NM_020309.4
Old probe design
New probe design
old design
new design
b
multi-region
chr3:187,668,883-187,670,421
1,539 bp.
chr3 (q27.3)
500 bases
hg38
SST/NM_001048.4
c
old design
new design
#genes
#probes per gene
d
sensitivity vs. #probes
ratio of rolony counts per nucleus
ratio of probe counts
MOBP
SST
SLC17A7
y=x
e
Old design
SLC17A7
SST
New design

## 3.5 Discussion

DART-FISH is a cost-effective technology capable of fast decoding on relatively large tissue sections. Using our protocol for padlock probe production from oligo pools, the cost of synthesis per gene scales sublinearly with the number of genes. Hence, oligo pricing will not hinder scaling the probe set to tens of thousands of transcripts. Moreover, DART-FISH does not need any specialized equipment for neither rolony generation nor decoding. The decoding process is relatively fast because it depends on the diffusion and hybridization of very short oligos and a strong signal can be obtained by 5-10 minutes of incubation with the fluorescent decoding probes at room temperature. Likewise, stripping and washing away the unbound decoding probes is straightforward and fast at room temperature. This process can be performed on a stationary glass-bottom petri dish or a coverslip mounted on a microscope and does not require reaction chambers or flow cells with sophisticated temperature control. The large size and the bright signal of the rolonies permit the use of 20x objective lenses for decoding which makes it possible to image centimeter-sized samples in a manageable time with an ordinary confocal microscope.

What distinguishes DART-FISH from other techniques of a similar class is how the cDNA molecules are treated. We demonstrated here that embedding the cDNA molecules in a polyacrylamide gel significantly enhances the retention of the cDNA throughout the rolony generation procedure and increases the sensitivity, a point not taken into account in previously published methods. Additionally, we introduced RiboSoma, a cDNA labeling technique, as a cell morphology marker which reveals more information about cell bodies than nuclear stains. We anticipate that this tool can be highly useful for cell body segmentation, particularly in thicker samples.

Due to its streamlined nature and simplicity, the basic DART-FISH chassis described here can be effectively extended in multiple ways. The workflow can be combined with antibody staining, for instance, to target extracellular factors such as matrix proteins and cell-cell communication molecules to enhance the definition of cell-cell interactions in pathological niches.

The thickness of tissue sections could be increased for higher resolution mapping of neighborhoods and cell connectivities; while increasing section thickness to 20-30um should be readily achievable, other strategies in sample mounting and handling may be necessary to increase the diffusion into even thicker sections (>100um). Padlock probes could also be designed to anneal directly to mRNA followed by circularization using an RNA-mediated DNA ligase, which would skip the cDNA synthesis and can improve the detection sensitivity.

## 3.6 Methods

### 3.6.1 Probe design

For short genes (length < 1.5kb), we defined the constitutive exon as the union of all isoforms in GencodeV41. For other genes, the constitutive exons were defined as regions in RefSeq where at least (33% for brain, 50% for kidney) of isoforms overlap. We used a modified version of ppDesigner [57] (https://github.com/Kiiaan/sppDesigner) to find padlock target sequences along the constitutive exons. ppDesigner was run on two settings: 1) no overlap between probes allowed, 2) overlap of up to 20nt allowed. Individual arms were constrained between 17nt and 22nt long with the total target sequences no longer than 40nt. The resulting target sequences were aligned to GRCh38/hg38 with BWA-MEM[58] and sequences with MAPQ<40 or secondary alignment were removed. We further removed probes that have GATC (DpnII recognition site). For the brain, a maximum of 50 probes per gene were selected prioritizing the non-overlapping set. For the kidney, a maximum of 40 probes per gene were selected with no overlap. Finally, the target sequences were concatenated with amplification primer sequences, universal sequence, and gene-specific decoder sequences to produce final padlock probe sequences (figure 3.1b) and were ordered as an oligo pool from Twist Bioscience (South San Francisco, CA). Amplification primer sequences can be found in Table S1.

To select a set of barcodes, we computationally created all possible barcodes in the compact format: an $n$ digit barcode with "1", "2" and "3" representing each of the three fluorescent channels and "0" indicating off cycles. For example, the barcode for *RORB* in Fig. 1c is "132000" in the 6-digit format. This amounted to 480 and 840 multi-color barcodes for brain and kidney, respectively. We then used a brute force algorithm to find the largest subset of barcodes, $Q$, in which every pair had a Hamming distance > 2. Followed by this, we created a graph, $G$, in which every possible barcode is a node, and pairs of nodes are connected with edges if their Hamming distance is 1. We then found a maximal independent set (MIS, networkx v2.6.2) that included the nodes in $Q$. This method ensures that every pair of barcodes in the

MIS have Hamming distance >1. Because the algorithm for finding MIS is random, we ran it 20,000 times and selected the largest MIS across the runs. For the brain, the MIS consisted of 159 barcodes, 121 of which were randomly assigned to the genes. For the kidney, the MIS had 269 barcodes. We randomly added 31 additional barcodes and counted the number of edges of the induced subgraph of $G$ with the selected nodes. We repeated this selection 20,000 times and proceeded with the run with the lowest edge count. 300 genes were randomly assigned to these barcodes.

### 3.6.2 Large-scale padlock probe production

Oligo pools are first amplified with probe set specific primers (Table S1, 0.3uM each) using KAPA HiFi following Twist Bioscience's protocol and purified to create 1st round amplicons. Then, the 1st round amplicon is PCR amplified on a 96-well plate (10pM template per reaction, 100ul reaction volume) using KAPA SYBR FAST and 1uM of each amplification primer until plateau. The PCR products were pooled and purified using QIAquick PCR purification kit (Qiagen 28106) on a vacuum manifold.

For the protocol in figure 3.4b, the purified amplicons were divided into parallel reactions (about 5ug each) and were digested with DpnII (1U/ul) in 1x NEBuffer DpnII (NEB R0543L) at 37°C for 3 hours and purified with QIAquick PCR purification kit. The purified products were digested with Lambda Exonuclease (0.5U/ul) in 1x buffer (NEB M0262L) for 2 hours and purified with Zymo ssDNA/RNA clean & concentrator kit. Finally, the library was digested with USER (0.0625U/ul, M5505L) in 1x NEBuffer DpnII in parallel reactions (about 2.5ug each) for 6 hours at 37°C followed by 3 hours at room temperature and purified with Zymo ssDNA/RNA clean & concentrator kit.

For the protocol in figure 3.4c, the purified amplicons were divided into parallel reactions of 50ul with 2ug of DNA template, and were digested with MboI (2.5U/ul) in 1x CutSmart buffer (NEB R0147M) at 37°C for 3 hours, heat inactivate for 20 minutes at 65°C. To the same tube, 2.5ul of Lambda Exonuclease and 2.5ul its reaction buffer (NEB M0262L) were added and

incubated at 37°C for 2 hours, then heat inactivated at 95°C for 20 minutes. Subsequently, each reaction was supplemented with 5.5ul of 0.5M EDTA and 5.5ul of USER (NEB M5505L), then incubated at 37°C for 9 hours and kept at 4°C overnight. Finally, the product was purified with Zymo ssDNA/RNA clean & concentrator kit.

### 3.6.3 DART-FISH rolony generation

#### 3.6.3.1 Reverse transcription and cDNA crosslinking

Tissue sections were fixed in 4% PFA in 1x PBS at 4°C for 1 hour, followed by two 3-minute washes with PBST (1x PBS and 0.1% Tween-20). Then, a series of 50%, 70%, 100%, and 100% ethanol were used to dehydrate the tissue sections at room temperature for 5 minutes each. Next, tissues were air dried for 5 minutes and in the meantime silicone isolators (Grace Bio-Labs, 664304) were attached around the tissue sections. Then, the tissue sections were permeabilized with 0.25% Triton X-100 in PBSR (1x PBS, 0.05U/ul Superase In, 0.2U/ul Enzymatics RNase Inhibitor) at room temperature for 10 minutes, followed by two chilled PBSTR (1x PBS, 0.1% Tween-20, 0.05U/ul Superase In, 0.2U/ul Enzymatics RNase Inhibitor) washes and a water wash. Next, the sections were digested with 0.01% pepsin in 0.1 N HCl (pre-warmed 37°C for 5 minutes) at 37°C for 90 seconds and washed with chilled PBSTR twice. Afterwards, acrydite-modified dT and N9 primers (Acr_dc7-AF488_dT20 and Acr_dc10-Cy5_N9, table S1) were mixed to a final concentration of 2.5 uM with the reverse-transcription mix (10U/uL SuperScript IV (SSIV) reverse transcriptase, 1x SSIV buffer, 250 uM dNTP, 40 uM aminoallyl-dUTP, 5 mM DTT, 0.05U/ul Superase In and 1U/uL Enzymatics RNase inhibitor). The sections with the mix were incubated at 4°C for 10 minutes and then transferred to a humidified 37°C oven for overnight incubation. After reverse transcription, tissue sections were washed with chilled PBSTR twice and incubated in 0.2 mg/mL Acryloyl-X, SE in 1x PBS at room temperature for 30 minutes. Then, the tissue sections were washed once with PBSTR, followed by incubation with 4% acrylamide solution (4% acrylamide/bis 37:1, 0.05U/uL Superase-In, and 0.2U/uL RNase inhibitor) at room temperature for 30 minutes. Subsequently, the acrylamide solution was

aspirated and gel polymerization solution (0.16% Ammonium persulfate and 0.2% TEMED in the 4% acrylamide solution) was added. Immediately, the tissues were covered with Gel Slick (Lonza #50640)-treated circular coverslips of 18 mm diameter (Ted Pella, 260369), transferred to an argon-filled chamber at room temperature and incubated for 30 minutes. After gel formation, the tissue sections were washed with 1x PBS twice and the coverslip was gently removed with a needle. At this point, the cDNA is crosslinked to the polyacrylamide gel.

#### 3.6.3.2 Padlock probe capture

After cDNA crosslinking in gel, remaining RNA was digested with RNase mix (0.25U/uL RNase H, 2.5% Invitrogen RNase cocktail mix, 1x RNase H buffer) at 37°C for 1 hour followed by two PBST washes, 3 minutes each. The padlock probe library was mixed with Ampligase buffer. Then, the mixture was heated to 85°C for 3 minutes and cooled on ice. Subsequently, the mixture was supplemented with 33.3U/uL Ampligase enzyme such that the final concentration of padlock probe library was 180 nM and 100 nM for the kidney and brain probe set, respectively, in 1x Ampligase buffer. Finally, the samples were incubated with probes at 37°C for 30 minutes, and then moved to a 55°C humidified oven for overnight incubation.

#### 3.6.3.3 RCA and rolony crosslinking

After padlock probe capture, the tissue sections were washed with 1x PBS three times, 3 minutes each and hybridized with RCA primer solution (0.5 uM rca_primer, 2x SSC, and 30% formamide) at 37°C for 1 hour. Then, the tissue sections were washed with 2x SSC twice and incubated with Phi29 polymerase solution (0.2 U/uL Phi29 polymerase, 1x Phi29 polymerase buffer, 0.02 mM aminoallyl-dUTP, 1 mg/mL BSA, and 0.25 mM dNTP) at 30°C in a humidified chamber for 7 hours. Afterwards, the tissue sections were washed with 1x PBS twice, 3 minutes each and the rolonies were crosslinked with 5 mM BS(PEG)9 in 1x PBS at room temperature for 1 hour. The crosslinking reaction was terminated with 1M Tris, pH 8.0 solution at room temperature for 30 minutes. Finally, samples were washed with 1x PBS twice and stored in a

4°C fridge until image acquisition.

### 3.6.4 DART-FISH image acquisition

The flow of image acquisition is as follows: anchor round, decoding rounds 1 to 7, DRAQ5 nuclear staining. All hybridizations were performed with 500nM of each of the fluorescent oligos in 2x SSC and 30% formamide for 15 minutes. Following hybridization, the unbound probes were washed with 4-5 washes with PBST each 2-3 minutes. Imaging was performed in PBST or 2xSSC. After each imaging round, stripping was performed with 80% formamide in 2x SSC and 0.1% Tween-20, 3 times each 3-5 minutes, followed by 2 quick washes with PBST to prepare for the next hybridization. In the anchor round imaging, all rolonies (through the universal sequence), N9 and dT RiboSoma (through the 5' handle on cDNA) are imaged. The probes DARTFISH_anchor_Cy3, dcProbe10_ATTO647N, and dcProbe7_AF488 are used for rolonies, N9 and dT, respectively. All the oligos are listed in table S1.

The imaging was performed on resonant-scanning confocal microscopes (Leica SP8 or Nikon AXR) using 20x immersion objectives (oil for Leica NA 0.75 and water for Nikon NA 0.95), with pinhole size of 2 airy units. Z-stacks covering 30-50um were taken with step size of about 2um with 2 or 3 times line averaging .

## 3.7 Acknowledgements

Chapter 3 is, in part, reprints of the material as it appears in Kalhor, K., Chen, C. J. ... & Zhang, K. (2024). Mapping human tissues with highly multiplexed RNA in situ hybridization. *Nature Communications*. The dissertation author was the primary investigator and co-first author of this paper.

The method described in section 3.4.2.2 was co-developed with Dr. Xuwen Li at Altos Labs.

# Chapter 4

# Developing an automated pipeline for processing multiplexed FISH data

In this chapter, I summarize my efforts to create a pipeline that can process the output of a DART-FISH experiment and create usable outputs for downstream analyses. I also touch on some efforts for optimal probe design for DART-FISH.

## 4.1 Motivation

The main readout for the multiplexed in situ hybridization techniques, including DART-FISH, is imaging. Because of the nature of combinatorial encoding, most the acquired images are not directly interpretable by humans. Rather, information about spatial gene expression needs to be computationally extracted from all of them simultaneously. This task, namely, computational extraction of gene expression spots from sequences of images is referred to as decoding.

Upon increasing the number of gene targets and sensitivity, multiplxed in situ hybridization techniques can be plagued with optical crowding. Crowding happens when the spots (e.g., rolonies) are too close that cannot be individually resolved by optical imaging. As such, decoding algorithms that rely on distinct spots will fall short and will lose sensitivity. Hence, creating decoding algorithms that are robust to optical crowding is a necessity for this growing field.

Aside from decoding, there are other challenges regarding data processing. A back-of-the-envelope calculation tells us the scale of imaging data that needs to be dealt with: 200FOVs

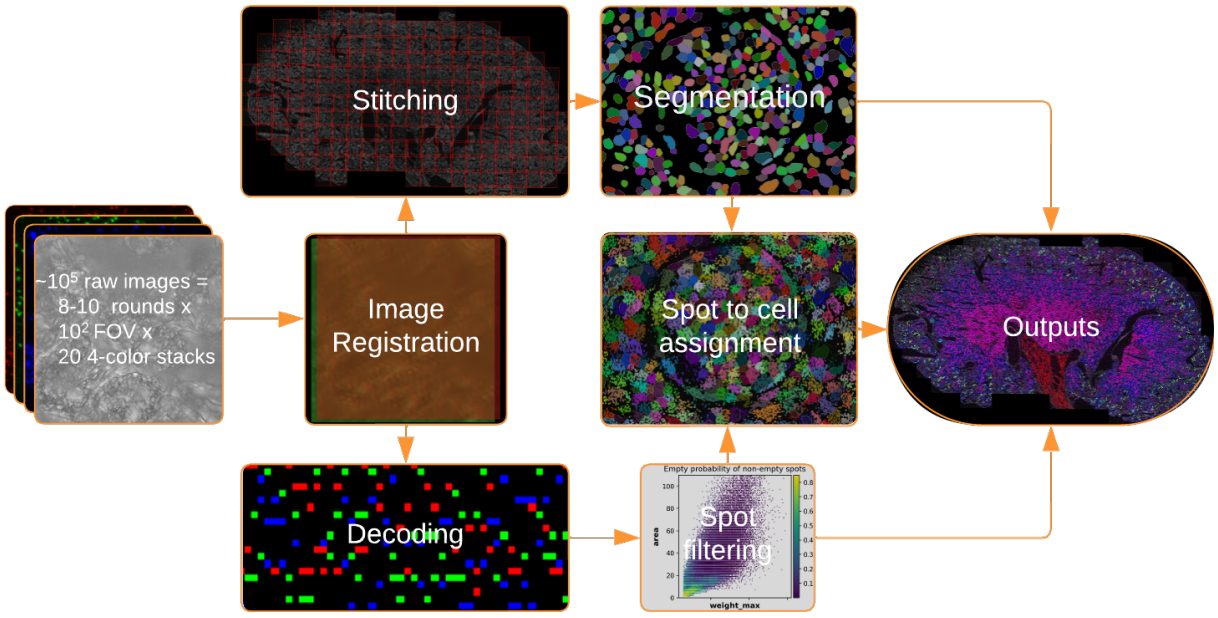



**Figure 4.1.** Schematics of the processing pipeline

x 8rounds x 20z-stacks x 4channels=128,000 raw images. Assuming that every image is 1024-by-1024 8-bit pixels, with no compression this amounts to 128 gigabytes of raw data. The large amount of data necessitates serious considerations regarding data handling, visualization and parallelized processing. For example, loading all the images of an FOV into the memory with an 8-bit structure occupies 640 megabytes of space. Though, most image processing algorithms, including a simple Gaussian filtering, require 64-bit data structures and thus about 5 gigabyte of memory. Hence, many of the processing steps in pipeline need to be memory conscious.

In the following sections, I describe the pipeline that I developed for processing DART-FISH images Figure 4.1. I created a decoding-by-deconvolution workflow that can extract spots from optically crowded areas. The pipeline is designed to be modular, so that whenever needed, modules could be modified or replaced by new ones. Moreover, I designed the pipeline to need little human oversight in different steps, with different checkpoints and outputs that can inform the quality of the data and the processing.

## 4.2 Image registration

The first step of the pipeline is image registration or image alignment. In order to decode, we need to trace the color of every rolony across multiple rounds. This requires all the images from different rounds to be aligned such that the pixel coordinates match the physical coordinates. In other words, we need pixel $(i, j)$ in all images to specify the same $(x, y)$ physical coordinate. However, raw images taken at different cycles tend to be slightly shifted, mainly in a non-predictable fashion. The reasons for this could be slight sample moving during between-cycle preparations or non-reproducible stage positions. Consequently, the raw images need to be computationally aligned.

Image registration between a "moving" image and a "fixed" image can be formulated as the problem of finding a mathematical transformation that when applied to the fixed image, the result closely resembles the moving image. The transformation can be parameterized in different ways; it could be a translation transformation with only 3 parameters (2 for 2D), or other more complex linear transformations such as rigid or affine; it could also be a complex non-linear and local B-spline. For our purposes, a translation transformation is enough for most cases, and for rare situations an affine transform suffices. As for the implementation, the pipeline uses SimpleElastix [59], which wrap powerful image processing libraries in Python.

To perform image registration, one needs a frame of reference with features that are generally conserved across different rounds. Fluorescent images with signals from rolonies are not suitable for this task, as their signal varies with rounds: rolonies change color or turn off. To address this, we collect transmitted light (brightfield, BF) images along with the fluorescent images. Because BF mainly captures the tissue structure, its features are conserved across different rounds. Hence, we select an arbitrary round as the reference, and then find the transform parameters for every other round by registering their BF image to the BF images of the reference. The transforms are then applied to all fluorescent channels to bring all channels in the same coordinate system.

## 4.3 Decoding: From images to transcripts

At the heart of the pipeline is the decoding algorithm. A decoding algorithm takes two main inputs: 1) fluorescent images from decoding rounds, 2) a codebook which contains the barcode for each gene (figure 3.1c). Every line in the codebook has this format: genename_barcode. For example, the entry SLC17A7_132000 means that rolonies from *SLC17A7* have value of "1" in round 1, "3" in round 2, "2" in round 3 and off in rounds 4 to 6. "1" corresponds to fluorescent probes that are labeled with Alexa Fluor 488, "2" corresponds to Cy3, and "3" to ATTO647N dyes (see table S1). Figure 4.2 shows an example raw data for DART-FISH decoding. In this figure, values of "1", "2", and "3" are represented by red, green, and blue colors, respectively. As seen in the figure, rolonies show up as bright round spots, and change colors across different cycles without moving.

Traditionally, there have been two main ways of looking at the decoding problem: spot-based decoding and pixel-based decoding [60, 61]. In spot-based methods, the first step is to identify the location of rolonies, then assign them to barcodes/genes. In pixel-based methods, every pixel is first assigned to a barcode/gene and then neighboring pixels are pooled together to identify rolonies. As I discuss below, in practice each approach comes with some pros and cons.

Pixel-based methods are more affected by imaging noise since their core operations happen at the pixel-level, while spot-based methods tend to be more robust to these sources of noise. On the other hand, the main hurdle for spot-based methods is the reliable detection of the spots. There are several methods for spot detection, some are based on the classic Laplacian of Gaussians (LoG) filtering [62], or more modern algorithms [63, 64]. Nevertheless, these methods are designed to detect spots from a single stack of images from one channel, and their reliability does not extend to more complex applications as in multiplexed hybridization, where the spots need to be tracked across multiple rounds and they may change color or may turn off altogether. For this reason, I focused my efforts on pixel-based decoders.

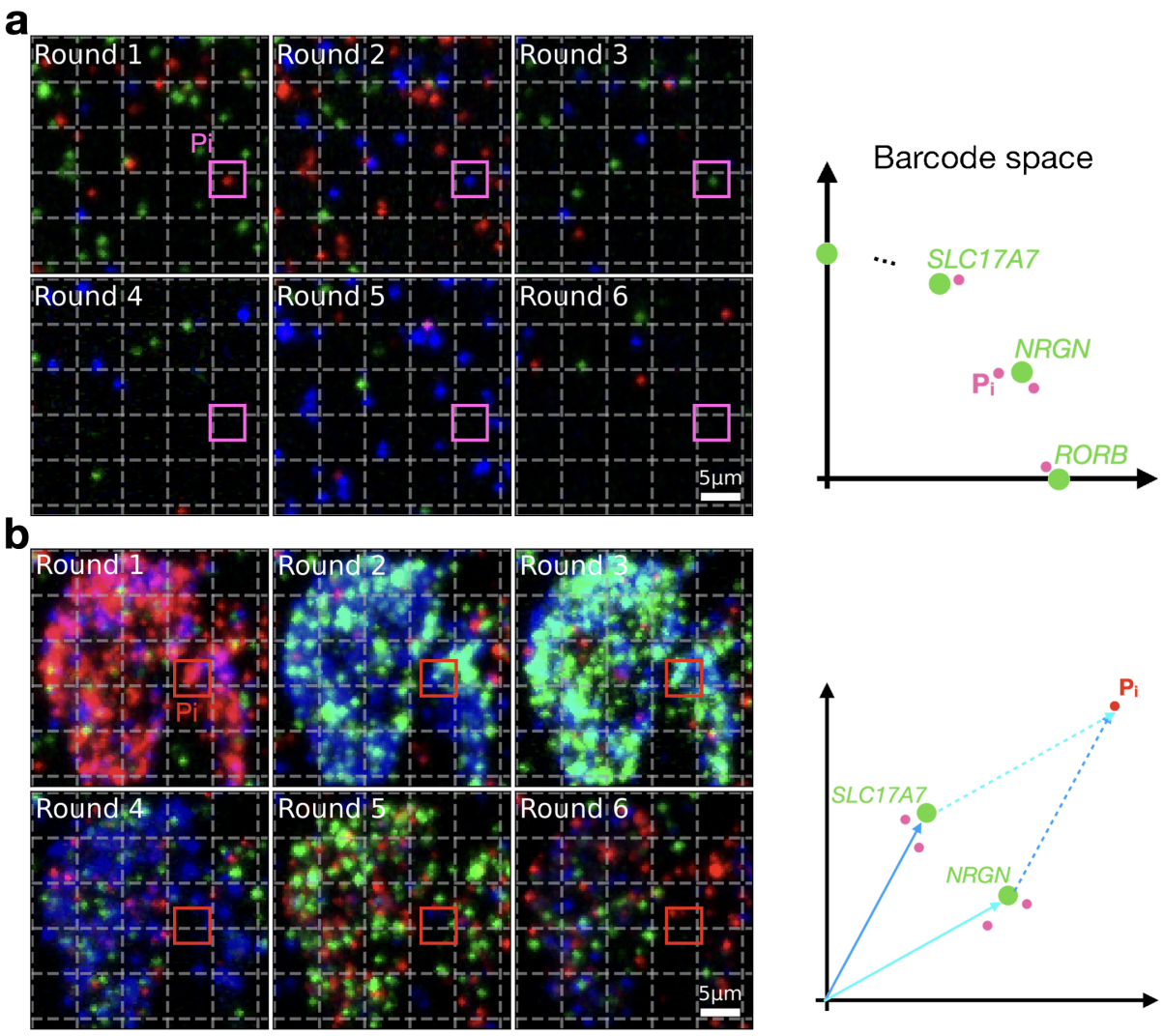



**Figure 4.2. a.** (left) Raw decoding images for the portion of a FOV with the codebook of figure 3.1c. (right) conceptual drawing of the barcode space and direct matching of pixels to barcodes. Barcodes are large nodes in green, the example pixel, $P_i$ in pink. **b.** Same as in a, in a FOV with dense and overlapping signal. Mixed signals are signified by mixed colors (e.g., magenta and cyan). Right panel shows how $P_i$ shows mixed signal from two distinct barcodes

### 4.3.1 Decoding by direct matching

The naive method for pixel-based decoding is direct matching (Figure 4.2a). That is, comparing the profile of a pixel directly with every barcode in the codebook and assigning the pixel to the barcode with highest level of similarity. To perform direct matching, two key measures need to be defined: 1) What we mean by the profile of a pixel, 2) what similarity

metric to use. To define the pixel profiles, many groups threshold the intensity values for every rounds and channel to convert them into a vector of binary values [19, 50]. This is followed by using Hamming distance as a similarity metric. To avoid setting threshold values, one could skip binarization and instead use normalized intensity values. In the widely used starfish package [60], the intensity vectors (real-valued vectors between 0 and 1, Methods) are compared to the barcodes (vectors consisting of 0s and 1s) using Euclidean distance to find the closest barcode to every pixel. Figure 4.2a gives an example of how this matching is done for a given pixel.

While useful for early generations of in situ transcriptomics data, this strategy is fundamentally limited to cases where spots are distinct and non-overlapping, i.e., sparse datasets. Figure 4.2b shows an example from a dense dataset in which these assumptions do not hold. In such cases, the intensity profile of a pixel can be a combination of multiple barcodes, and thus its direct comparison fails to identify the underlying barcodes.

### 4.3.2 Decoding by deconvolution

Targeting more genes with high sensitivity can result in optical overcrowding which may hinder rolony decoding. Physical expansion of the tissues [19, 52, 54, 65] has been used as an effective strategy to distance rolonies and reduce overcrowding but it leads to larger imaging areas, longer imaging time and thus lower throughput. A computational solution to the overcrowding problem can vastly increase the throughput. We reasoned that given the size of the rolonies (<1um)[51] and our pixel size ( 0.3um with 20x objective), each pixel will at most overlap a few rolonies. On the other hand, given that a small fraction of all possible barcodes are used, it may be possible to deconvolve mixtures of barcodes from fluorescent intensity values at the pixel level. To this end I developed the SparseDeconvolution (SpD) decoding algorithm: I formalized this deconvolution as a linear regression problem, where barcodes can combine linearly to form the observed pixel intensities (4.1). Because such a linear regression problem under-specified, it can have infinite solutions. Since the pixel intensities are generated by a limited number of barcodes, I regularized this regression problem under conditions that promote

sparsity (Methods, 4.2), that is, the solutions consist mainly of zeros.

$$X\mathbf{w} = \begin{bmatrix} | & | & | & | \\ x_1 & x_2 & \cdots & x_N \\ | & | & | & | \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}_{pix_i} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n*3} \end{bmatrix}_{pix_i} = \mathbf{y} \tag{4.1}$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left( ||\mathbf{y} - X\mathbf{w}||_2^2 + \alpha||\mathbf{w}||_1 + \alpha'||\mathbf{w}||_2^2) \right) \tag{4.2}$$

We solve this problem for every pixel and obtain initial weight maps for every single barcode (Figure 4.3a&b). This is followed by a filtering step in which an elbow filter is applied to the solution of every pixel. The elbow filter selects the top one or two largest features only if they are significantly larger than the rest of the non-zero features. In cases that the top one or two features do not satisfy this criterion, no feature is selected. Next, an ordinary (non-regularized) linear regression problem is solved using only the selected features to obtain unbiased weight maps. Next, neighboring pixels are aggregated and segmented to form spots (Figure 4.3c&d, Methods).

#### 4.3.2.1 Quality control

To control the quality of the deconvolution procedure, we extracted several features from every spot. These features include maximum weight and area, as well as shape descriptors including eccentricity and solidity. We also included empty barcodes in the codebook. Empty barcodes are those that were not used the probe set and are not going to produce rolonies, however, during decoding they are treated similarly to other barcodes. Any spot that is decoded to have an empty barcode must be an error occurred during decoding. To control the quality of the decoding, we use the features extracted from empty barcodes to remove other decoded spots with similar features (Methods). Empty rate, defined as the fraction of spots decoded as empty, should be kept lower than the fraction of empty barcodes in the codebook.

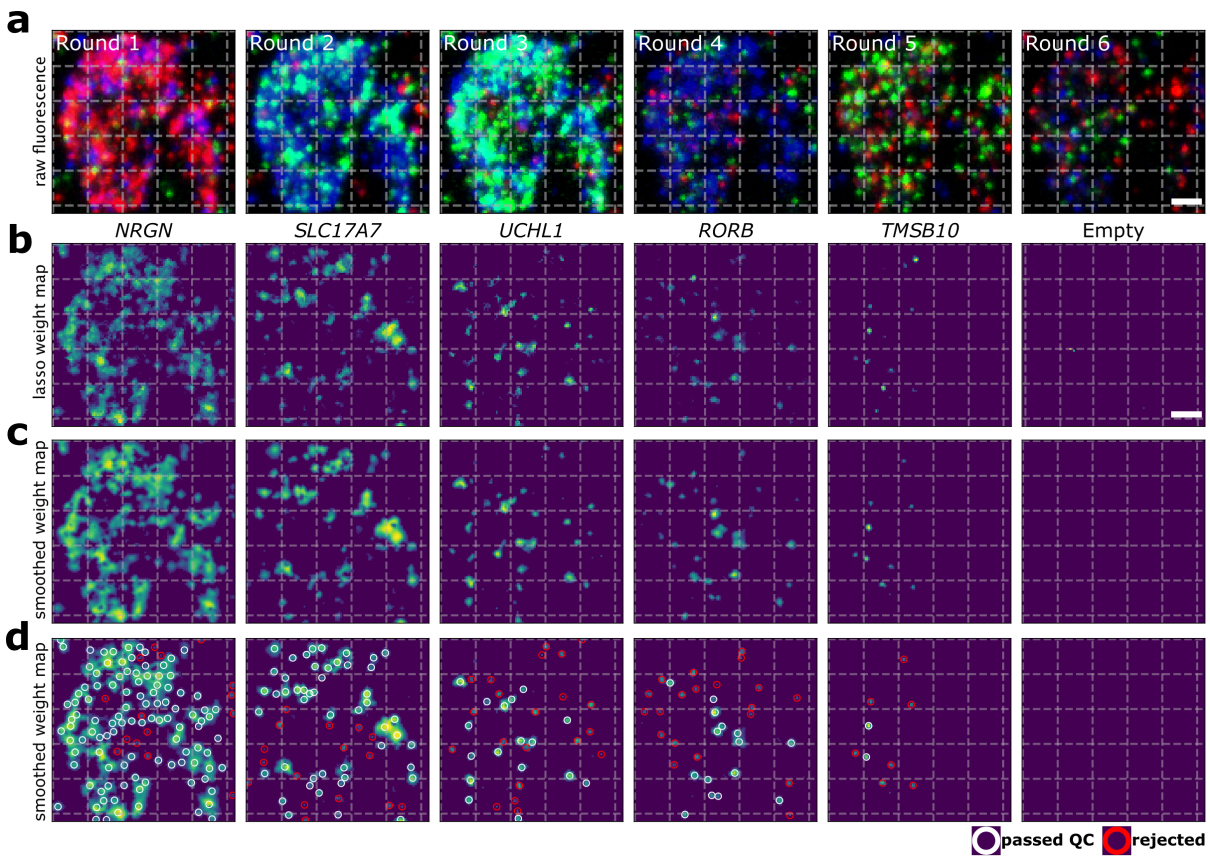



**Figure 4.3. a.** Example of decoding by FISH on the PA gel. The lower panel shows the maximum intensity projection of the fluorescent images across 6 decoding rounds and 3 channels (scale bar 5 um). **b.** Lasso maps. Lasso maps are the solutions to the optimization in 4.2 and represent the gene weights for each of *NRGN*, *SLC17A7*, *UCHL1*, *RORB*, *TMSB10*, and Empty barcodes in (a) (scale bar 5 um) **c.** Smoothed OLS maps for panel (a). Lasso weight maps (panel b) undergo pixel-wise elbow filtering to select the top 1 or 2 barcodes per pixel. Unbiased weights are then obtained by fitting an ordinary linear regression (OLS) using the selected barcodes (OLS maps). OLS maps are then smoothed with a Gaussian low pass filter. **d.** Spot detection on weight maps. For each gene, the local peaks are detected on the respective smoothed OLS map. These peaks then serve as markers for watershed segmentation. The centroids of the segmented areas are used as spot coordinates. White and red circles are drawn around high quality and rejected spots, respectively.

#### 4.3.2.2 Benchmarking

To put the performance of SpD into perspective, I performed a simulation and applied SpD and other decoding methods side-by-side. The advantage of simulated data compared to real data is that we can fully control the data quality and have complete ground truth (Methods).

The algorithms in the comparison included a naive algorithm that directly matches pixels to individual barcodes [60] and more sophisticated deconvolution algorithms [66, 67]. The results on synthetic data show that the performance of the direct matching quickly plunges at higher spot densities while deconvolution algorithms are more robust(Figure S2d). The deconvolution algorithms are harder to compare, as I used default parameters for all of them (including SpD) with no post-processing. Keeping these in mind, for a range of densities, SpD and ISTDECO show complementary performances, as ISTDECO is more sensitive while SpD is more specific (Figure S2d). Note that, unlike SpD, ISTDECO does not account for intensity variation between channels and is negatively affected by this common phenomenon.

Another interesting observation is that specificity, which is unobserved on real data, is related to empty rate. Assuming that decoding errors are uniformly distributed among barcodes one could estimate this relationship between specificity (SP, fraction of truly correct calls over all calls) and empty rate (ER, fraction of empty calls over all calls):

$$SP = 1 - ER/\lambda \tag{4.3}$$

where $\lambda$ is the fraction of empty barcodes in the codebook. Figure S3 shows that this estimation works well for a wide range of spot densities for various decoding methods. Consequently, one can keep specificity high by keeping the empty rate low. For example, the data shown in figure 4.3 has an empty rate of 0.25%; with 10 empty barcodes and 131 total barcodes, the estimated specificity is 0.967%.

In summary, I created a new decoding method, SpD, that addresses an increasingly important bottleneck in in situ hybridization techniques, that is optical crowding. With this capability, we could use lower magnification objectives to increase the throughput of imaging, while still being able decode transcript signals from mixed pixels.

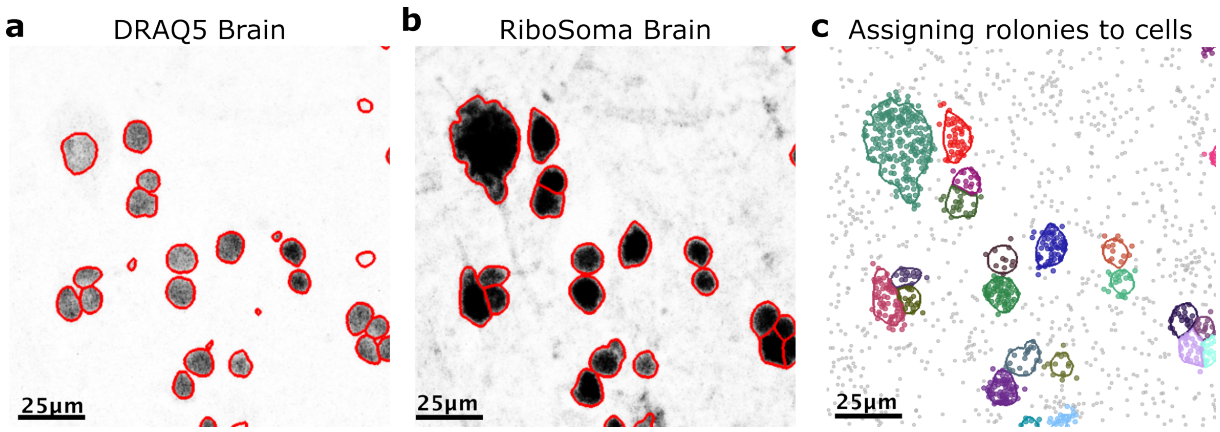



**Figure 4.4. a.** DRAQ5 nuclear staining in a human brain sample. Red outlines are segmented boundaries by cellpose. [69]. **b.** RiboSoma segmentation of the same FOV as in (a). **c.** Assigning decoded rolonies to the closest segmented cell. Transcripts that are too far from cell boundaries are discarded.

## 4.4 Cell segmentation and transcript assignment

To create gene expression profiles for the cells, transcripts need to be assigned to the cells they originated. This is typically achieved through performing cell segmentation followed by assigning transcripts to the closest cell. Misassignment of transcripts to cells will negatively impact downstream analysis, for example by complicating the cell annotation step.

In the field, it is typical to use a nuclear stain to perform cell segmentation [49]. The reason is primarily that there are non-hazardous fluorescent DNA dyes like DAPI (4',6-diamidino-2-phenylindole) that can easily be incorporated into in situ workflows. Moreover, in many cases nuclei have shapes that are simpler to segment (Figure 4.4a). Transcripts that fall inside segmented nuclei can be confidently assigned their cell of origin. For transcripts outside the nuclei, however, there is more uncertainty. A strategy widely used is to naively estimate the location of the cytoplasm by expanding the nuclear boundary by a constant amount, and discarding all transcripts that do not fall into the expanded regions [68]. Unfortunately, in most cases this strategy is inadequate. A large expansion radius is too large leads to leakage between cells, and a short radius results in the loss of many transcripts. Furthermore, there is no clear way for setting the radius, and one radius may not fit all cell types within a cell.

To mitigate this issue, I proposed to use RiboSoma for cell segmentation. RiboSoma is created by reverse-transcribing the total RNA of the cell. Since RNA is found in both nucleas and cytoplasm, RiboSoma can act as a cytoplasmic stain. In order to use RiboSoma, I fine-tuned a leading cell segmentation algorithm on composite images of RiboSoma and DRAQ5 nuclear stain [69, 70]. Figure 4.4b shows this new segmentation on a human brain FOV. Compared to the nuclear segmentation in figure 4.4a, RiboSoma confidently labeled a larger area and significantly improves cell segmentation.

Finally, we used this cell segmentation for assigning transcripts to cells. Figure 4.4c, shows the abundance of transcripts outside of the cells (gray dots) and how much better RiboSoma is informing this assignment compared to nuclear segmentation. In the data set shown in figure 4.4, my new segmentation increased the number of transcripts confidently assigned by 20%. Note that, this improvement is not uniform across all cells; rather, cell that have a larger cytoplasm to nuclear ratio benefit more.

## 4.5 Computational design of codebooks

As discussed before, in DART-FISH genes are represented by barcodes. The mapping between the genes and barcodes are reflected in the codebook. For example, in the codebook used in section 5.2, the barcode for gene *SLC12A1* is 0010230. The typical criteria for finding a set of barcodes are as follows: 1) each barcode is assigned to only one gene, 2) barcodes have as large a Hamming distance possible, 3) the barcodes represent all channels equally. In section 3.6 I described a heuristic algorithm that generates barcodes that satisfy these criteria. However, in practice, I have learned of some undesired cases that need to need further consideration in barcode design.

Let's start with an example. Consider the genes *SLC12A1* and *UMOD*. They are among the highest expressing genes in medulla of kidney. In fact, these genes are highly specific marker genes for the TAL (thick ascending limb) segment in the kidney. Since rolonies occupy physical

a

| Genes | Barcode |
|---|---|
| *SLC12A1* | x1=0010230 |
| *UMOD* | x2=0120002 |
| *BOC* | x3=0010032 |
| *EMCN* | x4=0120200 |
| *SLC12A1+UMOD* | x1+x2=01(1,2)023 |
| *BOC+EMCN* | x3+x4=01(1,2)023 |

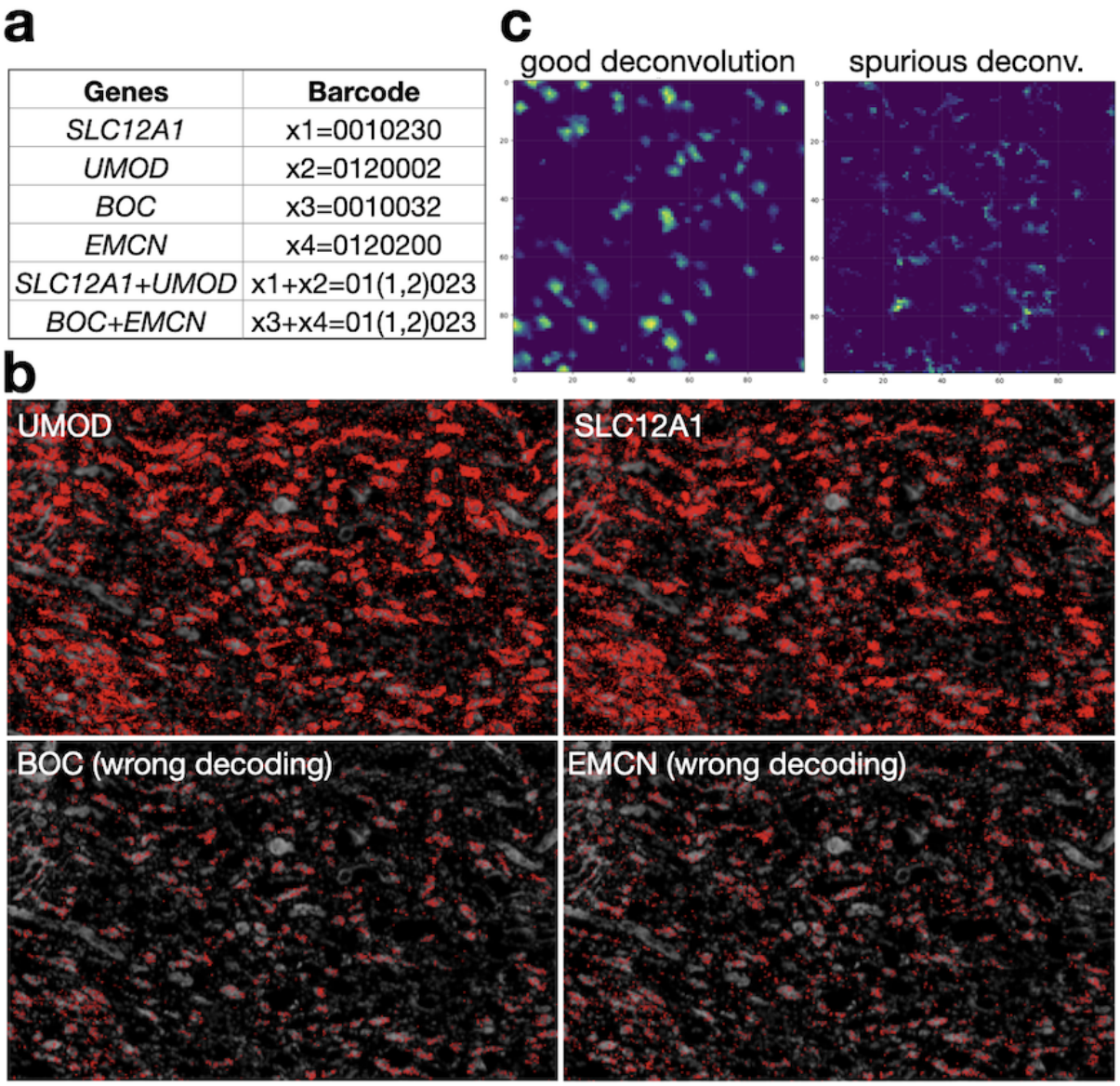



**Figure 4.5. a.** Table showing the barcodes of 4 genes. Linear combination of *SLC12A1* and *UMOD* equals to the combination of *BOC* and *EMCN* **b.** An example of decoded spots of the four genes in (a). *BOC* and *EMCN* are not expected to co-express with *SLC12A1* and *UMOD*, but the signal "leaks" from the top genes into the barcodes of the bottom two genes. **c.** Example of deconvolution weights. Left plot shows the weights coming from an underlying gene (real signal). Right plot shows an artifactual signal caused by non-unique combinations of multiple genes. The spurious signal tends to have irregular and disordered shapes.

space and these genes are highly expressed in the same cells, we expect a substantial overlap between the rolonies of these genes. Even though our SpD decoder 3.2.2 can extract barcodes from mixed pixels, it still has limitations.

Figure 4.5 shows one of these limitations. Panel (a) shows the barcodes for *SLC12A1* and *UMOD* as well as two other genes *EMCN* and *BOC* that are not expressed in TALs. As evident in panel (a), the linear mixture between *SLC12A1* and *UMOD* is identical to that of *EMCN* and *BOC*. Consequently, SpD may misidentify *SLC12A1-UMOD* mixtures as *EMCN-BOC* mixtures. I call this phenomenon barcode collision. Despite having strict filtering of the deconvolution weights, a small fraction of spot calls can still bear this mistake (figure 4.5b). This is a fundamental limitation of deconvolution the way we are performing it.

There are ways to mitigate this issue. For example, I have found that such spots tend to have irregular shape (figure 4.5c). Once the spurious spots are identified, a small subset can be manually selected to train a classifier on their shape descriptors. The classifier can then be a new level of filtering that is applied to all spots. This process works well however, it requires the investigator 1) to learn about the existence of the problem in the first place, 2) manually select training data for the classifier.

### 4.5.1 A cost function to evaluate codebooks

A more ideal solution to the collision issue would be to prevent it from happening altogether. One way is to assign the barcodes such that those genes that are co-expressed at high levels will have barcodes that have a unique mixture. That is, no other pair of barcodes will have that same mixture. In other words, not only do we distance barcodes from each other, but also do we distance the mixture of barcode pairs.

To do so, I created a cost function that takes a codebook as an input and outputs a real number that is higher when highly co-expressed genes have barcode collisions. The equation below shows this function more intuitively:

$$cost = \sum_{g_1,g_2} \text{Connectivity}\big(CB(g_1)+CB(g_2)\big) * \text{Cooccurrence}(g_1,g_2) \tag{4.4}$$

where $g_i$ is a gene, $CB$ is the assignment between genes and barcodes in the codebook. There

are two other important functions: "Connectivity" takes the barcode mixture of two genes and calculates the degree of collision given the assignment. "Cooccurrence" takes two genes and outputs a number between 0 and 1 showing the degree of co-expression between them.

The connectivity aspect can be computed through what I call a combination graph (combo graph). In the combo graph, every pair of barcodes (or genes) form a node. The value for nodes is the sum of their barcodes. The nodes are connected with $h_0$ edges (red) if with their values are identical and connected with $h_1$ edges (black) if the Hamming distance of their values is 1 (Figure 4.6a&b). It is up to the user how to define the connectivity value for every node (Connectivity$\big(CB(g_1)+CB(g_2)\big)$ in equation 4.4). In different runs, I have defined it as the number of $h_0$ edges or a combination of $h_0$ and $h_1$.

The co-occurrence score for gene pairs can be estimated from single-cell sequencing data. To create a scoring system, one needs to keep in mind that our concept of co-occurrence is different from correlation. Two genes may be highly correlated but expressed at low levels and thus less problematic. Hence, co-occurrence should also take into account the expression level of each gene. Furthermore, many genes have heterogeneous expression across the population, and we should be concerned with worst case scenarios for co-expression of two genes.

For the co-occurrence scoring, I first showed that if rolonies have the area $s$, and their are randomly dispersed in a plane with density $\lambda$, then the fraction of area that is covered by at least one rolony is $1-e^{-s\lambda}$. Similarly, the fraction of area in which two different types of rolonies with densities $\lambda_1$ and $\lambda_2$ overlap is $(1-e^{-s\lambda_1})(1-e^{-s\lambda_2})$. To actually calculate this score for pairs of genes, I estimated $\lambda_1$ and $\lambda_2$ from single cells, and calculated this score across all single cells and took the average of the top 10. Note that, the area $s$ may change the quantities but does not alter the order across all pairs of genes.

### 4.5.2 A heuristic algorithm for codebook optimization

Having defined the connectivity and co-occurrence, we need to find a way to minimize the cost function in equation 4.4. The minimum for this cost function does not have a closed-

form solution and I could not find an optimization algorithm with theoretical guarantees for convergence to global minima. Hence, I came up with a heuristic algorithm that reduces the cost in a greedy and iterative manner. The steps of the algorithm are as follows:

1. Randomly assign barcodes to genes
2. Calculate the cost function. Upon this calculation, create a list of gene pairs that contribute the most to the total cost
3. Take the gene pair $(g_1, g_2)$ that contributes the most to the cost. Randomly switch the barcode of $g_1$ or $g_2$ with another gene. Calculate the cost. If it is reduced, go to Step 2. If not, repeat this step. If all possible combinations are tested go to next step.
4. Repeat step 3 with the next gene pair in the list. If all gene pairs are tested and no reduction is achieved, stop.

Figure 4.6c shows the evolution of the cost function across 175 iterations of the algorithm. A visual output of the optimization can be seen in figure 4.6d in which a subset of a combo graph of an optimized codebook is depicted. Compared to a random codebook (figure 4.6b), there is a clear reduction in the number of edges. More quantitatively, we could look at $h_0$ and $h_1$ edges in the combo graphs for a given co-occurrence score. Figure 4.6e&f show this for a random assignment and 6 different runs of the codebook optimization algorithm. I can be seen that the algorithm significantly reduced the number of $h_0$ edges for high co-occurring gene pairs. As expected, a reduction of $h_1$ edges is seen only in runs that included the number of $h_1$ in their connectivity value.

### 4.5.3 Performance on synthetic data

To show that this codebook optimization method can yield positive results for decoding, I created a simulated spatial dateset consisting of 1000 cells with cell types taken a kidney atlas. I created multiple datasets and varied the number of molecules per cell from 10 to 150. I then create synthetic images from these datasets using random and optimized codebooks. Having the

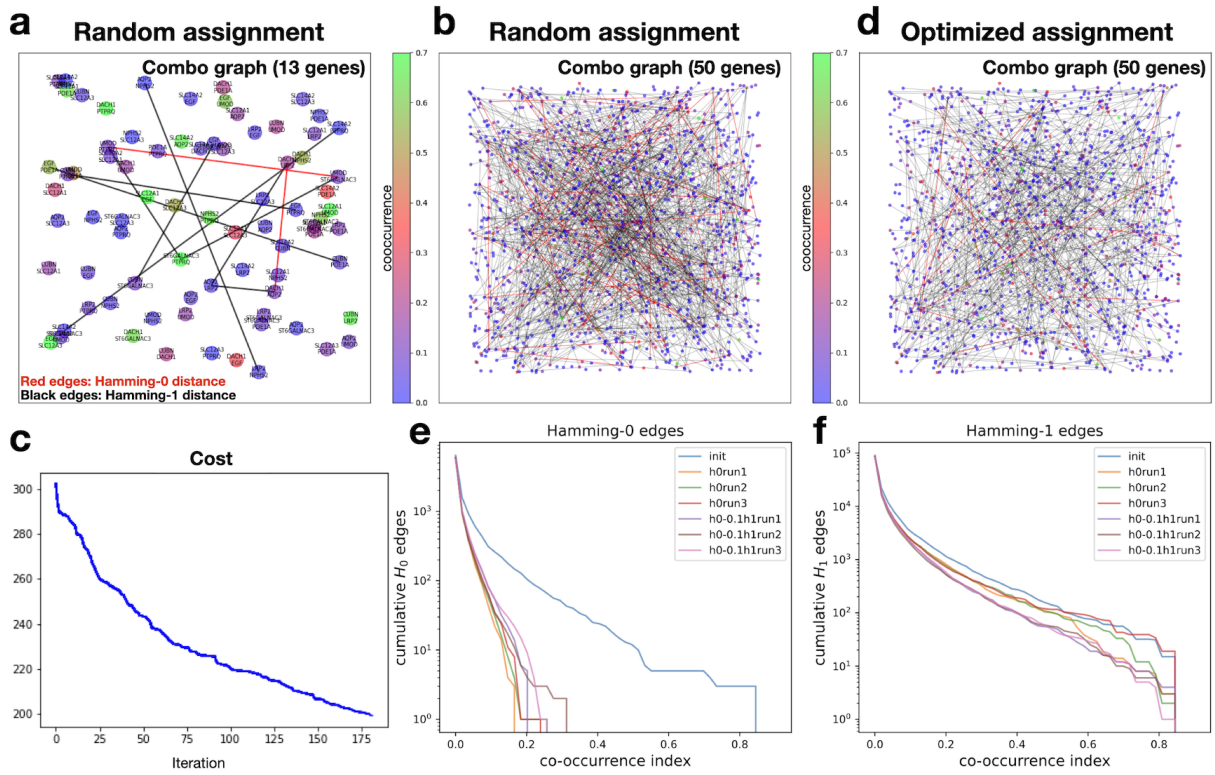



**Figure 4.6. a.** A zoomed-in view of a combo graph of a random codebook. The node labels denote the pair of genes, colored by their co-occurrence score. Black and red edges depict the $h_0$ and $h_1$ edges, respectively. The full combo graph contained 190 genes (about 18,000 nodes), but was subsetted to show the 78 nodes from 13 random genes. **b.** Same combo graph as in (a) but with a 50 gene subset (1225 nodes) **c.** The proposed optimization procedure reduces the cost function of the assignment problem. **d.** The optimized combo graph of the genes in (a) and (b). **e-f.** The number of $h_0$ (e) and $h_1$ (f) edges as a function of the co-occurrence score or index. For a co-occurrence value $CI$, the curves show the count of edges connected to gene pairs with co-occurrence index $> CI$. The connectivity function for curves labeled "h0run" only counts the number of $h_0$ edges, while for curves labeled "h0-0.1h1run", it uses this formula: $(\#h_0) + (\#h_1)/10$.

ground truth and the decoding results allows me to inspect the performance of the optimized codebooks as a whole or at the level of individual genes.

Figure 4.7a shows the total performance the codebooks across a range of spot densities. The performance is measured in terms of specificity, which is calculated as the number of correctly decoded spots divided by the number of all decoded spots. It is clear that the optimization lead to a positive total performance with an increasing effect at higher densities of spots. Overall, the average increase in sensitivity per gene was about 0.0036% (for 69 spots per cell which is a

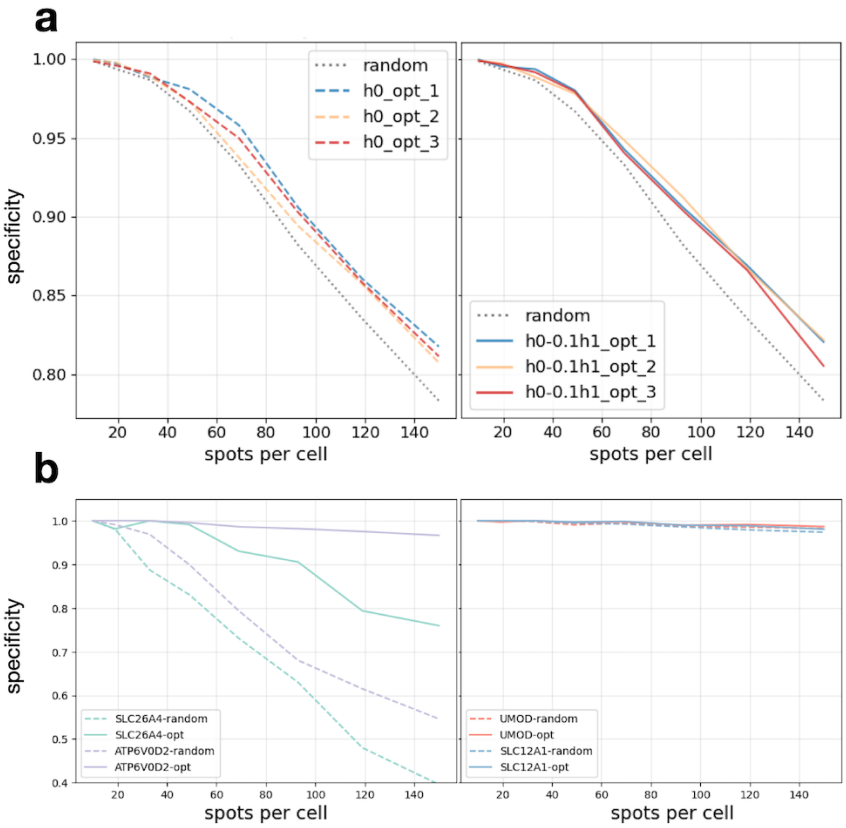



**Figure 4.7. a.** The specificity of decoded spots in synthetic datasets with varying spot densities. Left plot shows the performance of a random codebook and three optimized ones with $h_0$-only connectivity. The optimized codebooks in the right plot used both $h_0$ and $h_1$ in their connectivity. **b.** Same as (a) but focused on the performance of individual genes. These genes are involved in this collision in the random codebook: *SLC12A1+UMOD=SLC26A4+ATP6V0D2*

realistic number for DART-FISH). However, the average increase per gene weighted by the true frequency of genes is 3.625%. This is because the algorithm emphasizes more on the genes with higher expression since those genes will have a higher co-occurrence scores.

To see if I was successful in resolving cases that sparked the idea for this journey (see figure 4.5), I looked at collision of *SLC12A1* and *UMOD* again. In this example, the barcode collision happens between the mixture of these genes and *SLC26A4* and *ATP6V0D2*. Figure 4.7b shows that my algorithm can increase the specificity in detection of *SLC26A4* and *ATP6V0D2* up to 30%.

In summary, I created an algorithm for designing codebooks. My algorithm uses the

expression in single-cell datasets and assigns barcodes to genes to avoid barcode collisions. Using realistic simulations, I showed that my method outperforms the commonly used random assignment and increases the decoding sensitivity. Future work should move beyond pairwise collisions and consider higher order combinations of barcodes upon scoring codebooks.

## 4.6 Discussion

RCA-based in situ detection systems are prone to optical and physical overcrowding as more and more genes are detected with higher efficiency. To mitigate this issue, I developed a computational method (SpD) that used the redundancy in the barcode space to deconvolve mixed barcodes from single pixels. This strategy improved our decoding efficiency compared to naive decoding methods[60]. The utility of this method increases with higher redundancy in the barcode space by creating longer barcodes with more "on" cycles, and careful assignment of barcodes to genes such that genes that tend to co-express in the same cell types have unique barcode combinations. In addition, more sophisticated deconvolution methods that share information between neighboring pixels can potentially improve decoding efficiency[67, 66]. As the field is moving towards detecting more genes in parallel, pixel-based deconvolution methods like SpD could become increasingly relevant.

Upon developing SpD, I investigated its failure modes and showed that barcode collisions can become a challenging issue when dealing with highly expressed genes. Thus, I design a computational codebook optimization workflow that minimizes the chance of collision for pairs of genes. While I only obtained only a modest improvement in the specificity of the optimized probe sets (about 3.5%), I showed that the design of the codebook, even though often overlooked, is important and can have significant impact on the downstream results. In the future, more complex algorithms that encompass not only pairwise barcode combinations, but also higher order scenarios should be developed to utilize the redundancy in the barcode space to maximize the specificity of the decoding. Along with using multiple barcodes for each gene, it should be

possible to eliminate the barcode collision issue.

## 4.7 Methods

The DART-FISH datasets were processed by our custom pipeline. The source codes of the pipeline can be found in this Github page (https://github.com/Kiiaan/DF3D). Raw z-stack images with 4 channels (3 fluorescent channels and brightfield) from the microscope were registered to a reference round by affine transformation implemented in SimpleElastix[59] using the brightfield channel as the anchor. Then, each field of view (FOV) underwent decoding to obtain a list of candidate spots. Spots from all FOVs were pooled and filtered (See Sparse deconvolution (SpD) decoder for more details). To obtain the global position of the rolonies, the FOVs were stitched by applying FIJI's Grid/Collection Stitching plugin [71] (in headless mode) to the registered and maximum-projected brightfield images. Note that the theoretical positions of the FOVs, defined by the microscope, were used as initial positions for stitching.

Cell boundaries were segmented with Cellpose (v2.1.1) [69, 70]. The "cyto" model in Cellpose was fine tuned on each tissue by manually segmenting a handful of composite images of DRAQ5 (nuclei channel) and N9 cDNA stain (cyto channel) using the package's graphical user interface.

### 4.7.1 Details on sparse deconvolution (SpD) decoder

In DART-FISH, each gene is represented by a barcode that can be read out in $n$ rounds of 3-channel imaging. Each barcode is designed to emit fluorescence (be "on") in exactly $k$ rounds, each time in a single fluorescent channel and stay "off" in other rounds. We concatenate the rounds and channels and represent the barcodes as $3n$-dimensional vectors. In other words, barcode $i$ is represented by vector $x_i$ in which 1's are placed where "on" signal is expected, and 0's everywhere else. The codebook matrix X ($3n$x$N$) is then defined as $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, where $N$ is the total number of barcodes. In the same way, for every pixel we concatenate the fluorescent intensity values (scaled between 0 and 1) to create a $3n$-dimensional vector $\mathbf{y}$.

The fluorescence signal at each pixel can be sourced from more than one rolony if the

distance between neighboring rolonies is smaller than the optical resolution of the imaging system, or if 3-dimensional stacks are analyzed as maximum-projected 2D images. Nevertheless, because of physical constraints, only a handful of rolonies are expected to be the source of signal to each pixel. In this regard, because of the redundancy in the barcode space, combinations of barcodes in one pixel can be decomposed into their original composing barcodes. We formulated this problem as a regularized linear regression problem where a weighted sum of a few barcodes creates the observed signal intensity, where the vector $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$ shows the contribution of each barcode (equations 4.1 and 4.2) with most $w_s$ ($1 \leq s \leq N$) elements equal to 0. We initially used lasso to solve this problem ($\alpha'$=0 in equation 4.2) to promote the sparsity of $\mathbf{w}$, but later decided to use elastic net ([72] with a non-zero value for $\alpha'$ that is much smaller than $\alpha$ ($\alpha' = \alpha/100$) to increase stability. We call the solution to this problem $\hat{\mathbf{w}}_{lasso}$. Note that, we constrain the problem to positive weight values ($\hat{\mathbf{w}}_{lasso_s} \geq 0$ for every $s$). The regression problems are solved for all the foreground pixels ($||\mathbf{y}||_2 > 0.25$) individually. For every barcode $i$, we can construct an image with the estimated weight values as pixels: 0 for background and rejected pixels, and non-zero values from $\hat{\mathbf{w}}$. We call these images weight maps. Figure 4.3b shows weight maps constructed with $\hat{\mathbf{w}}_{lasso}$ which have not been filtered.

With our current barcode space, we can only confidently decompose bi-combinations. Hence, for every instance of the elastic net problem, we applied an elbow filter and accepted the solution only when the top one or two weights were significantly larger than other weights.

In more detail, for every pixel, the weights in $\hat{\mathbf{w}}_{lasso}$ are sorted in decreasing order. If the second largest weight is smaller than half of the top weight, then the top weight passes the elbow filter. Otherwise, if the third largest weight is smaller than 30% of the largest weight, the top two weights pass the elbow filter. All the values that do not pass the filter are set to zero. For accepted solutions, we performed an ordinary least square (OLS) regression using the top one or two weights to obtain unbiased weights ($\hat{\mathbf{w}}_{OLS}$). Figure 4.3c shows weight maps constructed with $\hat{\mathbf{w}}_{OLS}$ (OLS maps) after applying a Gaussian smoothing.

#### 4.7.1.1 Estimating channel-specific coefficients

So far, we have assumed that pixel intensities from different rounds and fluorescent channels all have the same scale and distribution. However, there is usually a variation among rounds and fluorescent channels, with some channel-rounds being brighter than others. To account for this effect, we model the channel-specific variations as a multiplicative factor that connects the weights at each pixel to intensities: $\mathbf{y} = \mathbf{c} \odot X\mathbf{w}$ where $\mathbf{c} = [c_1, c_2, \ldots, c_{3n}]^T$ is the channel coefficient vector and $\odot$ denotes element-wise multiplication. Suppose for a set of pixels $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \cdots, \mathbf{y}^{(P)}$ the true barcode weights $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \cdots, \mathbf{w}^{(P)}$ are given. For pixel $i$ and channel $j$, we could write: $y_j^{(i)} = c_j \sum_{b=1}^{N} X_{jb} w_b^{(i)} = c_j \sum_{b=1}^{N} (\mathbf{x}_j)_b w_b^{(i)}$ where $(\mathbf{x}_j)_b$ shows the b's element of the j's barcode. In this case, each $c_j$ can be estimated by solving an OLS problem between $y_j^{(.)}$ and $\sum_{b=1}^{N} (\mathbf{x}_j)_b w_b^{(.)}$. Conversely, if the channel coefficients are given, we can set up the decoding problem with normalized intensities: $\bar{\mathbf{y}} = \mathbf{y}/\mathbf{c} = X\mathbf{w}$ with $/$ being element-wise division. We estimate the channel coefficients in an iterative manner following the algorithm below:

1. Initialize $\mathbf{c} = \mathbf{1}$ (no channel variation)
2. Take a random sample of foreground pixels
3. Normalize the pixel intensities in the sample with $\mathbf{c}$
4. Run SpD on the normalized pixels
5. Keep pixels with one dominant unsaturated weight (weight in range 0.1 and 0.5) and obtain unbiased weights through OLS
6. Update the values of $\mathbf{c}$ by solving $3n$ OLS problems
7. Repeat steps 3-6 $n_{iter}$ times

We do this procedure for 2 iterations and apply the obtained values when decoding all fields of view.

#### 4.7.1.2 Setting the elastic net regularization parameter

Because of physical constraints, the solution to the deconvolution problem must be sparse, i.e., only a few non-zero weights should explain the observed intensities. The sparsity of the

solution is directly controlled by the L1 regularization term, $\alpha$ (equation 4.2). For a given pixel $\mathbf{y}$, higher values of $\alpha$ shrink the estimated weights ($||\hat{\mathbf{w}}_{lasso}||_1 \to 0$). Conversely, lower values of $\alpha$ allow more weights to be non-zero and $||\hat{\mathbf{w}}_{lasso}||_1$ to grow larger. In fact, one can show if the L2 regularization term, $\alpha' = 0$, the largest weight to be undetected for a pixel made purely from one barcode is $w_{max} = \frac{3n}{k}\alpha$ [73]. For instance, given $\alpha = 0.05$ and codebook parameters $n = 6, k = 3$, then $w_{max} = 0.3$. This means that a pixel composed of one barcode needs to have an underlying intensity $> 0.3$ to get a non-zero $\hat{\mathbf{w}}_{lasso}$. In other words, setting $\alpha$ too strictly will result in dimmer pixels to have $\hat{\mathbf{w}}_{lasso} = \mathbf{0}$, while setting $\alpha$ too loosely will result in spurious non-zero values in $\hat{\mathbf{w}}_{lasso}$ for brighter more complex pixels, potentially not passing the elbow filter and thus $\hat{\mathbf{w}}_{OLS} = \mathbf{0}$. To accommodate a wide range of rolony intensities, we choose $\alpha$ adaptively based on the pixel norm $||\mathbf{y}||_2$. First, we form a training data from a random subset of foreground pixels indexed by $i$. For a given pixel norm $u$, we find the alpha that maximizes a weighted sum of $||\hat{\mathbf{w}}^{(i)}_{OLS}||_1$ giving more weights to training pixels with closer norms to $u$:

$$\alpha(u) = \arg\max_{\alpha} \sum_i g\left(\frac{u - ||\mathbf{y}^{(i)}||_2}{\sigma}\right) ||\hat{\mathbf{w}}^{(i)}_{OLS}(\alpha)||_1 \tag{4.5}$$

Where $g(.)$ is the Gaussian function. In practice, for the training pixels we solve the sparse decoding problem for every value of $\alpha$ on a grid from 0.01 to 0.1 with a step size of 0.005, $\alpha_{train}$, to obtain estimated weights $\hat{\mathbf{w}}^{(i)}_{OLS}(\alpha)$. Then we create a grid of norms $\mathbf{u}_{train}$, spanning 0 and 2.8 with 50 steps. For every value of $u$ in $\mathbf{u}_{train}$, we solve equation 4.5 on the $\alpha_{train}$ grid. In other words, we create a lookup table connecting values of $\mathbf{u}_{train}$ to the best $\alpha$ in $\alpha_{train}$. For new pixels, $\alpha$ is determined by the closest norm in the lookup table.

### 4.7.2 Spot calling

To call spots, Gaussian smoothing is applied to individual OLS maps, followed by *peak_local_max* filter (scikit-image 0.19.3[74]) which returns a binary image with 1's at the local maxima of the smoothed OLS maps. These peaks are then used as markers for watershed

segmentation. From each segmented region, the following features are retained to be used in downstream steps: area, centroid, maximum and average intensity. This formed a list of candidate spots from each FOV.

### 4.7.3 Spot filtering

To control the specificity of the decoding procedure, we augmented the codebook with a number of barcodes (5-10% of the used barcodes) not used in the probe set (empty barcodes). After spot calling, we record the properties (e.g., area, maximum and average intensity) of spots with an empty barcode. Indeed, we see that empty spots tend to be smaller with lower average/maximum weight (Figure S2a-b). On a small fraction of spots from all fields of view, we train a random forest classifier (scikit-learn v1.1.3) with area, maximum and average weights as features to predict empty/non-empty labels (figure S2c). We applied the classifier to all spots and obtained emptiness probabilities and set a threshold on these probabilities (0.3-0.35).

### 4.7.4 Spot assignment to cells

The cell boundaries were computed by applying *find_boundaries* (scikit-image 0.19.3 [74]) to the segmentation mask. The distances of all spots were calculated to the closest boundary pixel. The distance was set to 0 if a spot was inside a boundary. A spot was assigned to its closest cell if the distance was less than or equal to 11um in the kidney, 3um for non-*MBP* and 0um for *MBP* spots in the brain.

### 4.7.5 Comparison of decoding methods

Datasets of varying levels of complexity were simulated to compare SpD with StarFish42 (pixel-based naive matching), BarDensr[67] and ISTDECO[66] (deconvolution-based methods). The synthetic datasets were constructed using the human brain codebook (3-on-3-off, 121 genes with 10 empty barcodes) with equal abundance of all genes and uniform spatial distribution of spots. The rolonies were modeled as Gaussian spots with peak intensity randomly chosen

to be between 0.25 and 0.7 and sigma between 2 and 2.5 pixels. To model channel-specific intensity variation, we randomly drew 18 channel-specific coefficients from a uniform distribution between 0.75 and 1.25 to scale their respective images, while clipping the intensity values above 1. We simulated multiple datasets varying the number of spots between 5*103 to 4*105 spots in a field of view of size 1024x1024 pixels. Different decoding methods were applied to the synthetic datasets with default settings to the extent possible, with no post-hoc filtering of the spots. The only exception was StarFish for which the distance threshold was set to 0.7 as a fair balance between specificity and sensitivity. Then, the ground truth spots were matched one-to-one to the decoded spots if the barcodes were identical and the centroids were closer than 6 pixels. Sensitivity is defined as the fraction of ground truth spots matched with a decoded spot. Specificity is defined as the fraction of matched decoded spots over all decoded spots. Empty rate is the fraction of empty barcodes among all decoded barcodes and is inversely related to specificity.

## 4.8 Acknowledgements

Chapter 4 is, in part, reprints of the material as it appears in Kalhor, K., Chen, C. J. ... & Zhang, K. (2024). Mapping human tissues with highly multiplexed RNA in situ hybridization. *Nature Communications*. The dissertation author was the primary investigator and co-first author of this paper.

The method described in section 4.5 was co-developed with Dr. Xuwen Li at Altos Labs.

# Chapter 5

# Spatial tissue mapping at single-cell resolution with RNA in situ hybridization

To showcase the utility of DART-FISH in gaining insights from real biological samples, we applied it to a variety of tissue sections. In the following, I summarize our efforts analyzing DART-FISH data from human brain, human kidney and mouse kidney.

## 5.1 Application of DART-FISH to human brain

To assess the performance of DART-FISH for profiling more than one hundred RNA species in large human tissue sections with fast image acquisition, we applied it to a 10um-thick, 6.9-by-4.3-$mm^2$ fresh-frozen post-mortem human M1C brain section (Figure 5.1a). The anatomy, function, and gene expression of M1C have been widely investigated at the single-cell level [75, 76], giving us a well-defined standard to compare across different studies. Note that archived human brain samples represent one of the most challenging sample types for spatial RNA mapping, due to the presence of high autofluorescence and in general, lower RNA quality.

We designed 5,097 padlock probes to target a selected panel of 121 genes containing known marker genes to resolve the spatial organization of excitatory and inhibitory neurons, as well as non-neuronal cells. The corresponding codebook followed a 3-on-3-off barcoding scheme. Imaging 6 rounds of decoding, the anchor round and the nuclear stain of this 30 $mm^2$ section of human M1C took about 10 hours. After image preprocessing and spot decoding

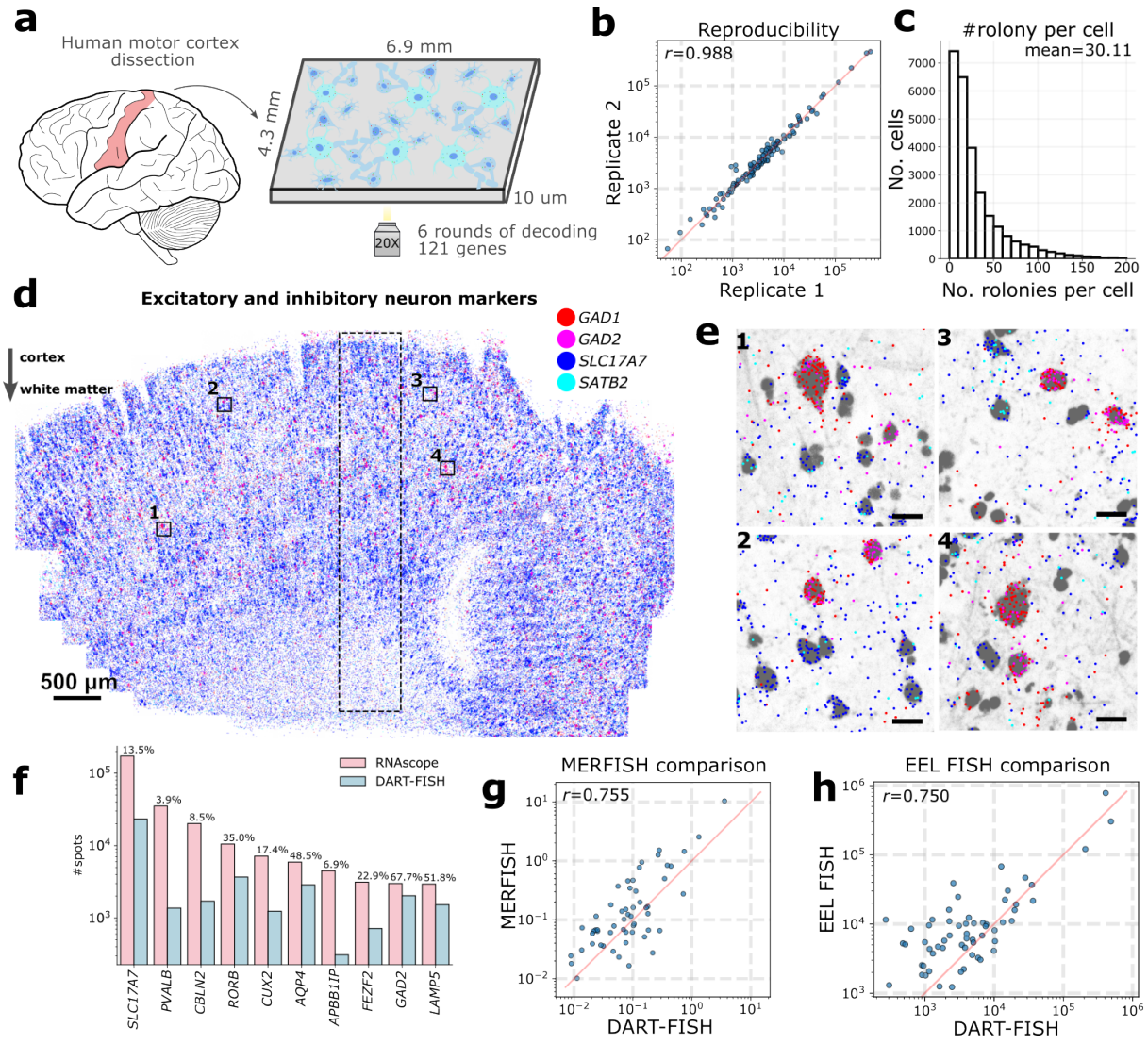



**Figure 5.1.** **(a)** Parallel sections were used from a post-mortem human M1C tissue block. Spatial distribution of 121 genes was measured by DART-FISH with 6 rounds of decoding. **(b)** Reproducibility between parallel tissue sections processed independently. Each dot represents the total count of each gene detected in each replicate. **(c)** Histogram for the number of high quality decoded rolonies per cell. **(d)** Segregation of of excitatory neuron (*SLC17A7* and *SATB*) and inhibitory neuron markers (*GAD1* and *GAD2*) in the whole tissue. The dashed rectangular box delineates the ROI in figure S5. **(e)** Zoomed-in views to show the segregation of excitatory and inhibitory markers at single-cell level in 4 ROIs indicated by the black squares in (d). Scale bars 20 um. **(f)** Quantitative comparison of counts for 10 marker genes in DART-FISH and RNAscope in equivalent ROIs. Percentages represent total spots detected in DART-FISH divided by total spots detected in RNAscope. **(g)** Comparing DART-FISH and MERFISH[77] (H18.06.006.MTG.4000.expand.rep2). Each dot represents the mean count per cell for the 56 shared genes. **(h)** Comparing DART-FISH and EEL FISH[68] (data from human visual cortex). Each dot represents the total count of the 60 shared genes.

by SpD, we obtained 2,008,260 transcripts (0.2% empty calls with 10 empty barcodes). The expression level of these 121 genes was highly consistent between two replicates (correlation coefficient $r^2 = 0.988$, Figure 5.1b), demonstrating a high reproducibility of DART-FISH.

We segmented the cells using RiboSoma and assigned the transcripts to the closest cell if the distance to the cell boundary was less than 3um (Methods, Figure 4.4c). Other transcripts were discarded from downstream analyses. Among the target genes, we noticed a higher fraction of *MBP* transcripts were found to be outside the cell bodies (93% outside, Figure S4a) while co-localizing with RiboSoma in the extrasomatic space of the cortex (Figure S4b). This observation may likely reflect the local translation of *MBP* transcripts at the axon-glia contact sites. Overall, we detected 26,646 cells with 802,361 transcripts that were assigned to a segmented cell with an average of 30 transcripts and 11 unique genes per cell (Figure 5.1c).

### 5.1.1 Benchmarking the specificity and sensitivity of DART-FISH

To assess spatial specificity of transcript localization, we first inspected the marker genes *SLC17A7* and *SATB2* in excitatory neurons and GAD1 and GAD2 in inhibitory neurons. As expected, the *SLC17A7* and *SATB2* transcripts were mainly aggregated in the soma of excitatory neurons with mutual exclusivity to GAD1 and GAD2 transcripts in inhibitory neurons (Figure 5.1d-e). We then compared the expression of 10 marker genes with the results of RNAscope generated on a parallel M1C tissue section (Methods). As shown in Figure S5, the spatial distribution of these marker genes in the same region demonstrates high concordance between RNAscope and DART-FISH. Specifically, the pan-excitatory neuron marker, *SLC17A7*, showed pronounced enrichment in the L2-L6 cortical layers. *CUX2*, *RORB*, and *FEZF2* were enriched in supragranular, granular, and infragranular layers of the neocortex, respectively, which is consistent with previous studies54–58. The observed localization of *CBLN2* in neocortical layers 2/3 and 5/6 neocortex also agrees with a previous report59. Collectively, these results indicate that DART-FISH can specifically map the spatial localization of these marker genes in human M1C.

To estimate the sensitivity of DART-FISH, we selected a similar region of interest (ROI) with equal area between RNAscope and DART-FISH samples and compared the number of transcripts of each gene. We found that the estimated sensitivity ranged from 3.9% to 67.7%, depending on the transcript (Figure 5.1f). We correlated our data to the publicly available MERFISH 60 and EEL FISH 60,61 datasets from the human brain (Pearson's r=0.755 and 0.750, respectively, Figure 5.1g and h), which we consider a high concordance given the differential probing efficiencies between different technologies, and the fact that samples from different regions were used for each technology. In summary, DART-FISH is a reproducible spatial transcriptomic method with the sensitivity and specificity to detect hundreds of RNA species in their spatial context, with potential for providing biologically meaningful insights to the human brain despite the high natural background autofluorescence.

### 5.1.2 Organization of cell types in the human primary motor cortex

To assess whether DART-FISH is able to resolve the organization of various cell types of human M1C, we set out to perform cell annotation by performing clustering on DART-FISH cells and matching them to the highest correlated subclass from a recent single-nucleus RNA sequencing (snRNA-seq) reference of M1C [78] (Methods, Figure 5.2a and b, Figure S4c). We resolved 20 subclasses from the major excitatory, inhibitory, and non-neuronal cell classes which constituted 24.3%, 10.6%, and 65.1%, respectively, in the M1C (Figure 5.2c-g). For excitatory neuronal subclasses, we successfully detected their laminar distribution, with L2/3 IT neurons localized at the superficial layer of the cortex and L6b/CT neurons deep in the cortex and close to the white matter (Fig. 3b-d), in line with the evolutionarily conserved organization of excitatory neurons in the mammalian M1C. Of note, L6 IT Car3 cells seem to be positioned more superficially than the L6 IT population, consistent with recent observations in human visual cortex and middle temporal gyrus [78, 79] (Figure 5.2d).

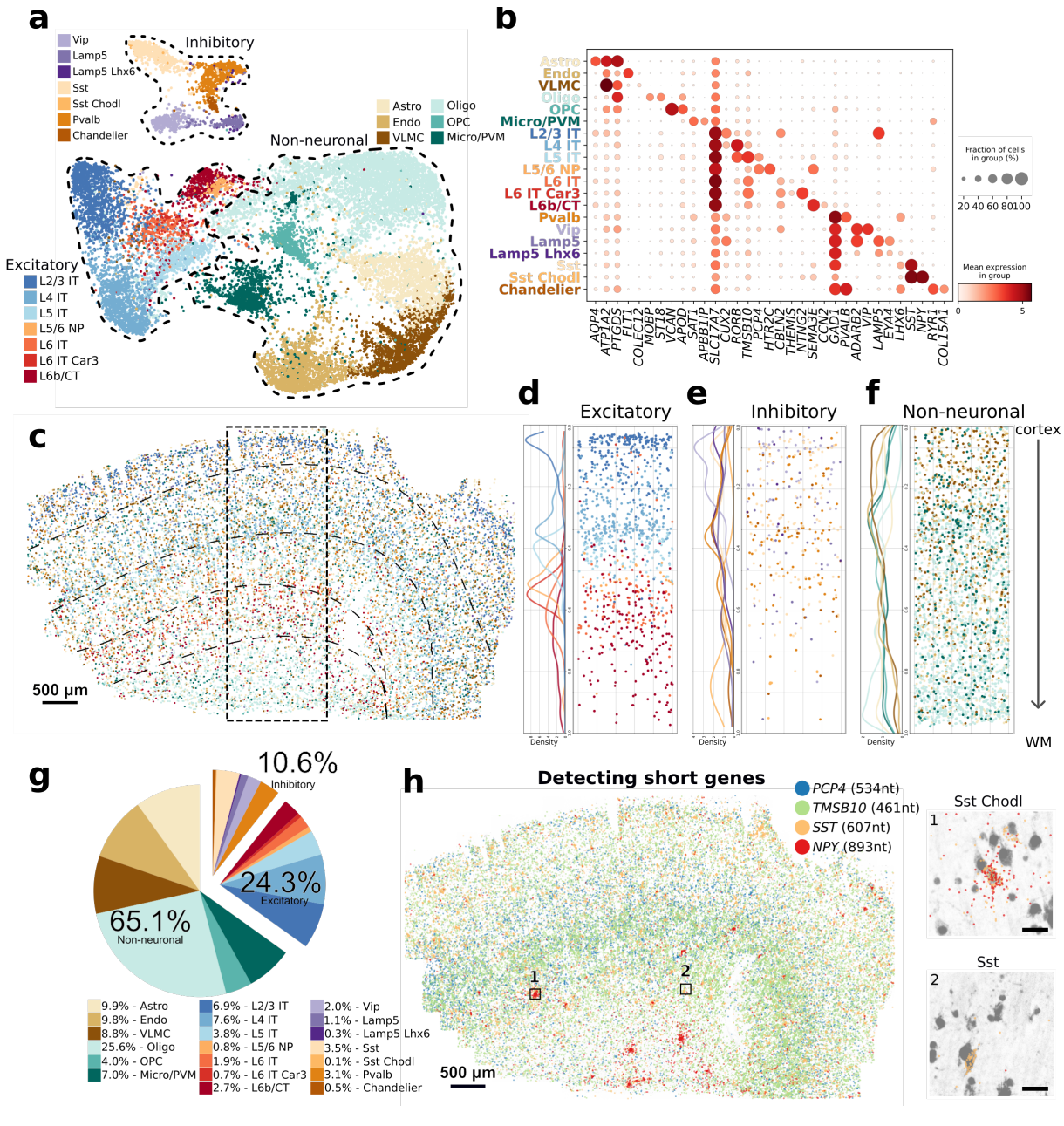
On the other hand, inhibitory neuronal subtypes generally showed wider spatial gradients along the cortical axis; for instance the Vip population was enriched in layer 2-4 as suggested

by previous studies in the mouse[76] (Figure 5.2b and e). Moreover, we observed some cells belonging to the excitatory neurons and inhibitory neurons localized in the white matter region, which may be the adult remnants of early generated subplate neurons discovered in previous studies[80]. For non-neuronal cells, we observed oligodendrocytes appearing at layer 4 and peaking in the white matter in spite of the uniform distribution of the oligodendrocyte progenitors across the tissue section (OPC, Figure 5.2f).

### 5.1.3 Detecting short genes enables detection of rare cells

We further assessed whether we could detect short genes (<1.5kb) with DART-FISH. smFISH-based methods rely on tiling sufficiently long RNA molecules with probes to generate detectable fluorescent signals. In contrast, DART-FISH requires only one padlock probe to bind successfully to the target to detect it. To boost our chances for detecting shorter genes, we allowed overlapping targets in our design strategy to obtain more probes for short RNA species [53] (figure 3.6b, *NPY* as an example). We compiled a list of 33 differentially expressed genes shorter than 1.5kb comprising well-studied genes as well as less well-known computationally-derived marker genes in the brain. For example, by targeting *SST* (length 607-nt) and *NPY* (893 nt), we could uncover a rare subclass of inhibitory neurons, Sst Chodl (0.1% abundance, Figure 5.2g), specified by the expression of these short neuropeptides (Figure 5.2b and h). Sst Chodl cells were found to be enriched in deeper layers, consistent with previous reports[81].

**Figure 5.2.** **(a)** UMAP plot of all annotated excitatory neurons (L2/3 IT, L4 IT, L5 IT, L5/6 NP, L6 IT, L6 IT Car3, and L6b/CT), inhibitory neurons (Pvalb, Vip, Lamp5, Lamp5 Lhx6, Sst, Sst Chodl, and Chandelier), and non-neuronal (Astro, Endo, VLMC, Oligo, OPC, and Micro/PVM) subclasses. Astro: astrocytes, Endo: endothelial cells, VLMC: vascular and leptomeningeal cells, Oligo: oligodendrocytes, OPC: oligodendrocyte precursor cells, Micro/PVM: microglia/perivascular macrophages, IT: intratelencephalic, CT: corticothalamic, NP: near-projecting. **(b)** Dot plot of marker gene expression across annotated subclasses. **(c)** Spatial distribution of all annotated cell types in the entire M1C tissue section from upper cortical layer at the top to the white matter (WM) at the bottom. The dashed rectangular box delineates the ROI in d-f. **(d), (e), (f)** show the density plot (left) and spatial distribution (right) of excitatory neurons, inhibitory neurons, and non-neuronal subclasses, respectively. **(g)** Pie chart depicting the relative frequency of annotated subclasses (n=1 section). **(h)** Spatial distribution of targeted short RNA species *PCP4*, *TMSB10*, *SST*, and *NPY* in the M1C tissue section. *PCP4* and *TMSB10* are layer 5 and layer 5-6 markers, respectively. Sst Chodl cells (0.1% abundance) are $SST^{+}$ $NPY^{+}$. Inset 1 shows an example of a Sst Chodl cell, while inset 2 is a $SST^{+}$ $NPY^{-}$ cell from the much frequent Sst subclass (abundance 3.5%). Inset scale bars 20um.


a
Vip
Lamp5
Lamp5 Lhx6
Sst
Sst Chodl
Pvalb
Chandelier
Inhibitory
Astro
Endo
VLMC
Oligo
OPC
Micro/PVM
Non-neuronal
Excitatory
L2/3 IT
L4 IT
L5 IT
L5/6 NP
L6 IT
L6 IT Car3
L6b/CT
b
Astro
Endo
VLMC
Oligo
OPC
Micro/PVM
L2/3 IT
L4 IT
L5 IT
L5/6 NP
L6 IT
L6 IT Car3
L6b/CT
Pvalb
Vip
Lamp5
Lamp5 Lhx6
Sst
Sst Chodl
Chandelier
AQP4
ATP1A2
PTGDS
FLT1
COL1A2
MOBP
ST18
VCAN
APOD
SAT1
APBB1IP
SLC17A7
CUX2
RORB
TMSB10
PCP4
HTR2C
CBLN2
THEMIS
NTNG2
SEMA3E
CLDN5
PVALB
ADARB2
VIP
LAMP5
EYA4
LHX6
SST
NPY
RELN
COL15A1
Fraction of cells in group (%)
20 40 60 80100
Mean expression in group
0
5
c
500 µm
d
Excitatory
Density
e
Inhibitory
Density
f
Non-neuronal
Density
cortex
WM
g
10.6%
Inhibitory
24.3%
Excitatory
65.1%
Non-neuronal
9.9% - Astro
9.8% - Endo
8.8% - VLMC
25.6% - Oligo
4.0% - OPC
7.0% - Micro/PVM
6.9% - L2/3 IT
7.6% - L4 IT
3.8% - L5 IT
0.8% - L5/6 NP
1.9% - L6 IT
0.7% - L6 IT Car3
2.7% - L6b/CT
2.0% - Vip
1.1% - Lamp5
0.3% - Lamp5 Lhx6
3.5% - Sst
0.1% - Sst Chodl
3.1% - Pvalb
0.5% - Chandelier
h
Detecting short genes
PCP4 (534nt)
TMSB10 (461nt)
SST (607nt)
NPY (893nt)
1
2
500 µm
Sst Chodl
Sst

In addition to these short neuropeptides, DART-FISH also detected other short RNA species including *PCP4* (534nt) and *TMSB10* (461nt) with pronounced localization (Figure 5.2h). *PCP4* is a known layer 5-6 marker in the mouse cerebral cortex while we propose *TMSB10* as a novel deep layer marker gene.

To quantify how well the targeted genes performed, we correlated their average expression at subclass level between DART-FISH and snRNA-seq (Methods, Figure S4d). We found 25 of 33 (75%) of the genes shorter than 1.5kb and 81 of 88 (92%) of the longer genes had higher correlations than 0.5. This is similar to a MERFISH data set targeting another region of the human cortex with 250 genes (88% with >0.5 Pearson's correlation, figure S4d). Taken together, we showed that DART-FISH can accurately map the distribution of all the main neuronal and non-neuronal subclasses in the human brain and can uncover rare cell populations by detecting short genes.

## 5.2 Application of DART-FISH to diseased human kidney

To demonstrate the applicability of DART-FISH to a clinically relevant tissue context, we next applied it to the human kidney. The kidney is composed of repetitive functional tissue units, called nephrons, with various closely organized cell types including endothelial, stromal, immune and epithelial cells that regulate the filtration of the blood as well as other homeostatic functions such as maintaining electrolyte and fluid balance (Figure 5.3a, Figure S6a). The homeostatic interactions between these cell types are perturbed in kidney disease and can lead to fibrosis and decline in kidney function [82]. We recently reported an atlas of cell types in healthy and diseased patients, and identified multiple maladaptive cell states that are associated with kidney disease [27] (see chapter 2). In the same study, we used sequencing-based spatial transcriptomics methods with 10um and 55um resolution to map cellular neighborhoods in healthy and diseased samples, respectively, which lacked the resolution needed to delineate the exact cellular composition, the boundaries and the positioning of cells within the neighborhoods.

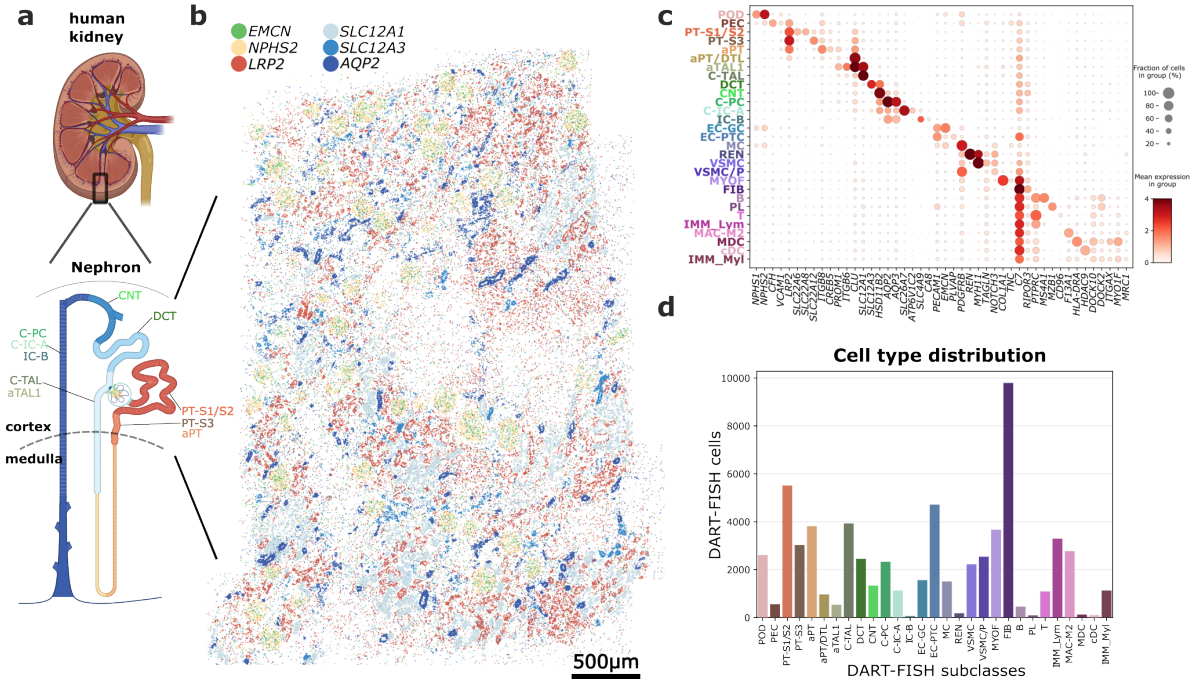



**Figure 5.3.** **(a)** Applying DART-FISH to a 4.9x3.8mm$^2$ section from the cortex of the human kidney (adapted from BioRender). The nephron schematics shows the expected epithelial subclasses in the section 102. **(b)** The spatial expression of key marker genes for the cortical segments: EMCN: glomerular capillary endothelial cells (EC-GC), *NPHS2*: podocytes (POD), *LRP2*: proximal tubules (PT), *SLC12A1*: cortical thick ascending limbs (C-TAL), *SLC12A3*: distal convoluted tubules (DCT), *AQP2*: cortical principal cells of the collecting duct (C-PC). **(c)** Dotplot of marker gene expression for the annotated subclasses. **(d)** Bar plot showing the number of cells from each annotated subclass in the human kidney from $n = 1$ section. High numbers of immune cells and fibroblasts are suggestive of inflammation and fibrosis.

We reasoned that the high spatial resolution provided by DART-FISH is complementary to the sequencing-based methods and can help define cellular niches more accurately.

Guided by the published single-nucleus reference atlas, we designed a panel of 300 genes with 6299 padlock probes following the 3-on-4-off barcoding scheme, focusing on the major healthy cell types of the kidney, immune cells and cell states implicated in kidney disease. We then performed DART-FISH on tissue sections from the kidney cortex of a patient with various clinical features including glomerulosclerosis, interstitial fibrosis, tubular atrophy, and chronic inflammation identified by a pathologist. Our gene panel correctly mapped spatial organization of cells in different regions of the nephron including glomeruli and cortical tubules (Figure 5.3b).

For instance, the transcripts *NPHS2* and *EMCN* which mark podocytes and glomerular capillary endothelial cells, respectively, are mainly found in the glomerular tuft of the round appearing renal corpuscles. We then compared our data with a Slide-seq dataset from a healthy individual. At the bulk level, the DART-FISH data is correlated with slide-seq (Pearson's r=0.609) with cells in DART-FISH demonstrating more copies of the targeted genes than Slide-seq beads 74 (median fold-change per gene=2.2 for the top 150 genes in slide-seq, figure S6b). The comparison also showed upregulation of markers of inflammation in the DART-FISH dataset, consistent with the underlying pathology in our sample (figure S6b). Hence, the spatial distribution of known kidney marker genes and their overall counts are consistent with kidney biology and prior data.

### 5.2.1 Organization of cell types at the single-cell level

To find the molecular identity of the cells in the human kidney, cell segmentation was performed using both RiboSoma and nuclear stains. We found RiboSoma to be superior to the nuclear stain in revealing tubular morphology and distinguishing the interstitial cells (figure S6c). Subsequently, with 30,000 segmented cells with an average of 30 detected transcripts and 20 unique genes per cell (figure S6d-e, empty rate <0.25% with 15 empty barcodes), the kidney DART-FISH data was annotated to cortical and altered cell types as identified in the single-cell kidney atlas [27] (figure S6g, figure S6f, figure S7, Methods). These annotated cell types were of the expected relative proportions and showed strong and specific differential expression of corresponding marker genes (Figure 5.3c, figure S6f). Thus, DART-FISH could confidently resolve >20 cell types and states in the human kidney.

Next, we investigated the neighborhoods formed by the healthy cell types. The complex archetypical structure of the renal corpuscle was successfully recapitulated, with podocytes (POD), glomerular capillary endothelial cells (EC-GC) and glomerular mesangial cells (MC) confined within the glomerular tuft, surrounded by parietal epithelial cells (PEC) or the outer layer of the Bowman's capsule and juxtaposed with the renin-secreting cells (REN) in the wall of the arterioles (Figure 5.4a, figure S6a). We also detected medullary rays with the

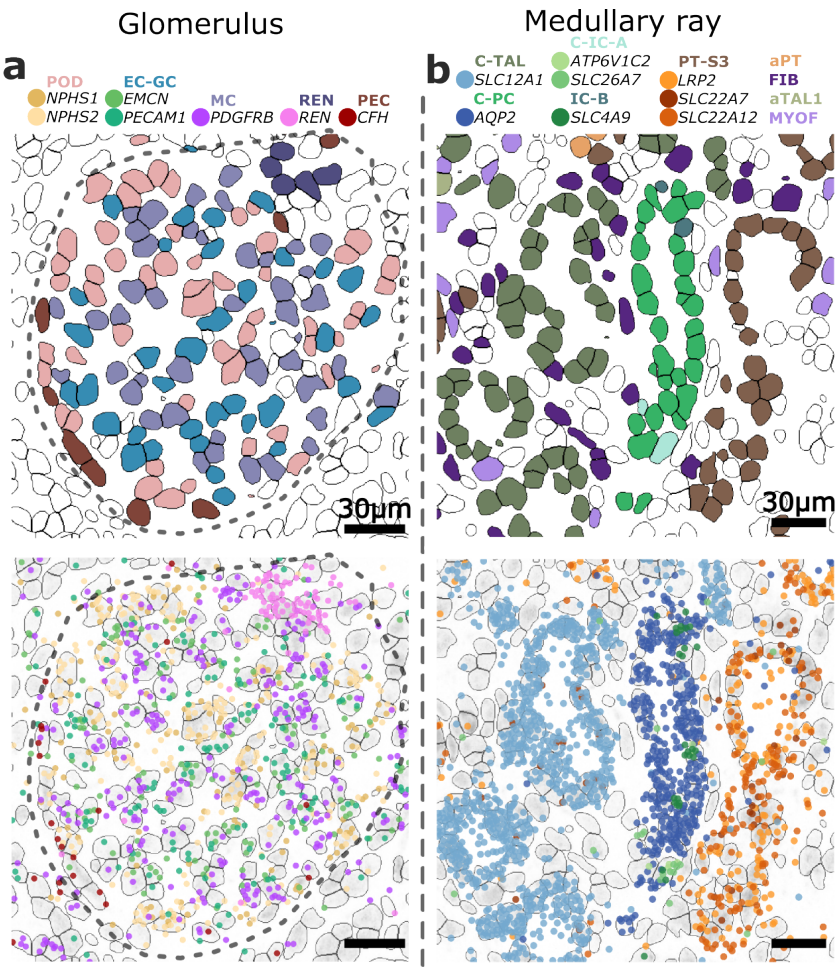



**Figure 5.4. (a)** An example of a glomerulus with part of the juxtaglomerular apparatus. (top) cells colored by the annotated subclass, (bottom) marker genes corresponding to the subclasses. Each dot represents one rolony. Dashed line delineates the boundary of the renal corpuscle. **(b)** Example of a medullary ray with a bundle of TALs, PT-S3, and collecting ducts. Note that for clarity, some cell types, i.e., aPT, FIB, aTAL1 and MYOF are plotted (top) but their corresponding marker genes are omitted (bottom).

characteristic bundling of the tubules of cortical thick ascending limb (C-TAL), the S3 segment of proximal tubules (PT-S3) and collecting ducts (Figure 5.4b). Further, collecting ducts comprising intermixed principal cells (PC) and alpha- and beta-intercalated cells (C-IC-A and IC-B) could be clearly resolved. These results show that our cell type annotations closely match the known structures within the human kidney.

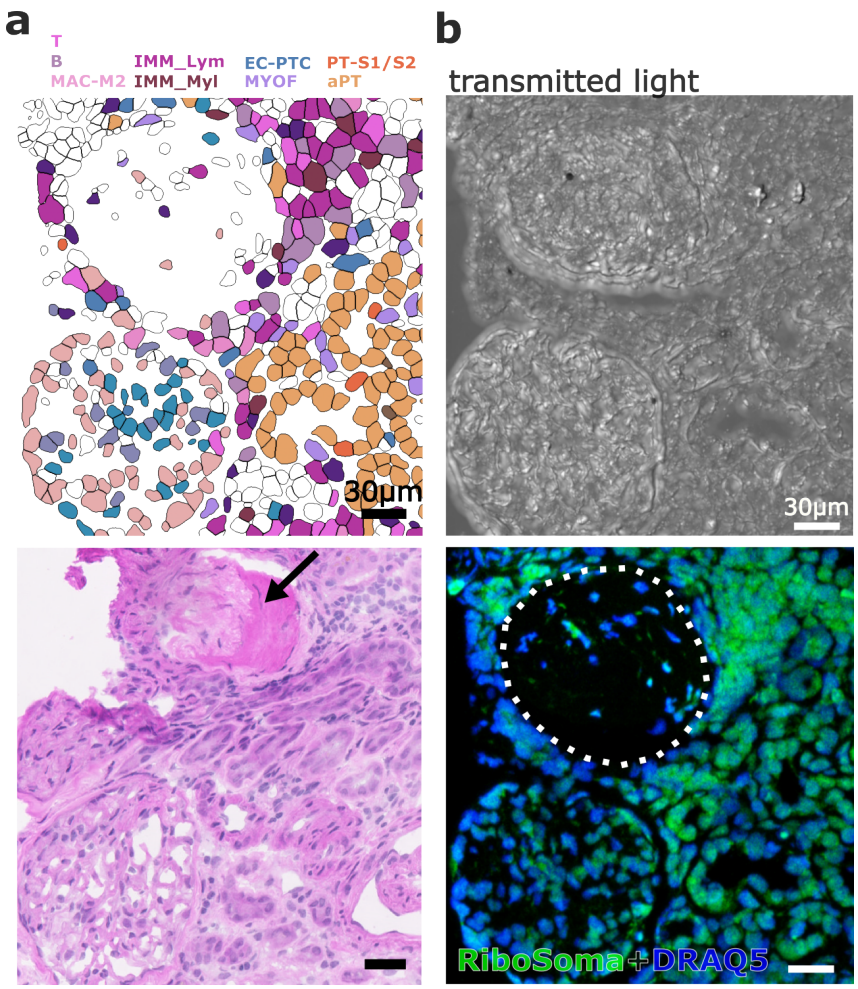



**Figure 5.5. (a)** Example of a pathological niche with inflammation, a sclerosed glomerulus and altered proximal tubule cells adjacent to a more normal glomerulus (top). The same area on an H&E-stained parallel section from the same tissue block confirms the decellularization and inflammation observed in DART-FISH. The black arrow points to the sclerotic glomerulus. **b** Transmitted light (top) and RiboSoma overlaid with nuclear stain (bottom) of the same ROI as in (a). The cells in the sclerosed glomerulus (dashed line) are mostly replaced by scar tissue as shown by the occupied space in the transmitted light view.

### 5.2.2 Profiling histopathologically abnormal cells and neighborhoods

To compare the tissue morphology obtained from DART-FISH with a clinically relevant histological stain, we performed Hematoxylin and Eosin (H&E) staining on a parallel section from the same tissue block. In an area with putative inflammation on the H&E slide, we observed an abundance of immune cells of both lymphoid and myeloid origin on the DART-FISH section

(Figure 5.5a). These immune cells surround a sclerotic glomerulus, which in contrast to a more normal glomerulus, is depleted from cells and is instead fibrotic (shown by an arrow in figure 5.5a). In DART-FISH, this phenomenon can be clearly detected by contrasting the low cell numbers revealed by RiboSoma and the physically occupied space through the accompanying transmitted light image (Supplementary figure 5.5b). Thus, by paired H&E staining we showed that DART-FISH can capture different pathological phenomena with a molecular resolution beyond that of the traditional histology.

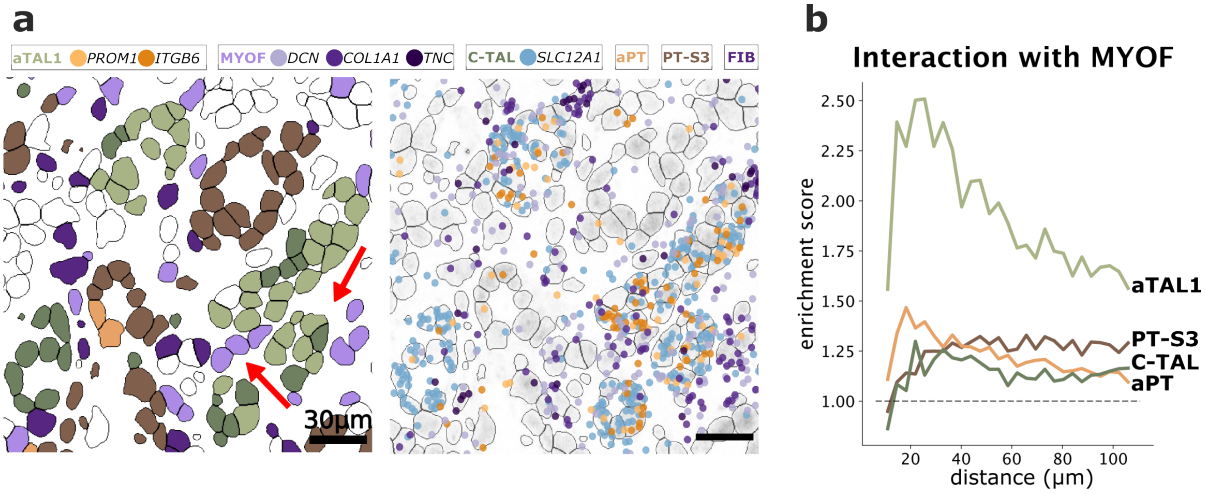



**Figure 5.6.** **(a)** Example of a pathological niche composed of aTAL1 cells and myofibroblasts. The left plot shows the cell type annotations, and the right plot shows the expression of relevant marker genes. Red arrows point toward densities of MYOF and aTAL1 cells. **(b)** Plot showing the co-occurrence enrichment [83] of some cell types with MYOF at a range of distances, suggesting an interaction between MYOF with aTAL1 cells whereas there is no apparent co-occurrence enrichment between MYOF and aPT, or healthy PT-S3 and C-TAL.

In addition to healthy cell types, DART-FISH was also able to reveal distinct pathological cell states. This includes a population of myofibroblasts (MYOF) expressing matrisome genes including *COL1A1*, *TNC*, *DCN* and *POSTN*, suggestive of their ECM-producing role in kidney fibrosis [27, 84]. Furthermore, we detected altered PT (aPT) and TAL (aTAL1) populations, both of which expressed PROM1, in line with recent findings [27, 85]. To determine whether these pathological cell states form distinctive niches, computational methods were applied to find pairs of cell types that showed enrichment in their spatial colocalization [83]. Interestingly,

in neighborhoods around MYOFs, there was an increased presence of aTAL1 cells compared to C-TAL and aPT (Figure 5.6). This observation indicates a possible interplay between the maladaptive repair of TALs and fibrosis. We speculate that there are a variety of cellular neighborhoods associated with adaptive repair and fibrosis that could be defined through further studies. All in all, these results demonstrate how DART-FISH as a single-cell resolution spatial transcriptomic technique can be used to interrogate neighborhoods of cell types and states defined by single-cell RNA sequencing studies in diseased human tissues.

## 5.3 Organ-scale imaging with DART-FISH in mouse kidney

### 5.3.1 Motivation

So far in the data presented, we have focused on tissues that constitute a small portion of the entire organ. However, this cannot typically inform us about all the compartments that are involved in tissue function, neither can it resolve all the compartments affected by a disease condition. On the other hand, a condition that affects one region of an organ, may differentially alter other regions. Moreover, an insult in one part of an organ may lead to an adaptive change in another part. Thus, to obtain a comprehensive view of tissue function or dysfunction in disease, it is most useful to image entire organs.

As discussed in chapter 3, DART-FISH enables data acquisition from centimeter-sized samples within a reasonable time frame (1-2 days of decoding time). Since human organs are generally larger than a few centimeters, upon preparation of post-mortem blocks they have to be divided into smaller pieces. On the other hand, many mouse organs are small enough to fit on a microscopy slide. Therefore, while organ-scale spatial transcriptomics may not be within reach for human tissues, imaging entire organs from smaller model organisms like mouse fits well within the specifications of DART-FISH.

Here, I demonstrate organ-scale imaging on a longitudinal section of the kidney containing the main anatomical structures from cortex to papillary tip. Using a panel of 170 genes, I

could detect all epithelial cell types along the nephron involved in the kidney function as well as the accompanying vasculature. Gene expression alone can segment the kidney into regions in line with their anatomical structures. Additionally, in the panel were a number of genes involved in pathology, such as genes for injury, fibrosis, immune signalling and function. Using these genes, I identified spatial neighborhoods of injury and repair with varying levels of enrichment for markers of inflammation and fibrosis.

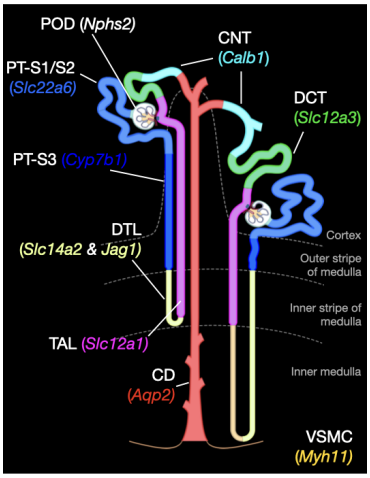



**Figure 5.7.** Schematic of the nephron with marker genes

### 5.3.2 All nephron components in a single dataset

DART-FISH was applied to a kidney section from a mouse with 16 months of age. To image the entire tissue, it was tiled with 219 FOVs, each of which covering an area of 540um-by-540um. Considering the overlap between adjacent FOVs, this amounts to an area of about 50mm$^2$. In this area, 10 million transcripts were decoded from the 170 genes in the panel. Cell segmentation counted 371,000 cells in the dataset. On average, there were 26 decoded transcripts per cell.

The nephron components are visualized in the schematic in Figure 5.7. About 50 genes out of the 170 genes in the panel are specific to the components of the nephron, from glomerular capsule to the collecting ducts. The spatial expression of nine of these genes as well as a vascular

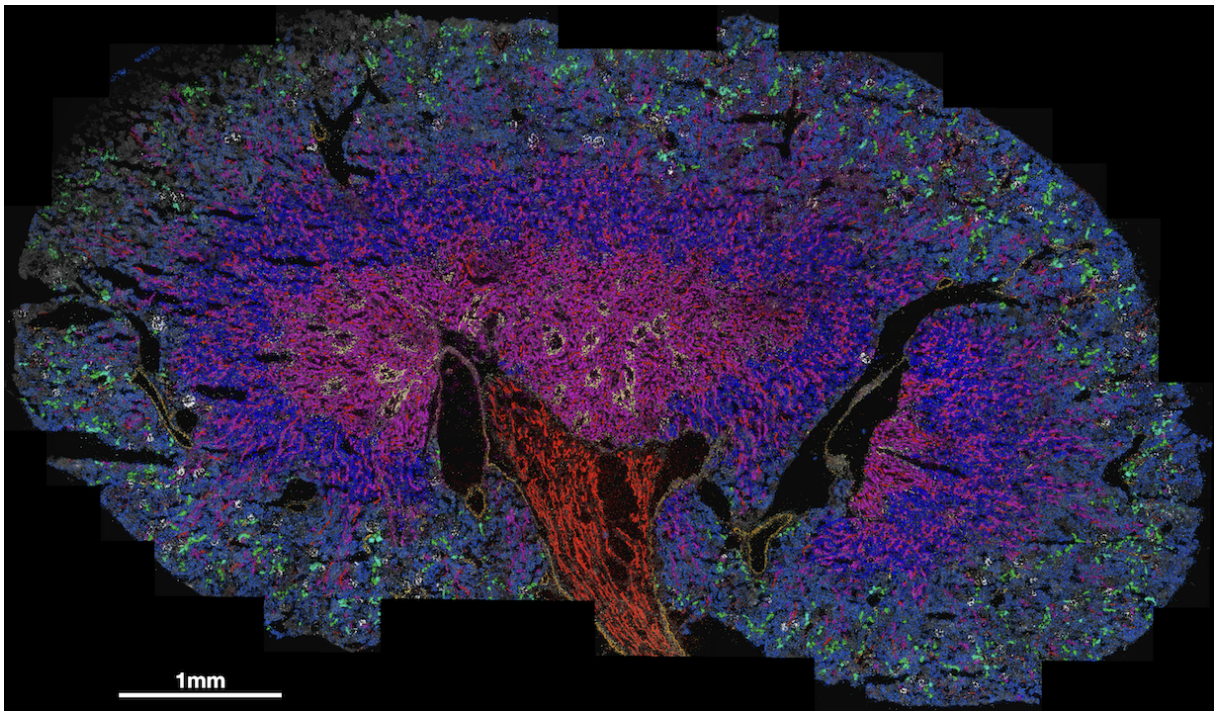



**Figure 5.8.** Organ-scale image of a mouse kidney. Ten genes are plotted for each main epithelial segment and vasulature. This image consists of 1.8 million dots, each of which representing a rolony from these genes. The black and white background is N9 cDNA and the color key for the genes follows figure 5.7. In detail, *Nphs2* for podocytes in glomeruli in white. *Slc22a6* for S1 and S2 segments of proximal tubules in light blue. *Cyp7b1* for S3 segment of proximal tubules in dark blue. *Slc14a2* and *Jag1* for descending thin limb in light yellow. *Slc12a1* in magenta for thick ascending limb. *Slc12a3* in green for distal convoluted tubules. *Calb1* in cyan for connecting tubules. *Aqp2* in red for principal cells of the collecting ducts. *Myh11* in gold for smooth muscle cells of the vasculature. Note that, ascending thin limb cells are not plotted here due to not having a unique and strong marker gene in the panel.

marker is plotted in Figure 5.8. These genes broadly mark all the main cell types involved in the kidney function, that is, filtering the blood and producing urine.

As expected from kidney physiology [35], glomeruli, the S1/S2 segment of proximal tubules, distal convoluted tubules and connecting tubules are confined to the cortex. The composition of cell types changes drastically moving toward the depth of the kidney. The S3 segment of proximal tubules have a strong presence in the outer stripe of outer medulla, the descending thin limbs and thick ascending limbs have high abundance in the inner stripe of outer medulla and the inner medulla is strong in collecting ducts. In fact, by just clustering the local

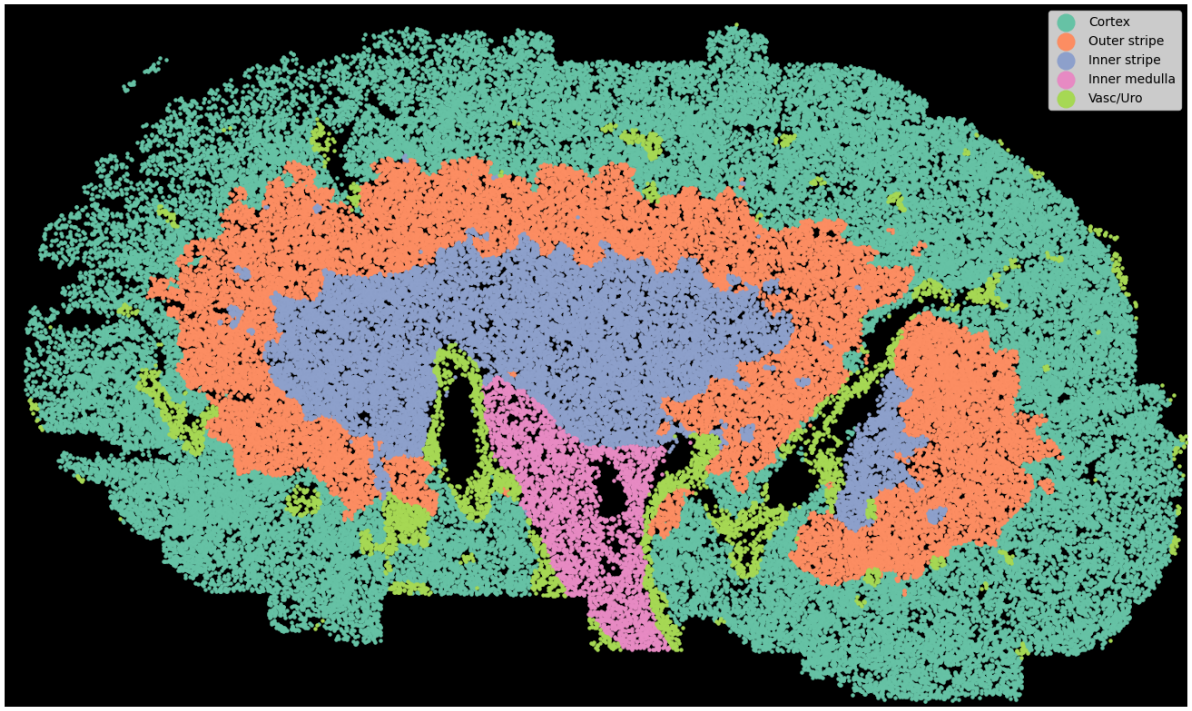



**Figure 5.9.** Anatomical domains of the kidney. To obtain these domains, gene expression neighborhoods were formed and clustered into 5 major groups. A neighborhood is defined as a circle of diameter 135um around a center enclosing a number of rolonies. The centers were sampled randomly from spots within the tissue. Based on the expression of marker genes and their spatial organization, the clusters were annotated as cortex, outer and inner stripes of the outer medulla, inner medulla, and vasculature/urothelium.

gene expression, I could resolve these anatomical regions across this data set (Figure 5.9).

### 5.3.3 Systematic identification of non-epithelial and injury domains

As a result of the age of the mouse donor (16 months), we expect to see some level of aging-associated injury and damage. As previously mentioned, 50 genes in the probe set are specific to healthy epithelial process. I refer to these genes as "healthy genes". To systematically identify areas that might be enriched for non-healthy or non-epithelial processes, I selected neighborhoods that had a low representation of healthy genes and were enriched in non-healthy genes (sum of expression of injury genes $\geq 50$ and sum of expression of healthy genes $\leq 50$). Then clustering algorithms were applied to identify stable and distinct populations of cells

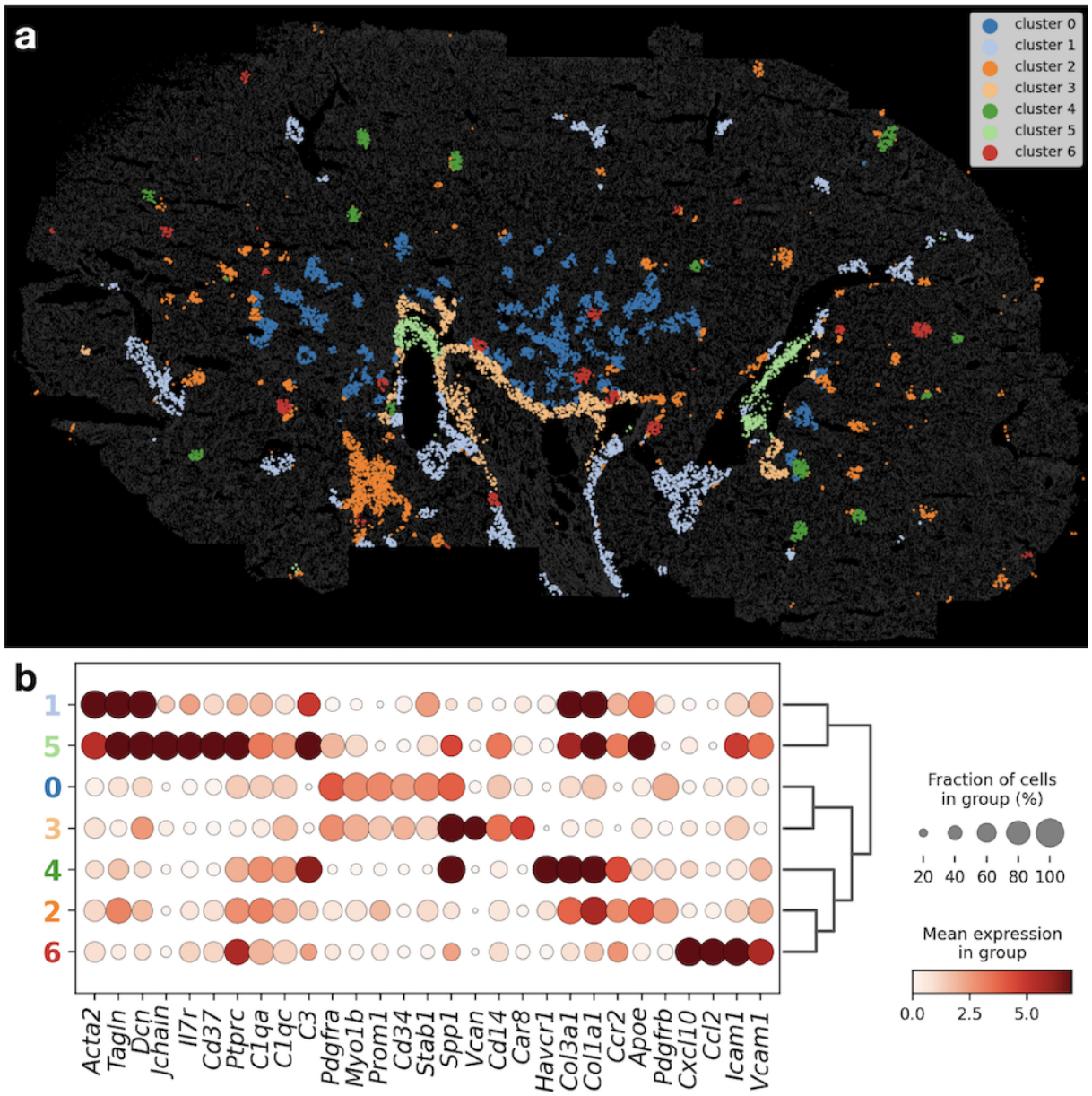



**Figure 5.10.** Non-epithelial domains in mouse kidney. **a.** Neighborhoods were selected based on high expression of non- and low expression of healthy epithelial genes. Stable clustering (as in section 5.1) divided the neighborhoods into 7 groups. **b.** Dot plot of the genes that are differentially expressed in each cluster. These distinguishing genes include inflammatory markers, signalling molecules and matrisome genes.

(normalization of gene expression and clustering was done similar to section 5.1).

The clusters obtained from this analysis are depicted in Figure 5.10a and their corresponding distinguishing genes are in figure 5.10b. These clusters span a variety of non-epithelial

structures. Some are structural; for instance, cluster 1 closely follows the Vasculature/Urothelium domain of figure 5.10. The genes that standout the most are related to VSMCs (*Tagln*), myofibroblasts (*Acta2*) and fibroblasts (*Dcn*). They clearly indicate the presence of vascular and perivascular structures. Similarly, cluster 0 seems to show neighborhoods of fibroblasts and endothelial cells in the medulla. Whether or not this cluster is age-related or normal will need experiments with young controls.

On the other hand, some clusters have clear signs of pathology. For instance, cluster 5, closely related to cluster 1, is also enriched in markers of inflammation, including monocytes, B cells, plasma cells and T cells. Moreover, neighborhoods of cluster 5 are concentrated in regions close to vascular cells of cluster 0. These properties are reminiscent of perivascular cell clusters with characteristics of tertiary lymphoid organs that are associated with kidney aging [86, 87, 88]. The identity of cluster 3 is not immediately clear. It is structurally contiguous to clusters 1 (vascular and perivascular cells) and 5 (the likely tertiary lymphoid organ) however it does not express the same fibroblast and contractility markers. It is however, strongly expressing *Spp1*, or Osteopontin, a gene with complex function both in normal cases and injured cases in the kidney and other tissues [89].

Cluster 6 is highly enriched in genes *Cxcl10, Ccr2, Icam1* and *Vcam1* with slightly less presence of monocyte genes (*C1qa, C1qb, C1qc*). Zooming-in on an area of cluster 6 neighborhoods (Figure 5.11) shows us an area with cells expressing healthy epithelial genes (figure 5.11a-b). These cells are juxtaposed with stromal cells, likely endothelial cells and fibroblasts that are not highly active in extracellular matrix remodeling (figure 5.11d). These stromal cells are expressing high levels *Cxcl10, Ccr2, Icam1*. These genes are implicated in the recruitment of monocytes and lymphocytes [90, 91]. Based on these patterns, we could conclude that this is a site of forthcoming inflammation. Therefore, cluster 6 marks inflammatory but not fibrotic niches. Based on the gene expression, the inflammation in these niches is mediated by chemokines such as *Cxcl10, Ccr2*.

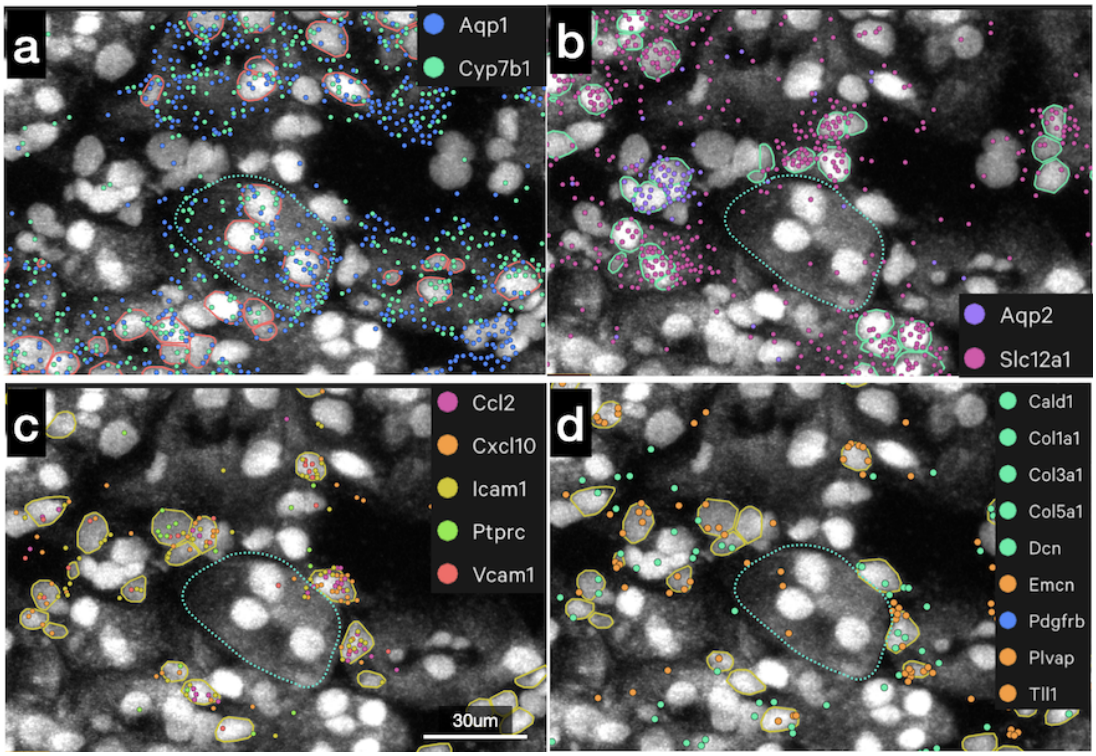



**Figure 5.11.** Example of a pre-fibrotic niche of cluster 6. **a and b.** Healthy epithelial genes. Dashed line is drawn around a proximal tubule. **c.** Cells expressing immune recruiting chemokines and other signalling molecules. *Ptprc* is an immune cell marker. **d.** Endothelial and matrisome-producing fibroblast genes. The chemokine expression cells of panel c are fibroblasts and endothelial cells and are not expressing high levels of extracellular proteins. Plots are screenshots of Xenium Explorer v2.0 by 10x Genomics.

### 5.3.4 Discussion

In this section, I used two main capabilities of DART-FISH, namely, its large area imaging and single-cell resolution detection, to systematically identify anatomical and pathological cellular neighborhoods across a section of an entire murine organ. By forming neighborhoods that aggregate gene expression from multiple cells, various properties of these niches could be studies. Including all genes will mark anatomical regions in the kidney (figure 5.9) while separating out the injury-related genes will uncover niches that are undergoing injury or repair (figure 5.10). Furthermore, upon the systematic identification of these niches, one could zoom-in and study their composition at the single-cell levels. This unbiased strategy is useful for extracting interesting phenomena from large data sets.

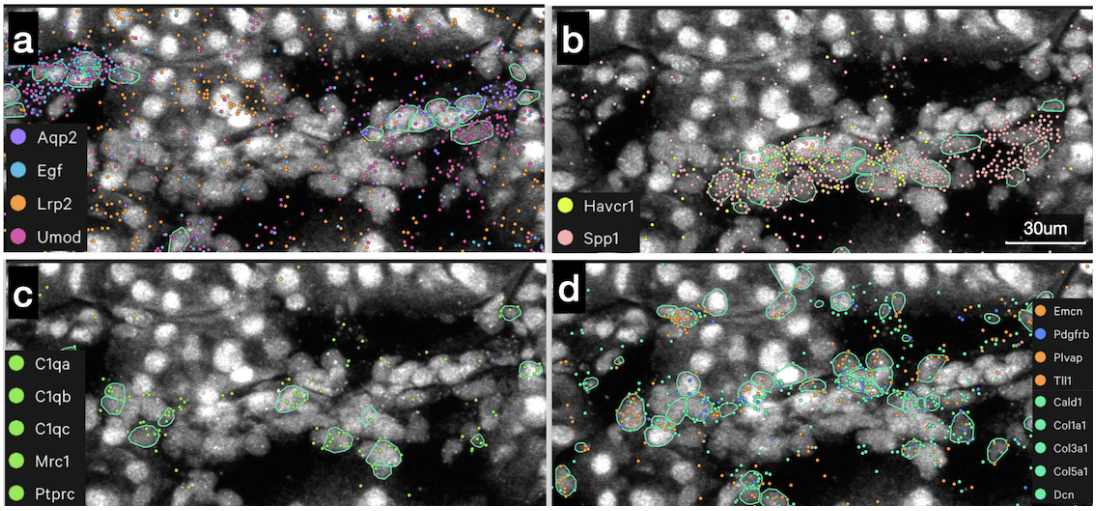



**Figure 5.12.** Example of a fibrotic and inflamed niche of cluster 4. **a.** Healthy epithelial genes. **b.** Injury-associated genes. Loss of epithelial markers are seen in cells positive for injury genes. **c.** Immune genes. **d.** Endothelial and matrisome-producing fibroblast genes. Strong expression of these genes are observed compared to figure 5.11d.

Increasing the sample size is necessary for rigorously investigating of the mechanisms of disease. Different samples in case-control studies and time-course studies can be combined at neighborhood level before clustering to identify shared and recurring patterns across all samples. Then, the incidence of these neighborhoods can be compared across conditions with proper statistics. Perhaps, with a rich dataset, different neighborhoods can be computationally aligned across disease and repair trajectories. Furthermore, because the imaging is done at an organ scale, the neighborhoods can be stratified by their anatomical location, e.g., position along the corticomedullary axis, or their proximity to important structures like blood vessels.

## 5.4 Methods

### 5.4.1 Cell annotation

We used anndata [92, 93, 94](v0.8.0) and scanpy [92] (v1.9.1) to handle and analyze the data. The data normalization was performed using analytic Pearson residuals [95] (clipped at 40) with a lower bound placed on gene-level standard deviations [96]. Clustering was done with the Leiden algorithm implemented in scanpy.

#### 5.4.1.1 Annotating the Brain data set

Cells with counts less than 5 and more than 300 were removed (2980 out of 26348). The top 100 highly variable genes were used for normalization (*scanpy.experimental.pp.highly_variable_gene(., flavor="pearson_residuals")*), embedding and annotations. PCA was performed on pearson residuals, and the neighborhood graph was created with this command *scanpy.pp.neighbors(., n_neighbors=20, n_pcs=15, metric='cosine')*. Single-nucleus RNA-seq reference from Jorstad et al. [78] was subsetted to M1C cells and normalized in the same way as DART-FISH. Pax6 and Scng subclasses were removed since we did not design our probe set to target those. Average normalized counts (centroids) were computed for every other subclass in the "within_area_subclass" slot and all clusters of DART-FISH. To annotate the DART-FISH clusters at the class level (excitatory, inhibitory, non-neuronal), we first correlated each cluster to all single-nucleus subclasses, and assigned that cluster to the class of the most highly correlated subclass. Annotation of each class was done separately.

For excitatory neurons, all DART-FISH cells that had a class label of "excitatory" and had at least 20 transcripts were kept (5957 cells). We realized that the Leiden clustering was unstable and by mere shuffling of the order of cells, we would obtain very different clusters. We reasoned that by removing some cells that tend to move between clusters, we could get more stable clusters and have more confidence in their annotation. To find cells that don't stably cluster, we ran clustering 20 times, every time shuffling the order of the cells. For every cell, we

calculated the number of times it was co-clustered with every other cell and took the average of the non-zero values as the co-clustering index (CCI). A perfect CCI of 20 means that the cell is clustered with the same partners in every clustering instance, while lower values show deviations from this limit. We removed the cells with a CCI smaller than 6 and repeated this filtering procedure for three more iterations. The final results show a more stable clustering of the remaining 5101 cells. We then constructed a new neighborhood graph using newly computed principal components (*n_neighbors=10, n_pcs=15*), followed by Leiden clustering. The cluster centroids were calculated and correlated to the reference subclass centroids. We assigned clusters to their maximally correlated reference subclass if we could also see differential expression of their marker genes (scanpy's *rank_genes_groups*), otherwise we labeled them as NA. Of note, the DART-FISH population labeled as L6b/CT was highly correlated with reference subclasses L6b and L6 CT (Figure S4c) and showed expression of marker genes from both subclasses.

For inhibitory neurons and non-neuronal cells, the clustering was more stable to begin with, and we started by constructing the neighborhood matrix (For inhibitory neurons: *n_neighbors=20, n_pcs=10.* For non-neuronal cells: *n_neighbors=25, n_pcs=15*) and clustering. Then clusters were assigned to the reference subclass with maximum Pearson's correlation if the marker genes matched, or otherwise were labeled as NA.

### 5.4.2 Gene concordance analysis

The RNA portion of the SNARE-seq2 (snare) dataset from Bakken et al[75] and Plongthongkum et al 50 was used in this section. First, the snare data was subsetted to the DART-FISH genes. Then, DART-FISH and snare data were both normalized (*scanpy.pp.normalize_total(., target_sum=1000)*) followed by log-normalization (*scanpy.pp.log1p(.)*). The average normalized gene expression was calculated for all subclasses. For each gene, the concordance was defined as the Pearson's correlation between the average expressions across the subclasses between the DART-FISH and snare data (top panel of Supplementary Fig. 5c). The same analysis was performed for a MERFISH data set from Fang et al[77] (sample H18.06.006.MTG.250.expand.rep1)

with the following details: the subclass labels from metadata column "cluster_L2" were renamed to be consistent with DART-FISH annotations. In particular, subclasses L6b and L6 CT were merged, and subclass L5 ET was removed. Note that subclasses Sst Chodl, Chandelier and Lamp5 Lhx6 were not annotated in the MERFISH dataset and were removed from the DART-FISH analysis for consistency. The rest of the analysis was carried out with 242 shared genes between the datasets (bottom panel of Figure S4d).

### 5.4.3 Annotating the kidney data set

Cells with less than 5 and more than 100 transcripts were filtered (2024 out of 65565). The top 250 highly variable genes were kept for downstream analyses (*scanpy.experimental.pp.highly_variable_gene(., flavor='pearson_residuals')*). PCA was performed on pearson residuals, and the neighborhood graph was constructed (*scanpy.pp.neighbors(., n_neighbors=20, n_pcs=20, metric='cosine')*) followed by Leiden clustering (l1 clustering). From the kidney reference atlas 74, degenerative, cycling, transitioning and medullary cell types were removed. The counts were transformed to Pearson residuals and the remaining subclass level 1 and level 2 centroids were calculated. We then calculated the Pearson correlations between subclass level 1 centroids and cluster centroids and assigned each l1 cluster to the subclass level 1 with maximum correlation. We then subclustered each of the l1 clusters and assigned those to subclass level 2 identities with maximum correlation, only if the relevant marker genes were expressed. Through this procedure we could not resolve PT-S1 and PT-S2 subtypes separately; thus, we labeled the clusters that were highly correlated with these populations as PT-S1/S2. Similarly, for immune cells, this procedure could confidently resolve MAC-M2 cells and the general myeloid (IMM_Myl) and lymphoid (IMM_Lym) populations. To annotate the immune cells at higher level of granularity, we updated their subclass level 2 labels with the following strategy: Each DART-FISH cell with subclass level 1 label "IMM" was separately correlated with the following immune subtypes in the reference atlas: B, PL, T, MAC-M2, MDC, cDC. The immune subtypes with highest and 2nd highest correlation were kept. If the highest correlation was larger than 0.4 and the ratio of the

highest to the 2nd highest correlation was larger than 1.25, the label was updated to that of the highest correlated subtype, otherwise it remained unchanged.

### 5.4.4 Cell-cell interaction analysis

We used squidpy.gr.co_occurrence function (v1.2.4.dev27+gb644428) with n_splits=1 and an interval between 7um and 110um [83].

## 5.5 Acknowledgements

Chapter 5 is, in part, reprints of the material as it appears in Kalhor, K., Chen, C. J. ... & Zhang, K. (2024). Mapping human tissues with highly multiplexed RNA in situ hybridization. *Nature Communications*. The dissertation author was the primary investigator and co-first author of this paper.

# Appendix

# Supplementary information

**Table S1.** Oligo sequences for DART-FISH

| name | sequence | Usage |
|---|---|---|
| pAP1V41U | G*T*AGACTGGAAGAGCACTGTU | Amplification of kidney probe set |
| AP2V4 | /5Phos/TAGCCTCATGCGTATCCGAT | Amplification of kidney probe set |
| AP1V7U | A*A*GCAAGATTCTCGTCGAG/3deoxyU/ | Amplification of brain probe set |
| AP2V7 | /5Phos/TG TAA GGC ACA TCT CGG ATC | Amplification of brain probe set |
| RE_DpnII_V7N | GCACATCTCGGATCNNNN | Amplification of brain probe set |
| Acr_dc7-AF488_dT20 | /5Acryd/CATGGATTCGCGGAGGATCATTTTTTTTTTTTTTTTV*N | Reverse-transcription primer |
| Acr_dc10-Cy5_N9 | /5Acryd/CCGATAGTCACGATCTGTGGNNNNNNNN*N | Reverse-transcription primer |
| rca_primer | GATATCGGGAAGCTGA*A*G | RCA primer |
| DARTFISH_anchor_Cy3 | /5Cy3/CTTCAGCTTCCCGATATCCG | anchor probe |
| dcProbe7-AF488 | /5Alex488N/TGATCCTCCGCGAATCCATG | dT cDNA stain |
| dcProbe10-ATTO647N | /5ATTO647NN/CCACAGATCGTGACTATCGG | N9 cDNA stain |
| dcProbe0-AF488 | /5Alex488N/TGTATCGCGCTCGATTGGCA | decoding probe |
| dcProbe0-Cy3 | /5Cy3/CGTATCGGTAGTCGCAACGC | decoding probe |
| dcProbe0-ATTO647N | /5ATTO647NN/ACGCTACGGAGTACGCCACT | decoding probe |
| dcProbe1-AF488 | /5Alex488N/TCTTGCGTGCGATACGGAGT | decoding probe |
| dcProbe1-Cy3 | /5Cy3/AACGGTATTCGGTCGTCATC | decoding probe |
| dcProbe1-ATTO647N | /5ATTO647NN/CTGGTTCGGGCGTACCTAAC | decoding probe |
| dcProbe2-AF488 | /5Alex488N/AGAACTTGCGCGGATACACG | decoding probe |
| dcProbe2-Cy3 | /5Cy3/CTACTTCGTCGCGTCAGACC | decoding probe |
| dcProbe2-ATTO647N | GACGAACGGTCGAGATTTAC/3ATTO647NN/ | decoding probe |
| dcProbe3-AF488 | /5Alex488N/GAATTGTCCGCGCTCTACGA | decoding probe |
| dcProbe3-Cy3_2 | /5Cy3/TCGTACTTCGACGGCACTCA | decoding probe |
| dcProbe3-ATTO647N | /5ATTO647NN/AACTGCGACCGTCGGCTTAC | decoding probe |
| dcProbe4-AF488 | /5Alex488N/CGGAATACGTCGTTGACTGC | decoding probe |
| dcProbe4-Cy3 | /5Cy3/TACCATTCGCGTGCGATTCC | decoding probe |
| dcProbe4-ATTO647N_2 | /5ATTO647NN/ACTCTACCGGCAATCGCGTC | decoding probe |
| dcProbe5-AF488 | /5Alex488N/GAGTGTCGCGCAACTTAGCG | decoding probe |
| dcProbe5-Cy3 | /5Cy3/ACGTCTGCGTACCGGCTTAG | decoding probe |
| dcProbe5-ATTO647N | /5ATTO647NN/CATGCGATTAACCGCGACTG | decoding probe |
| dcProbe6-AF488_2 | /5Alex488N/CTTGCGGCGACAGTCGAACA | decoding probe |
| dcProbe6-Cy3 | /5Cy3/TCGTAACCCGTGCGAAGTGC | decoding probe |
| dcProbe6-ATTO647N | /5ATTO647NN/CTCTCGTAGCGTGCGATGAG | decoding probe |

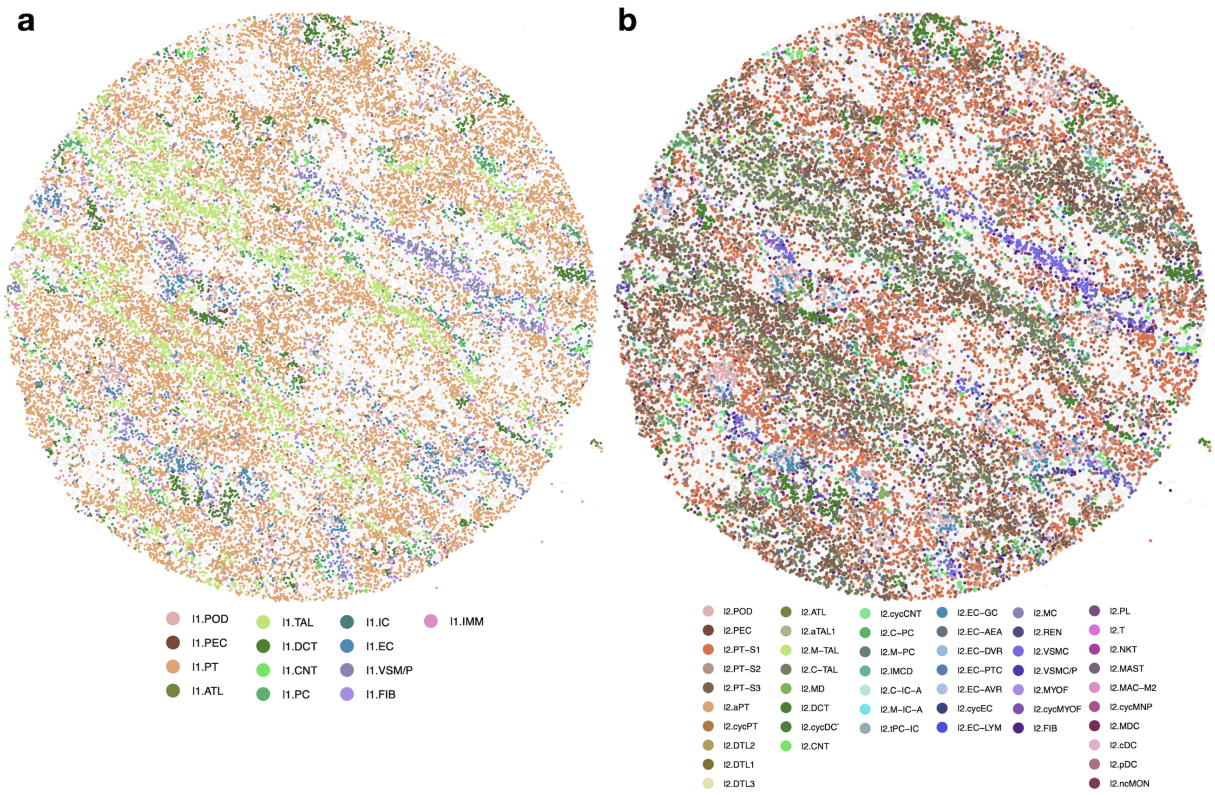



**Figure S1. Annotation of slide-seq data. a.** Annotation of a slide-seq puck at subclass level 1 using RCTD [31]. **b.** Annotation of a slide-seq puck at subclass level 2 using RCTD.

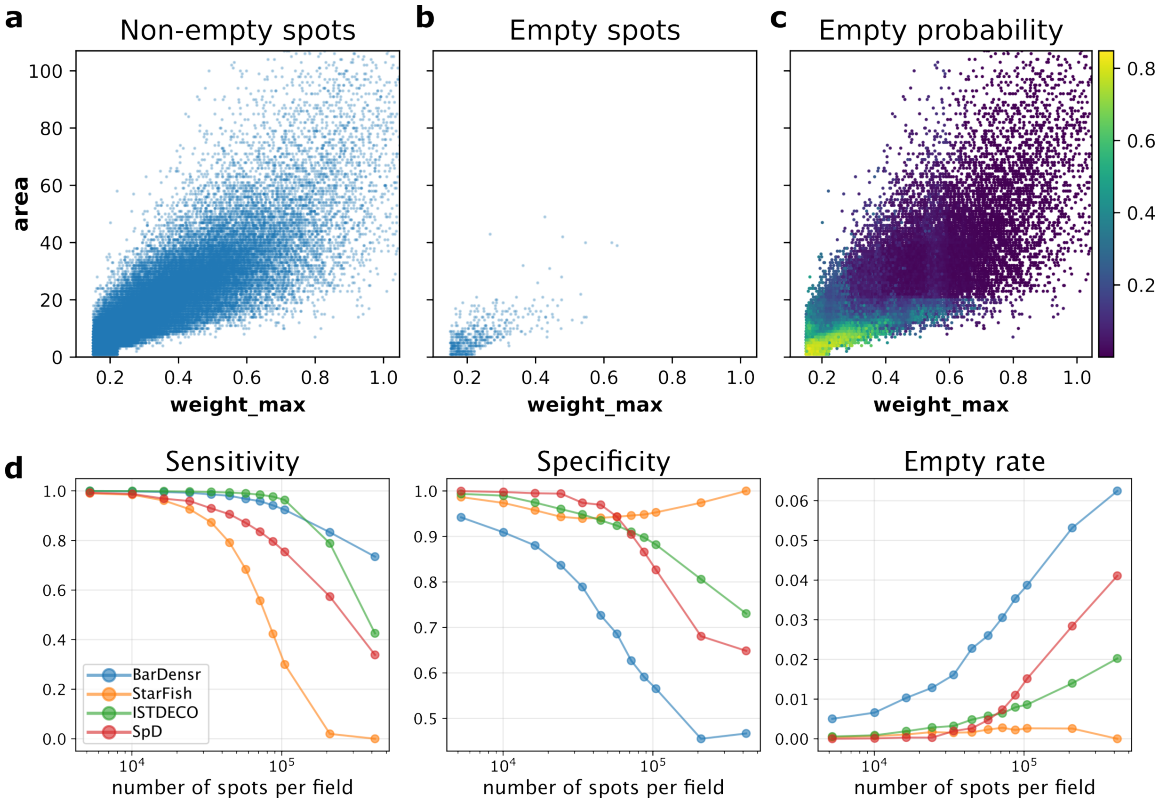



**Figure S2. Quality control and performance of SpD. a-c.** Scatter plots of two main features extracted from segmented spots with valid barcodes representing genes (a) or empty (unused) barcodes (b). Empty barcodes tend to be smaller in area and have lower weights than valid barcodes. (c) Emptiness probabilities inferred from a random forest that was trained to distinguish empty from non-empty spots based on the extracted features (maximum weight, average weight, area, ...). A cutoff is later set on the empty probabilities to keep high quality spots. **d.** Comparison of SpD with StarFish (naive matching), BarDensr and ISTDECO (deconvolution-based methods) on synthetic images with varying degrees of difficulty.

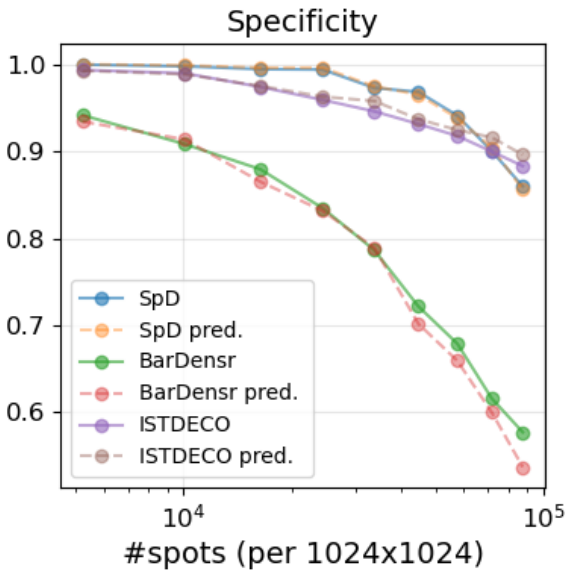



**Figure S3. Estimating specificity from empty rate** The layout is similar to figure S2b. The dashed line represent the estimation of specificity using empty rate in equation 4.3.

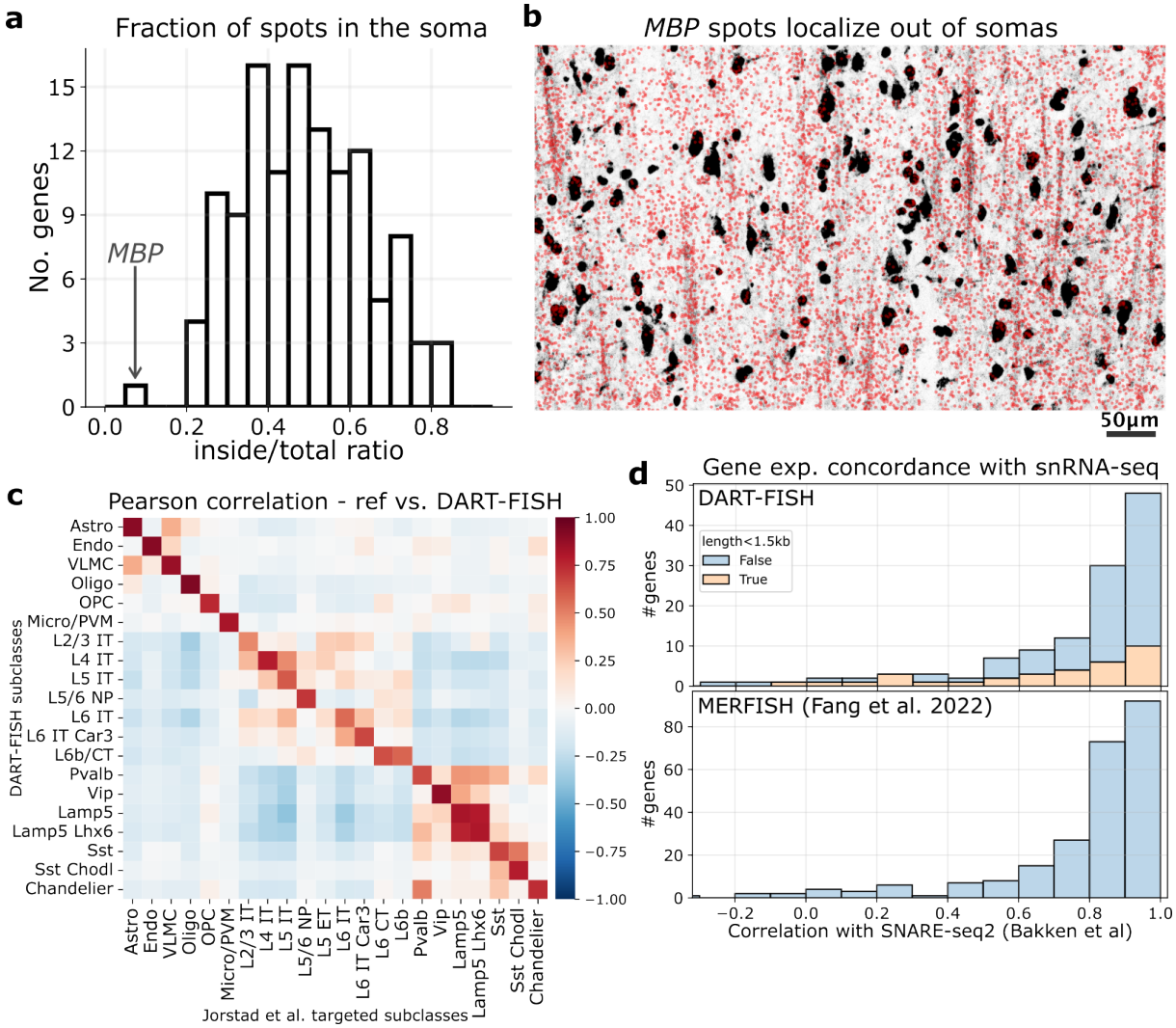



**Figure S4. Cell annotation accuracy in human M1C. (a)** Histogram showing the fraction of spots inside the segmented cells for each gene. *MBP* encoding Myelin basic protein has the lowest fraction of spots inside the cells. **(b)** An example of *MBP* being expressed outside the soma. Every red dot is a decoded *MBP* transcript on the background of RiboSoma (contrast is increased for clarity). *MBP* spots seem to co-localize with the RiboSoma signal over long threads that resemble axons. **(c)** Pearson's correlation of DART-FISH subclasses with the snRNA-seq reference subclasses used for annotation1 **(d)** Histogram of concordance values for genes in DART-FISH (top) and MERFISH (bottom, sample H18.06.006.MTG.250.expand.rep1[77]). Concordance is defined as the Pearson's correlation of expression levels across subclasses between SNARE-seq2[75] and the spatial assay. The histogram for DART-FISH is color coded to show the performance of short genes (constitutive exon length <1.5kb)

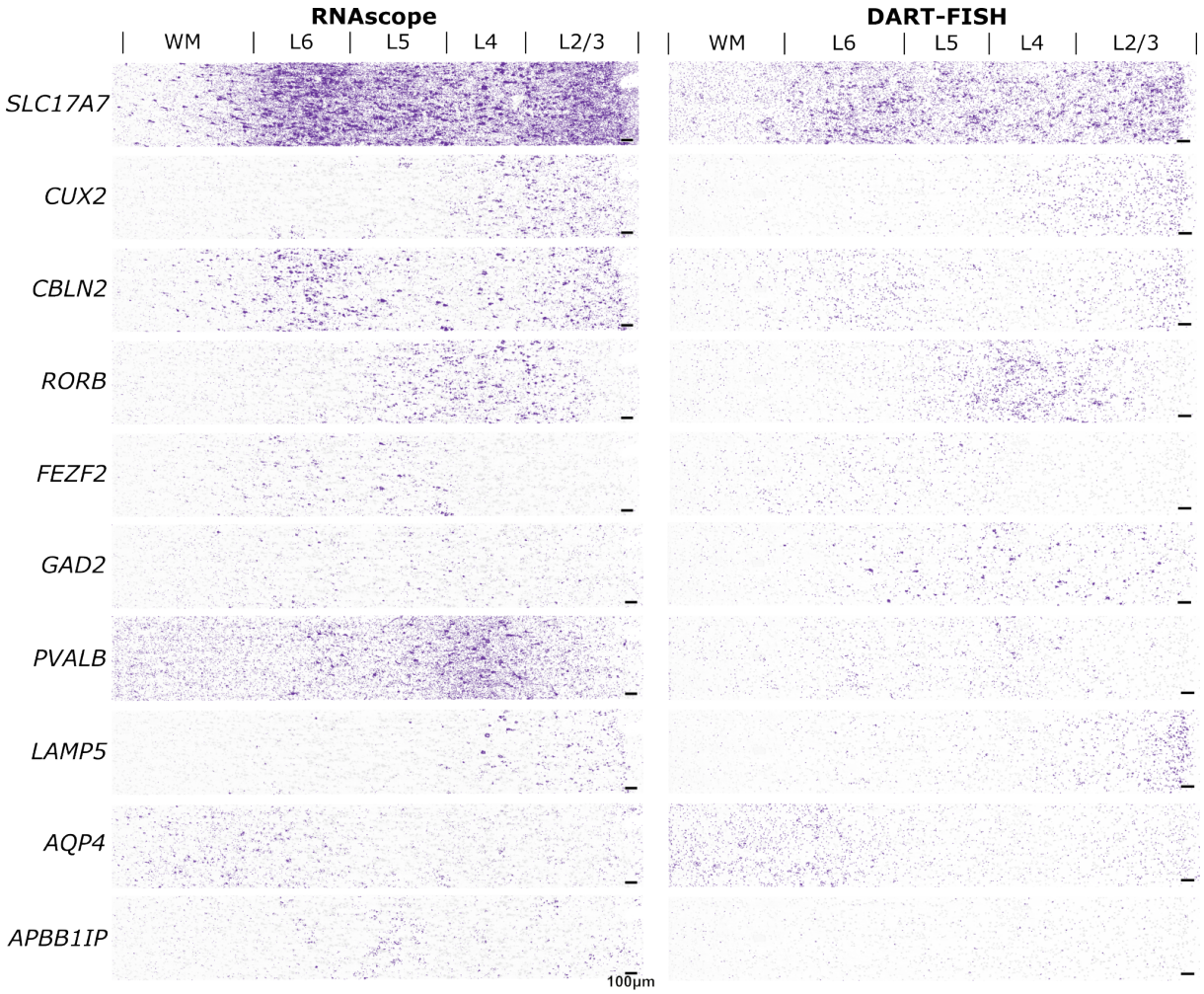
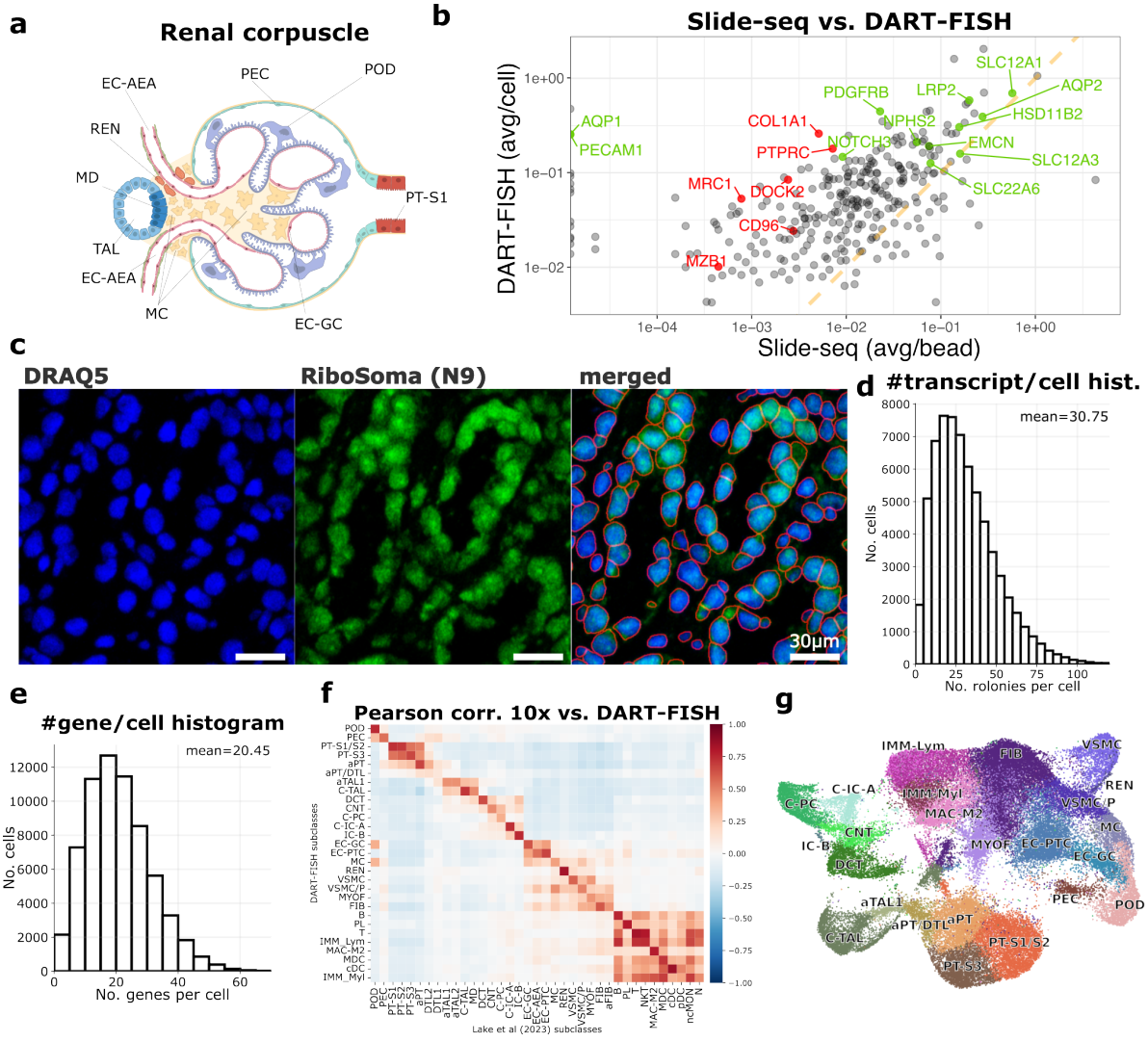



**Figure S5. Validation of DART-FISH by RNAscope**. Spatial distribution of 10 marker genes across the cortical layers measured by RNAscope (left) and DART-FISH (right). Scale bar 100um.

**Figure S6. Human kidney segmentation and annotation (a)** Diagram of the cell types composing the renal corpuscle and the juxtaglomerular apparatus [97]. **(b)** Scatter plot comparing average gene counts per bead in Slide-seq (Puck_200903_06 from a healthy patient[27] with average counts per cell in DART-FISH (Pearson's $r = 0.609$). Green dots represent canonical cell type markers while red dots are immune markers, suggesting high inflammation in the DART-FISH samples. The orange line indicates equal average counts across the two technologies. The top 150 highly expressed genes in slide-seq had on median 2.2x lower average transcripts per bead than average transcripts per cell in DART-FISH. **(c)** RiboSoma (randomly primed cDNA, middle) resolves tubular morphology better than the nuclear stain (left) and enhances cell segmentation (right). **(d)** Histogram of the number of rolonies per cell in >65,000 cells. There are on average 30 decoded transcripts per cell. **(e)** Histogram of the number of detected genes per cell in the kidney, averaging at 20 unique genes per cell. **(f)** Pearson's correlation of average DART-FISH subclasses with the average snRNA-seq reference subclasses used for annotation [27]. **(g)** UMAP of all annotated subclasses. PEC: parietal epithelial cells, aPT: altered proximal tubules, DTL: descending thin limbs, aTAL: altered thick ascending limbs, DCT: distal convoluted tubules, CNT: connecting tubules, C-IC-A: cortical intercalated cell type A, IC-B: intercalated cell type B, EC-PTC: peritubular capillary endothelial cell, MC: mesangial cell, REN: renin-positive juxtaglomerular granular cell, VSMC: vascular smooth muscle cell, VSMC/P: vascular smooth muscle cell/pericyte, FIB: fibroblast, MYOF: Myofibroblast, MAC-M2: M2 macrophage, IMM-Lym: lymphoid cell, IMM-Myl: myeloid cell.


a
Renal corpuscle
EC-AEA
PEC
POD
REN
MD
PT-S1
TAL
EC-AEA
MC
EC-GC
b
Slide-seq vs. DART-FISH
DART-FISH (avg/cell)
Slide-seq (avg/bead)
1e+00
1e−01
1e−02
1e−04
1e−03
1e−02
1e−01
1e+00
SLC12A1
AQP2
PDGFRB
LRP2
HSD11B2
AQP1
COL1A1
NPHS2
PECAM1
NOTCH3
EMCN
PTPRC
SLC12A3
MRC1
DOCK2
SLC22A6
CD96
MZB1
c
DRAQ5
RiboSoma (N9)
merged
30µm
d
#transcript/cell hist.
mean=30.75
No. cells
No. rolonies per cell
e
#gene/cell histogram
mean=20.45
No. cells
No. genes per cell
f
Pearson corr. 10x vs. DART-FISH
DART-FISH subclasses
Lake et al (2023) subclasses
g
IMM-Lym
FIB
VSMC
C-IC-A
IMM-Myl
REN
C-PC
MAC-M2
VSMC/P
CNT
MC
IC-B
MYOF
EC-PTC
EC-GC
DCT
PEC
POD
aTAL1
aPT
aPT/DTL
C-TAL
PT-S1/S2
PT-S3

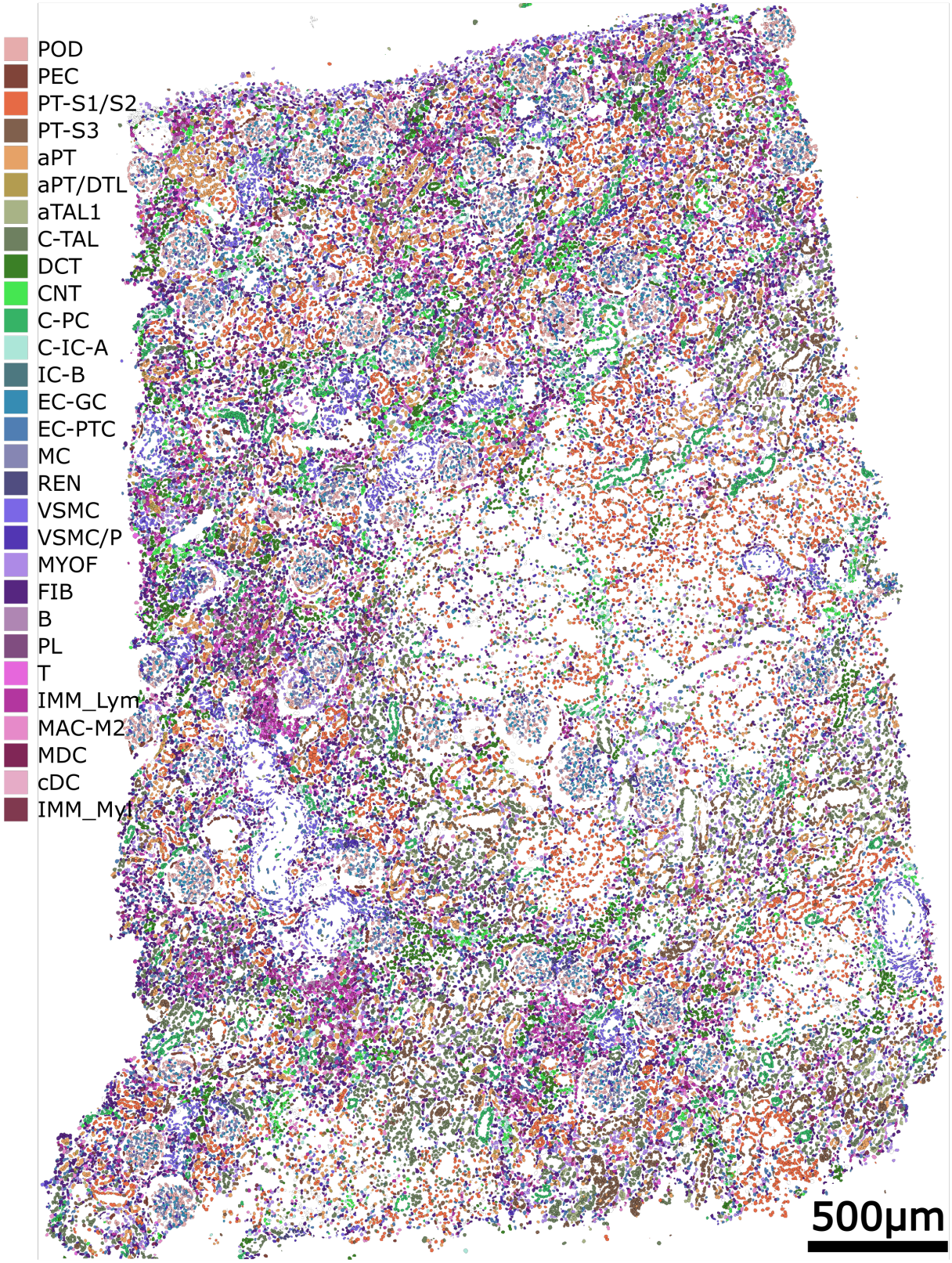



**Figure S7.** Annotated cells in the human kidney

# Bibliography


[1] Christina Karlsson Rosenthal. The beginning: Invention of the microscope. *Nature Cell Biology*, 11(S1):S6–S6, October 2009.

[2] Nicole Rusk. The fluorescence microscope: First fluorescence microscope, First epifluorescence microscope, The dichroic mirror. *Nature Cell Biology*, 11(S1):S8–S9, October 2009.

[3] Christiaan Van Ooij. Recipe for fluorescent antibodies: Immunofluorescence. *Nature Cell Biology*, 11(S1):S10–S11, October 2009.

[4] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M. D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M.

Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie vandeVondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, November 2008.

[5] Sarah B. Ng, Emily H. Turner, Peggy D. Robertson, Steven D. Flygare, Abigail W. Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E. Eichler, Michael Bamshad, Deborah A. Nickerson, and Jay Shendure. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, September 2009.

[6] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.

[7] Peter J. Park. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, October 2009.

[8] Wendy A. Bickmore. The Spatial Organization of the Human Genome. *Annual Review of Genomics and Human Genetics*, 14(1):67–84, August 2013.

[9] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure, and Michael J Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1):30–35, January 2010.

[10] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, May 2012.

[11] Kinga Matuła, Francesca Rivello, and Wilhelm T. S. Huck. Single-Cell Analysis Using Droplet Microfluidics. *Advanced Biosystems*, 4(1):1900188, January 2020.

[12] Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. Multiplex

single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, May 2015.

[13] Chengxiang Qiu, Beth K. Martin, Ian C. Welsh, Riza M. Daza, Truc-Mai Le, Xingfan Huang, Eva K. Nichols, Megan L. Taylor, Olivia Fulton, Diana R. O'Day, Anne Roshella Gomes, Saskia Ilcisin, Sanjay Srivatsan, Xinxian Deng, Christine M. Disteche, William Stafford Noble, Nobuhiko Hamazaki, Cecilia B. Moens, David Kimelman, Junyue Cao, Alexander F. Schier, Malte Spielmann, Stephen A. Murray, Cole Trapnell, and Jay Shendure. A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature*, 626(8001):1084–1093, February 2024.

[14] Ian H. De Boer, Charles E. Alpers, Evren U. Azeloglu, Ulysses G.J. Balis, Jonathan M. Barasch, Laura Barisoni, Kristina N. Blank, Andrew S. Bomback, Keith Brown, Pierre C. Dagher, Ashveena L. Dighe, Michael T. Eadon, Tarek M. El-Achkar, Joseph P. Gaut, Nir Hacohen, Yongqun He, Jeffrey B. Hodgin, Sanjay Jain, John A. Kellum, Krzysztof Kiryluk, Richard Knight, Zoltan G. Laszik, Chrysta Lienczewski, Laura H. Mariani, Robyn L. McClelland, Steven Menez, Dennis G. Moledina, Sean D. Mooney, John F. O'Toole, Paul M. Palevsky, Chirag R. Parikh, Emilio D. Poggio, Sylvia E. Rosas, Matthew R. Rosengart, Minnie M. Sarwal, Jennifer A. Schaub, John R. Sedor, Kumar Sharma, Becky Steck, Robert D. Toto, Olga G. Troyanskaya, Katherine R. Tuttle, Miguel A. Vazquez, Sushrut S. Waikar, Kayleen Williams, Francis Perry Wilson, Kun Zhang, Ravi Iyengar, Matthias Kretzler, Jonathan Himmelfarb, Richard Knight, Stewart Lecker, Isaac Stillman, Sushrut Waikar, Gearoid Mcmahon, Astrid Weins, Samuel Short, Nir Hacohen, Paul Hoover, Mark Aulisio, Leslie Cooperman, Leal Herlitz, John O'Toole, Emilio Poggio, John Sedor, Stacey Jolly, Paul Appelbaum, Olivia Balderes, Jonathan Barasch, Andrew Bomback, Pietro A. Canetta, Vivette D. d'Agati, Krzysztof Kiryluk, Satoru Kudose, Karla Mehl, Jai Radhakrishnan, Chenhua Weng, Laura Barisoni, Theodore Alexandrov, Tarek Ashkar, Daria Barwinska, Pierre Dagher, Kenneth Dunn, Michael Eadon, Michael Ferkowicz, Katherine Kelly, Timothy Sutton, Seth Winfree, Steven Menez, Chirag Parikh, Avi Rosenberg, Pam Villalobos, Rubab Malik, Derek Fine, Mohammed Atta, Jose Manuel Monroy Trujillo, Alison Slack, Sylvia Rosas, Mark Williams, Evren Azeloglu, Cijang (John) He, Ravi Iyengar, Jens Hansen, Samir Parikh, Brad Rovin, Chris Anderton, Ljiljana Pasa-Tolic, Dusan Velickovic, Jessica Lukowski, George (Holt) Oliver, Joseph Ardayfio, Jack Bebiak, Keith Brown, Taneisha Campbell, Catherine Campbell, Lynda Hayashi, Nichole Jefferson, Robert Koewler, Glenda Roberts, John Saul, Anna Shpigel, Edith Christine Stutzke, Lorenda Wright, Leslie Miegs, Roy Pinkeney, Rachel Sealfon, Olga Troyanskaya, Katherine Tuttle, Dejan Dobi, Yury Goltsev, Blue Lake, Kun Zhang, Maria Joanes, Zoltan Laszik, Andrew Schroeder, Minnie Sarwal, Tara Sigdel, Ulysses Balis, Victoria Blanc, Oliver He, Jeffrey Hodgin, Matthias Kretzler, Laura Mariani, Rajasree Menon, Edgar Otto, Jennifer Schaub, Becky Steck, Chrysta Lienczewski, Sean Eddy, Michele Elder, Daniel Hall, John Kellum, Mary Kruth, Raghav Murugan, Paul Palevsky, Parmjeet Randhawa, Matthew Rosengart, Sunny Sims-Lucas, Mary Stefanick, Stacy Stull, Mitchell Tublin, Charles Alpers, Ian De Boer, Ashveena Dighe, Jonathan Himmelfarb, Robyn Mcclelland, Sean Mooney, Stuart Shankland, Kayleen Williams, Kristina Blank, Jonas Carson, Frederick Dowd, Zach Drager,

Christopher Park, Kumar Sharma, Guanshi Zhang, Shweta Bansal, Manjeri Venkatachalam, Asra Kermani, Simon Lee, Christopher Lu, Tyler Miller, Orson Moe, Harold Park, Kamalanathan Sambandam, Francisco Sanchez, Jose Torrealba, Toto Robert, Miguel Vazquez, Nancy Wang, Joe Gaut, Sanjay Jain, Anitha Vijayan, Randy Luciano, Dennis Moledina, Ugwuowo Ugochukwu, Francis Perry Wilson, and Sandy Alfano. Rationale and design of the Kidney Precision Medicine Project. *Kidney International*, 99(3):498–510, March 2021.

[15] Jeffrey R. Moffitt, Emma Lundberg, and Holger Heyn. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics*, 23(12):741–759, December 2022.

[16] Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, 10(9):857–860, September 2013.

[17] Je Hyuk Lee, Evan R. Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L. Yang, Thomas C. Ferrante, Richard Terry, Sauveur S. F. Jeanty, Chao Li, Ryoji Amamoto, Derek T. Peters, Brian M. Turczyk, Adam H. Marblestone, Samuel A. Inverso, Amy Bernard, Prashant Mali, Xavier Rios, John Aach, and George M. Church. Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science*, 343(6177):1360–1363, March 2014.

[18] Michael G. Mohsen and Eric T. Kool. The Discovery of Rolling Circle Amplification and Rolling Circle Transcription. *Accounts of Chemical Research*, 49(11):2540–2550, November 2016.

[19] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, April 2015.

[20] Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*, 11(4):360–361, April 2014.

[21] Arjun Raj, Patrick Van Den Bogaard, Scott A Rifkin, Alexander Van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, October 2008.

[22] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, July 2016.

[23] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, March 2019.

[24] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, September 2019.

[25] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela J. Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, 39(3):313–319, March 2021.

[26] David W. McKellar, Madhav Mantri, Meleana M. Hinchman, John S. L. Parker, Praveen Sethupathy, Benjamin D. Cosgrove, and Iwijn De Vlaminck. Spatial mapping of the total transcriptome by in situ polyadenylation. *Nature Biotechnology*, 41(4):513–520, April 2023.

[27] Blue B. Lake, Rajasree Menon, Seth Winfree, Qiwen Hu, Ricardo Melo Ferreira, Kian Kalhor, Daria Barwinska, Edgar A. Otto, Michael Ferkowicz, Dinh Diep, Nongluk Plongthongkum, Amanda Knoten, Sarah Urata, Laura H. Mariani, Abhijit S. Naik, Sean Eddy, Bo Zhang, Yan Wu, Diane Salamon, James C. Williams, Xin Wang, Karol S. Balderrama, Paul J. Hoover, Evan Murray, Jamie L. Marshall, Teia Noel, Anitha Vijayan, Austin Hartman, Fei Chen, Sushrut S. Waikar, Sylvia E. Rosas, Francis P. Wilson, Paul M. Palevsky, Krzysztof Kiryluk, John R. Sedor, Robert D. Toto, Chirag R. Parikh, Eric H. Kim, Rahul Satija, Anna Greka, Evan Z. Macosko, Peter V. Kharchenko, Joseph P. Gaut, Jeffrey B. Hodgin, KPMP Consortium, Richard Knight, Stewart H. Lecker, Isaac Stillman, Afolarin A. Amodu, Titlayo Ilori, Shana Maikhor, Insa Schmidt, Gearoid M. McMahon, Astrid Weins, Nir Hacohen, Lakeshia Bush, Agustin Gonzalez-Vicente, Jonathan Taliercio, John O'toole, Emilio Poggio, Leslie Cooperman, Stacey Jolly, Leal Herlitz, Jane Nguyen, Ellen Palmer, Dianna Sendrey, Kassandra Spates-Harden, Paul Appelbaum, Jonathan M. Barasch, Andrew S. Bomback, Vivette D. D'Agati, Karla Mehl, Pietro A. Canetta, Ning Shang, Olivia Balderes, Satoru Kudose, Laura Barisoni, Theodore Alexandrov, Yinghua Cheng, Kenneth W. Dunn, Katherine J. Kelly, Timothy A. Sutton, Yumeng Wen, Celia P. Corona-Villalobos, Steven Menez, Avi Rosenberg, Mohammed Atta, Camille Johansen, Jennifer Sun, Neil Roy, Mark Williams, Evren U. Azeloglu, Cijang He, Ravi Iyengar, Jens Hansen, Yuguang Xiong, Brad Rovin, Samir Parikh, Sethu M. Madhavan, Christopher R. Anderton, Ljiljana Pasa-Tolic, Dusan Velickovic, Olga Troyanskaya, Rachel Sealfon, Katherine R. Tuttle, Zoltan G. Laszik, Garry Nolan, Minnie Sarwal, Kavya Anjani, Tara Sigdel, Heather Ascani, Ulysses G. J. Balis, Chrysta Lienczewski, Becky Steck, Yougqun He, Jennifer Schaub, Victoria M. Blanc, Raghavan Murugan, Parmjeet Randhawa, Matthew Rosengart, Mitchell Tublin, Tina Vita, John A. Kellum, Daniel E. Hall, Michele M. Elder, James Winters, Matthew Gilliam, Charles E. Alpers, Kristina N. Blank, Jonas Carson, Ian H. De Boer, Ashveena L. Dighe, Jonathan Himmelfarb, Sean D. Mooney, Stuart Shankland, Kayleen Williams, Christopher Park, Frederick Dowd, Robyn L. McClelland, Stephen Daniel, Andrew N. Hoofnagle, Adam Wilcox, Shweta Bansal, Kumar Sharma, Manjeri Venkatachalam, Guanshi Zhang, Annapurna Pamreddy, Vijaykumar R. Kakade, Dennis

Moledina, Melissa M. Shaw, Ugochukwu Ugwuowo, Tanima Arora, Joseph Ardayfio, Jack Bebiak, Keith Brown, Catherine E. Campbell, John Saul, Anna Shpigel, Christy Stutzke, Robert Koewler, Taneisha Campbell, Lynda Hayashi, Nichole Jefferson, Roy Pinkeney, Glenda V. Roberts, Michael T. Eadon, Pierre C. Dagher, Tarek M. El-Achkar, Kun Zhang, Matthias Kretzler, and Sanjay Jain. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970):585–594, July 2023.

[28] Song Chen, Blue B. Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, December 2019.

[29] Nongluk Plongthongkum, Dinh Diep, Song Chen, Blue B. Lake, and Kun Zhang. Scalable dual-omics profiling with single-nucleus chromatin accessibility and mRNA expression sequencing 2 (SNARE-seq2). *Nature Protocols*, 16(11):4992–5029, November 2021.

[30] Yingkun Zhang, Xinrui Lin, Zhixian Yao, Di Sun, Xin Lin, Xiaoyu Wang, Chaoyong Yang, and Jia Song. Deconvolution algorithms for inference of the cell-type composition of the spatial transcriptome. *Computational and Structural Biotechnology Journal*, 21:176–184, 2023.

[31] Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandrina Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4):517–526, April 2022.

[32] Lise Bankir, Lucile Figueres, Caroline Prot-Bertoye, Nadine Bouby, Gilles Crambert, J. Howard Pratt, and Pascal Houillier. Medullary and cortical thick ascending limb: similarities and differences. *American Journal of Physiology-Renal Physiology*, 318(2):F422–F442, February 2020.

[33] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, Rani E. George, Nico Pierson, Long Cai, and Guo-Cheng Yuan. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1):78, December 2021.

[34] Donald D Newmeyer and Shelagh Ferguson-Miller. Mitochondria. *Cell*, 112(4):481–490, February 2003.

[35] Barry M. Brenner and Floyd C. Rector, editors. *Brenner & Rector's the kidney*. Saunders Elsevier, Philadelphia, 8th ed edition, 2008. OCLC: ocm72774314.

[36] Adina R. Buxbaum, Gal Haimovich, and Robert H. Singer. In the right place at the right time: visualizing and understanding mRNA localization. *Nature Reviews Molecular Cell Biology*, 16(2):95–109, February 2015.

[37] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, March 2019.

[38] A.C. Croce and G. Bottiroli. Autofluorescence spectroscopy and imaging: a tool for biomedical research and diagnosis. *European Journal of Histochemistry*, December 2014.

[39] B. Banerjee, B. E. Miedema, and H. R. Chandrasekhar. Role of basement membrane collagen and elastin in the autofluorescence spectra of the colon. *Journal of Investigative Medicine: The Official Publication of the American Federation for Clinical Research*, 47(6):326–332, July 1999.

[40] Robim M. Rodrigues, Peter Macko, Taina Palosaari, and Maurice P. Whelan. Autofluorescence microscopy: A non-destructive tool to monitor mitochondrial toxicity. *Toxicology Letters*, 206(3):281–288, October 2011.

[41] Mats Nilsson, Helena Malmgren, Martina Samiotaki, Marek Kwiatkowski, Bhanu P. Chowdhary, and Ulf Landegren. Padlock Probes: Circularizing Oligonucleotides for Localized DNA Detection. *Science*, 265(5181):2085–2088, September 1994.

[42] Paul M. Lizardi, Xiaohua Huang, Zhengrong Zhu, Patricia Bray-Ward, David C. Thomas, and David C. Ward. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nature Genetics*, 19(3):225–232, July 1998.

[43] Xiaoyin Chen, Yu-Chi Sun, George M Church, Je Hyuk Lee, and Anthony M Zador. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Research*, 46(4):e22–e22, February 2018.

[44] Xiaoyan Qian, Kenneth D. Harris, Thomas Hauling, Dimitris Nicoloutsopoulos, Ana B. Muñoz-Manchado, Nathan Skene, Jens Hjerling-Leffler, and Mats Nilsson. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nature Methods*, 17(1):101–106, January 2020.

[45] Kian Kalhor, Chien-Ju Chen, Ho Suk Lee, Matthew Cai, Mahsa Nafisi, Richard Que, Carter R. Palmer, Yixu Yuan, Yida Zhang, Xuwen Li, Jinghui Song, Amanda Knoten, Blue B. Lake, Joseph P. Gaut, C. Dirk Keene, Ed Lein, Peter V. Kharchenko, Jerold Chun, Sanjay Jain, Jian-Bing Fan, and Kun Zhang. Mapping human tissues with highly multiplexed RNA in situ hybridization. *Nature Communications*, 15(1):2511, March 2024.

[46] Ho Suk Lee. *Biomolecular Processing and Quantitative Localization of Single Cell Analysis*. PhD thesis, UC San Diego, 2015.

[47] Matthew Cai. *Spatial mapping of single cells in human cerebral cortex using DARTFISH: A highly multiplexed method for in situ quantification of targeted RNA transcripts*. PhD thesis, UC San Diego, 2019.

[48] Kevin L. Gunderson, Semyon Kruglyak, Michael S. Graige, Francisco Garcia, Bahram G. Kermani, Chanfeng Zhao, Diping Che, Todd Dickinson, Eliza Wickham, Jim Bierle, Dennis Doucet, Monika Milewski, Robert Yang, Chris Siegmund, Juergen Haas, Lixin Zhou, Arnold Oliphant, Jian-Bing Fan, Steven Barnard, and Mark S. Chee. Decoding Randomly Ordered DNA Arrays. *Genome Research*, 14(5):870–877, May 2004.

[49] Daniel Gyllborg, Christoffer Mattsson Langseth, Xiaoyan Qian, Eunkyoung Choi, Sergio Marco Salas, Markus M Hilscher, Ed S Lein, and Mats Nilsson. Hybridization-based *in situ* sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. *Nucleic Acids Research*, 48(19):e112–e112, November 2020.

[50] Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):eaat5691, July 2018.

[51] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, 10(3):442–458, March 2015.

[52] Fei Chen, Asmamaw T Wassie, Allison J Cote, Anubhav Sinha, Shahar Alon, Shoh Asano, Evan R Daugharthy, Jae-Byum Chang, Adam Marblestone, George M Church, Arjun Raj, and Edward S Boyden. Nanoscale imaging of RNA with expansion microscopy. *Nature Methods*, 13(8):679–684, August 2016.

[53] Guiping Wang, Jeffrey R. Moffitt, and Xiaowei Zhuang. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Scientific Reports*, 8(1):4847, March 2018.

[54] Shahar Alon, Daniel R. Goodwin, Anubhav Sinha, Asmamaw T. Wassie, Fei Chen, Evan R. Daugharthy, Yosuke Bando, Atsushi Kajita, Andrew G. Xue, Karl Marrett, Robert Prior, Yi Cui, Andrew C. Payne, Chun-Chen Yao, Ho-Jun Suk, Ru Wang, Chih-Chieh (Jay) Yu, Paul Tillberg, Paul Reginato, Nikita Pak, Songlei Liu, Sukanya Punthambaker, Eswar P. R. Iyer, Richie E. Kohman, Jeremy A. Miller, Ed S. Lein, Ana Lako, Nicole Cullen, Scott Rodig, Karla Helvie, Daniel L. Abravanel, Nikhil Wagle, Bruce E. Johnson, Johanna Klughammer, Michal Slyper, Julia Waldman, Judit Jané-Valbuena, Orit Rozenblatt-Rosen, Aviv Regev, IMAXT Consortium, George M. Church, Adam H. Marblestone, Edward S. Boyden, H. R. Ali, M. Al Sa'd, S. Alon, S. Aparicio, G. Battistoni, S. Balasubramanian, R. Becker, B. Bodenmiller, E. S. Boyden, D. Bressan, A. Bruna, Marcel Burger, C. Caldas, M. Callari, I. G. Cannell, H. Casbolt, N. Chornay, Y. Cui, A. Dariush, K. Dinh, A. Emenari, Y. Eyal-Lubling, J. Fan, A. Fatemi, E. Fisher, E. A. González-Solares, C. González-Fernández, D. Goodwin, W. Greenwood, F. Grimaldi, G. J. Hannon, O. Harris, S. Harris, C. Jauset, J. A. Joyce, E. D. Karagiannis, T. Kovačević, L. Kuett, R. Kunes, A. Küpcü Yoldaş, D. Lai, E. Laks, H. Lee, M. Lee, G. Lerda, Y. Li, A. McPherson, N. Millar, C. M. Mulvey, F. Nugent, C. H. O'Flanagan, M. Paez-Ribes, I. Pearsall, F. Qosaj, A. J. Roth, O. M. Rueda, T. Ruiz, K. Sawicka, L. A. Sepúlveda, S. P. Shah, A. Shea, A. Sinha, A. Smith, S. Tavaré, S. Tietscher, I. Vázquez-García, S. L. Vogl, N. A. Walton, A. T. Wassie, S. S. Watson, J. Weselak, S. A. Wild, E. Williams, J. Windhager, T. Whitmarsh, C. Xia, P. Zheng, and X. Zhuang. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*, 371(6528):eaax2656, January 2021.

[55] Jin Billy Li, Yuan Gao, John Aach, Kun Zhang, Gregory V. Kryukov, Bin Xie, Annika Ahlford, Jung-Ki Yoon, Abraham M. Rosenbaum, Alexander Wait Zaranek, Emily LeProust, Shamil R. Sunyaev, and George M. Church. Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Research*, 19(9):1606–1615, September 2009.

[56] Yu-Chi Sun, Xiaoyin Chen, Stephan Fischer, Shaina Lu, Huiqing Zhan, Jesse Gillis, and Anthony M. Zador. Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections. *Nature Neuroscience*, May 2021.

[57] Dinh Diep, Nongluk Plongthongkum, Athurva Gore, Ho-Lim Fung, Robert Shoemaker, and Kun Zhang. Library-free methylation sequencing with bisulfite padlock probes. *Nature Methods*, 9(3):270–272, March 2012.

[58] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, May 2013. arXiv:1303.3997 [q-bio].

[59] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 574–582, Las Vegas, NV, USA, June 2016. IEEE.

[60] Shannon Axelrod, Matthew Cai, Ambrose Carr, Jeremy Freeman, Deep Ganguli, Justin Kiggins, Brian Long, Tony Tung, and Kevin Yamauchi. starfish: scalable pipelines for image-based transcriptomics. *Journal of Open Source Software*, 6(61):2440, May 2021.

[61] Cecilia Cisar, Nicholas Keener, Mathew Ruffalo, and Benedict Paten. A unified pipeline for FISH spatial transcriptomics. *Cell Genomics*, 3(9):100384, September 2023.

[62] Arthur Imbert, Wei Ouyang, Adham Safieddine, Emeline Coleno, Christophe Zimmer, Edouard Bertrand, Thomas Walter, and Florian Mueller. FISH-quant v2: a scalable and modular tool for smFISH image analysis. *RNA*, 28(6):786–795, June 2022.

[63] Bastian Th Eichenberger, YinXiu Zhan, Markus Rempfler, Luca Giorgetti, and Jeffrey A Chao. deepBlink: threshold-independent detection and localization of diffraction-limited spots. *Nucleic Acids Research*, 49(13):7292–7297, July 2021.

[64] Ella Bahry, Laura Breimann, Marwan Zouinkhi, Leo Epstein, Klim Kolyvanov, Nicholas Mamrak, Benjamin King, Xi Long, Kyle I. S. Harrington, Timothée Lionnet, and Stephan Preibisch. RS-FISH: precise, interactive, fast, and scalable FISH spot detection. *Nature Methods*, 19(12):1563–1567, December 2022.

[65] Fei Chen, Paul W. Tillberg, and Edward S. Boyden. Expansion microscopy. *Science*, 347(6221):543–548, January 2015.

[66] Axel Andersson, Ferran Diego, Fred A. Hamprecht, and Carolina Wählby. ISTDECO: In Situ Transcriptomics Decoding by Deconvolution. preprint, Bioinformatics, March 2021.

[67] Shuonan Chen, Jackson Loper, Xiaoyin Chen, Alex Vaughan, Anthony M. Zador, and Liam Paninski. BARcode DEmixing through Non-negative Spatial Regression (BarDensr). *PLOS Computational Biology*, 17(3):e1008256, March 2021.

[68] Lars E. Borm, Alejandro Mossi Albiach, Camiel C.A. Mannens, Jokubas Janusauskas, Ceren Özgün, David Fernández-García, Rebecca Hodge, Ed S. Lein, Simone Codeluppi, and Sten Linnarsson. Scalable *in situ* single-cell profiling by electrophoretic capture of mRNA. preprint, Neuroscience, January 2022.

[69] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, January 2021.

[70] Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature Methods*, 19(12):1634–1641, December 2022.

[71] Stephan Preibisch, Stephan Saalfeld, and Pavel Tomancak. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics*, 25(11):1463–1465, June 2009.

[72] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[73] Bradley Efron and Trevor Hastie. *Computer age statistical inference: algorithms, evidence, and data science*. Number 6 in Institute of Mathematical Statistics monographs. Cambridge University Press, Cambridge, United Kingdom New York, USA Port Melbourne, Australia New Delhi, India Singapore, student edition edition, 2021.

[74] Stéfan Van Der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, June 2014.

[75] Trygve E. Bakken, Nikolas L. Jorstad, Qiwen Hu, Blue B. Lake, Wei Tian, Brian E. Kalmbach, Megan Crow, Rebecca D. Hodge, Fenna M. Krienen, Staci A. Sorensen, Jeroen Eggermont, Zizhen Yao, Brian D. Aevermann, Andrew I. Aldridge, Anna Bartlett, Darren Bertagnolli, Tamara Casper, Rosa G. Castanon, Kirsten Crichton, Tanya L. Daigle, Rachel Dalley, Nick Dee, Nikolai Dembrow, Dinh Diep, Song-Lin Ding, Weixiu Dong, Rongxin Fang, Stephan Fischer, Melissa Goldman, Jeff Goldy, Lucas T. Graybuck, Brian R. Herb, Xiaomeng Hou, Jayaram Kancherla, Matthew Kroll, Kanan Lathia, Baldur Van Lew, Yang Eric Li, Christine S. Liu, Hanqing Liu, Jacinta D. Lucero, Anup Mahurkar, Delissa McMillen, Jeremy A. Miller, Marmar Moussa, Joseph R. Nery, Philip R. Nicovich, Sheng-Yong Niu, Joshua Orvis, Julia K. Osteen, Scott Owen, Carter R. Palmer, Thanh Pham, Nongluk Plongthongkum, Olivier Poirion, Nora M. Reed, Christine Rimorin, Angeline Rivkin, William J. Romanow, Adriana E. Sedeño-Cortés, Kimberly Siletti, Saroja Somasundaram, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Xinxin Wang, Fangming

Xie, Anna Marie Yanny, Renee Zhang, Seth A. Ament, M. Margarita Behrens, Hector Corrada Bravo, Jerold Chun, Alexander Dobin, Jesse Gillis, Ronna Hertzano, Patrick R. Hof, Thomas Höllt, Gregory D. Horwitz, C. Dirk Keene, Peter V. Kharchenko, Andrew L. Ko, Boudewijn P. Lelieveldt, Chongyuan Luo, Eran A. Mukamel, António Pinto-Duarte, Sebastian Preissl, Aviv Regev, Bing Ren, Richard H. Scheuermann, Kimberly Smith, William J. Spain, Owen R. White, Christof Koch, Michael Hawrylycz, Bosiljka Tasic, Evan Z. Macosko, Steven A. McCarroll, Jonathan T. Ting, Hongkui Zeng, Kun Zhang, Guoping Feng, Joseph R. Ecker, Sten Linnarsson, and Ed S. Lein. Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*, 598(7879):111–119, October 2021.

[76] Meng Zhang, Stephen W. Eichhorn, Brian Zingg, Zizhen Yao, Kaelan Cotter, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature*, 598(7879):137–143, October 2021.

[77] Rongxin Fang, Chenglong Xia, Jennie L. Close, Meng Zhang, Jiang He, Zhengkai Huang, Aaron R. Halpern, Brian Long, Jeremy A. Miller, Ed S. Lein, and Xiaowei Zhuang. Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH. *Science*, 377(6601):56–62, July 2022.

[78] Nikolas L. Jorstad, Jennie Close, Nelson Johansen, Anna Marie Yanny, Eliza R. Barkan, Kyle J. Travaglini, Darren Bertagnolli, Jazmin Campos, Tamara Casper, Kirsten Crichton, Nick Dee, Song-Lin Ding, Emily Gelfand, Jeff Goldy, Daniel Hirschstein, Katelyn Kiick, Matthew Kroll, Michael Kunst, Kanan Lathia, Brian Long, Naomi Martin, Delissa McMillen, Trangthanh Pham, Christine Rimorin, Augustin Ruiz, Nadiya Shapovalova, Soraya Shehata, Kimberly Siletti, Saroja Somasundaram, Josef Sulc, Michael Tieu, Amy Torkelson, Herman Tung, Edward M. Callaway, Patrick R. Hof, C. Dirk Keene, Boaz P. Levi, Sten Linnarsson, Partha P. Mitra, Kimberly Smith, Rebecca D. Hodge, Trygve E. Bakken, and Ed S Lein. Transcriptomic cytoarchitecture reveals principles of human neocortex organization. *Science*, 382(6667):eadf6812, October 2023.

[79] Nikolas L. Jorstad, Janet H. T. Song, David Exposito-Alonso, Hamsini Suresh, Nathan Castro-Pacheco, Fenna M. Krienen, Anna Marie Yanny, Jennie Close, Emily Gelfand, Brian Long, Stephanie C. Seeman, Kyle J. Travaglini, Soumyadeep Basu, Marc Beaudin, Darren Bertagnolli, Megan Crow, Song-Lin Ding, Jeroen Eggermont, Alexandra Glandon, Jeff Goldy, Katelyn Kiick, Thomas Kroes, Delissa McMillen, Trangthanh Pham, Christine Rimorin, Kimberly Siletti, Saroja Somasundaram, Michael Tieu, Amy Torkelson, Guoping Feng, William D. Hopkins, Thomas Höllt, C. Dirk Keene, Sten Linnarsson, Steven A. McCarroll, Boudewijn P. Lelieveldt, Chet C. Sherwood, Kimberly Smith, Christopher A. Walsh, Alexander Dobin, Jesse Gillis, Ed S. Lein, Rebecca D. Hodge, and Trygve E. Bakken. Comparative transcriptomics reveals human-specific cortical features. *Science*, 382(6667):eade9516, October 2023.

[80] Jerold J. M. Chun and Carla J. Shatz. Interstitial cells of the adult neocortical white matter are the remnant of the early generated subplate neuron population. *Journal of Comparative Neurology*, 282(4):555–569, April 1989.

[81] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, Darren Bertagnolli, Jeff Goldy, Nadiya Shapovalova, Sheana Parry, Changkyu Lee, Kimberly Smith, Amy Bernard, Linda Madisen, Susan M Sunkin, Michael Hawrylycz, Christof Koch, and Hongkui Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335–346, February 2016.

[82] Benjamin J. Stewart, John R. Ferdinand, and Menna R. Clatworthy. Using single-cell technologies to map the human immune system — implications for nephrology. *Nature Reviews Nephrology*, 16(2):112–128, February 2020.

[83] Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, February 2022.

[84] Christoph Kuppe, Mahmoud M. Ibrahim, Jennifer Kranz, Xiaoting Zhang, Susanne Ziegler, Javier Perales-Patón, Jitske Jansen, Katharina C. Reimer, James R. Smith, Ross Dobie, John R. Wilson-Kanamori, Maurice Halder, Yaoxian Xu, Nazanin Kabgani, Nadine Kaesler, Martin Klaus, Lukas Gernhold, Victor G. Puelles, Tobias B. Huber, Peter Boor, Sylvia Menzel, Remco M. Hoogenboezem, Eric M. J. Bindels, Joachim Steffens, Jürgen Floege, Rebekka K. Schneider, Julio Saez-Rodriguez, Neil C. Henderson, and Rafael Kramann. Decoding myofibroblast origins in human kidney fibrosis. *Nature*, 589(7841):281–286, January 2021.

[85] Yoshiharu Muto, Parker C. Wilson, Nicolas Ledru, Haojia Wu, Henrik Dimke, Sushrut S. Waikar, and Benjamin D. Humphreys. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature Communications*, 12(1):2190, April 2021.

[86] Yuan Huang, Christina R. Caputo, Gerda A. Noordmans, Saleh Yazdani, Luiz Henrique Monteiro, Jaap Van Den Born, Harry Van Goor, Peter Heeringa, Ron Korstanje, and Jan-Luuk Hillebrands. Identification of Novel Genes Associated with Renal Tertiary Lymphoid Organ Formation in Aging Mice. *PLoS ONE*, 9(3):e91850, March 2014.

[87] Md. Abdul Masum, Osamu Ichii, Yaser Hosny Ali Elewa, Yuki Otani, Takashi Namba, and Yasuhiro Kon. Vasculature-Associated Lymphoid Tissue: A Unique Tertiary Lymphoid Tissue Correlates With Renal Lesions in Lupus Nephritis Mouse Model. *Frontiers in Immunology*, 11:595672, December 2020.

[88] Yuki Sato, Akiko Mii, Yoko Hamazaki, Harumi Fujita, Hirosuke Nakata, Kyoko Masuda, Shingo Nishiyama, Shinsuke Shibuya, Hironori Haga, Osamu Ogawa, Akira Shimizu, Shuh Narumiya, Tsuneyasu Kaisho, Makoto Arita, Masashi Yanagisawa, Masayuki Miyasaka, Kumar Sharma, Nagahiro Minato, Hiroshi Kawamoto, and Motoko Yanagita. Heterogeneous fibroblasts underlie age-dependent tertiary lymphoid tissues in the kidney. *JCI Insight*, 1(11), July 2016.

[89] Yuansheng Xie, Minoru Sakatsume, Shinichi Nishi, Ichiei Narita, Masaaki Arakawa, and Fumitake Gejyo. Expression, roles, receptors, and regulation of osteopontin in the kidney. *Kidney International*, 60(5):1645–1657, November 2001.

[90] Triet M Bui, Hannah L Wiesolek, and Ronen Sumagin. ICAM-1: A master regulator of cellular responses in inflammation, injury resolution, and tumorigenesis. *Journal of Leukocyte Biology*, 108(3):787–799, September 2020.

[91] Satish L. Deshmane, Sergey Kremlev, Shohreh Amini, and Bassel E. Sawaya. Monocyte Chemoattractant Protein-1 (MCP-1): An Overview. *Journal of Interferon & Cytokine Research*, 29(6):313–326, June 2009.

[92] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, December 2018.

[93] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. anndata: Annotated data. preprint, Bioinformatics, December 2021.

[94] Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso, Ilia Kats, Mikaela Koutrouli, Scverse Community, Philipp Angerer, Volker Bergen, Pierre Boyeau, Maren Büttner, Gokcen Eraslan, David Fischer, Max Frank, Justin Hong, Michal Klein, Marius Lange, Romain Lopez, Mohammad Lotfollahi, Malte D. Luecken, Fidel Ramirez, Jeffrey Regier, Sergei Rybakov, Anna C. Schaar, Valeh Valiollah Pour Amiri, Philipp Weiler, Galen Xing, Bonnie Berger, Dana Pe'er, Aviv Regev, Sarah A. Teichmann, Francesca Finotello, F. Alexander Wolf, Nir Yosef, Oliver Stegle, and Fabian J. Theis. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology*, 41(5):604–606, May 2023.

[95] Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, 22(1):258, September 2021.

[96] Saket Choudhary and Rahul Satija. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology*, 23(1):27, January 2022.

[97] Kidney Precision Medicine Project. Schemata of the human nephron and renal corpuscle developed by the Kidney Precision Medicine Project, 2021.