Title

Physiology From Anatomy Using Spatial Transcriptomic Mapping

Author

Hemminger, Zachary Edward

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

“Physiology From Anatomy Using Spatial Transcriptomic Mapping”

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

by

Zachary Edward Hemminger

2022

ABSTRACT OF THE DISSERTATION

"Physiology From Anatomy Using Spatial Transcriptomic Mapping"

by

Zachary Edward Hemminger

Doctor of Philosophy in Biochemistry, Molecular and Structural Biology

University of California, Los Angeles, 2022

Professor Roy Wollman, Chair

Understanding the physiology of complex systems like tissues and organs is likely impossible without detailed structural maps of the anatomy, especially in the context of perturbations. Spatial transcriptomic techniques like Multiplexed Error Robust Fluorescence In Situ Hybridization or MERFISH have ushered in methods that are capable of generating these detailed anatomical maps for small regions of interest. Existing work primarily focuses on technological development, and few if any have compared perturbed to wild-type conditions. Here we present three cases of increasing difficulty where MERFISH can be used to compare a perturbed state to wild type. Existing spatial transcriptomic approaches, including MERFISH, lack the scale necessary to generate anatomical maps of large tissues and whole organs. Here we present Dimensionally Reduced Fluorescence In Situ Hybridization or dredFISH which allows the generation of detailed anatomical maps at scales far exceeding existing other approaches. Together the fundamental shift towards comparing biological conditions as well as the technological improvements in scale will provide a wealth of detailed anatomical maps which should provide unique physiological insights which likely would have been missed.

The dissertation of Zachary Edward Hemminger is approved.

Margot Quinlan

Xia Yang

Albert Courey

Roy Wollman, Committee Chair

University of California, Los Angeles

2022

Dedication

To my family, friends, mentors, and partner for all of the support and encouragement that they have given me.

Table of Contents

List of Figures

Acknowledgments

This thesis is the result of an accumulation of support that has been given over the course of my life. I am grateful for the support of my family who not only supported me physically, financially, and emotionally but also contributed to my work ethic and my never-ending need to ask questions. I would like to thank my partner, Noelle Alexa Novales, for the constant encouragement as well as being there to bounce ideas off of I would not have been able to accomplish all that I have if not for you. For our dog Rambo (Meatball), who was always there to brighten my day and warm any cold feet.

I would like to thank the QCBio community as well as the Modeling and Microscopy community for which I have learned so much. Specifically, Alex Hoffmann, Aaron Meyer, Eric Deeds, Pavak Shah, and Amjad Askary whose discussions influenced our work as well as how I think about biology as a whole. To our collaborators, I thank you for the opportunity to expand our interests and the enjoyment that came from working with them.

For mentors, I have learned more from you than I ever knew existed. For my undergraduate mentors Jennifer Green, Kristin Picardo, and Bob Curtis thank you for giving me my first opportunity to undergo research and many of the technical skills that allowed me to succeed. For my early graduate mentors Jen and Alon Oyler-Yaniv, Rob Foreman, and Maeve Nagle Thank you for always sharing your knowledge and your way of thinking about science.

I would like to thank my advisor and mentor, Roy Wollman, for giving me the opportunity to develop the skills that I lacked and for allowing me to work on projects that I was passionate about even at the cost of changing the entire focus of his lab. I would also like to thank Gaby Tam, the work we accomplished would not have been possible without your role. Lastly, I would like to thank Wollman Lab members past and present: Evan Maltz, Thomas Underwood, Fangming Xie, as well as all of the undergraduate students who I had the pleasure of working

with during my time. Thank you all for your support as well as for making the time I have spent here an enjoyable experience.

Vita

2013 - 2017 B.S. in Biology & B.S. Chemistry Concentration in Biochemistry,

St. John Fisher College, Rochester, NY

2017 - 2019 M.S. in Biochemistry, Molecular and Structural Biology,

University of California at Los Angeles, Los Angeles, CA

Publications

Nagle, M. P., Tam, G. S., Maltz, E., **Hemminger, Z**. & Wollman, R. Bridging scales: From cell biology to physiology using in situ single-cell technologies. *Cell Syst* **12**, 388–400 (2021)

Lantz, C. ***et al.*** ClipsMS: An Algorithm for Analyzing Internal Fragments Resulting from Top-Down Mass Spectrometry. *J. Proteome Res.* **20**, 1928–1935 (2021)

Littman, R., **Hemminger, Z.**, *et al.* Joint cell segmentation and cell type annotation for spatial transcriptomics. *Mol. Syst. Biol.* **17**, e10108 (2021)

Song, D., Li, K., **Hemminger, Z**., Wollman, R. & Li, J. J. scPNMF: sparse gene encoding of single cells to facilitate gene selection for targeted gene profiling. *Bioinformatics* **37**, i358–i366 (2021)

Hemminger, Z., Walsh, P., Curtis, R. & Picardo, K. Traditional Biocidal Replacement Viability of Microcrystalline Silver Chloride. *J. Nanomed. Nanotechnol.* **08**, (2017)

Chapter 1

Bridging scales: from cell biology to physiology using in situ single-cell technologies

Nagle, Maeve P; Tam, Gabriela S; Maltz, Evan; Hemminger, Zachary; Wollman, Roy

Abstract

Biological organization crosses multiple spatial scales: from molecular, cellular, to tissues and organs. The proliferation of molecular profiling technologies enables increasingly detailed cataloging of the components at each scale. However, the scarcity of spatial profiling has made it challenging to bridge across these scales. Emerging technologies based on highly multiplexed *in situ* profiling are paving the way to study the spatial organization of cells and tissues in greater detail. These new technologies provide the data needed to cross the scale from cell biology to physiology and identify the fundamental principles that govern tissue organization. Here, we provide an overview of these key technologies and discuss the present and future insights these powerful techniques enable.

Introduction

In biology, structure and function are tightly linked. For example, it is the structure of a protein that determines its function, and not simply its amino-acid composition. To solve a protein structure the x, y, and z coordinates of each atom are determined, the local organization identified (e.g. alpha-helix, beta-sheets) and the different domains of the proteins are defined. It is the detailed understanding of the spatial organization of the different amino acids that make up the protein that allows researchers to build a model that explains how its structure (and the dynamics of that structure) determines its function. Similarly, at the cellular level, a list of all the molecules in a cell is insufficient to understand a cell's function. Historically, the electron micrographs obtained by cell biology pioneers such as Palade and Porter in the 1950s were key

to defining cellular organelles and determining the structural organization of the cell (Palade and Porter, 1954). In the 70 years that followed, modern cell biology connected these structural insights to the molecular composition of a cell providing key understanding of how the spatial organization of the molecules that make up a cell determines the cell's function. At the next level, the connection between an organ's anatomy (i.e. structure) and its physiology (i.e. function) has always been a core perspective used to investigate tissues and organs. Histological sections observed using light microscopy have been a key tool that enabled understanding of organ function through insights into their microstructure. However, similar to Palade and Porter's electron micrograph, existing histological approaches lack sufficient molecular details. Histological staining is often based on a combination of non-specific dyes and a handful of molecular markers and does not provide sufficient information to fully understand the complex molecular and cellular structure of the organ. Therefore, while anatomical information is ubiquitous, the lack of spatio-molecular details limits the ability to connect a structure to its function across biological scales.

Body

From dissociative to spatial measurements

Technological advances in single-cell measurements allow the cataloging of all cells into types, subtypes, and states. These catalogs provide key insights into the cellular composition of different organs. The most widespread single-cell technology is undoubtedly single-cell RNA sequencing (scRNA-seq) (Tang et al., 2009). Named "method of the year" for 2013 (2014), scRNA-seq has since become a fixture across many biology labs and has led to many new biological insights due to the ease of analyzing large numbers of cells in a short time frame. In scRNA-seq, cells are dissociated from each other, isolated, barcoded, and sequenced. Due to its dissociative nature, scRNAseq is especially suited for the task of cell classification. However,

this dissociative approach loses the spatial context of cells. Therefore, while this technique provides an invaluable new vocabulary of cell type taxonomy, the lack of spatial information limits its use for organ-scale structure-function analysis.

In recent years, new technologies have been developed that measure the characteristics of single cells *in situ* (in the original site). These technologies link the detailed compositional information obtained through dissociative measurement with spatial histological measurements. These new measurement technologies have transformative potential as they provide the missing data on organs' molecular and cellular structures. By bridging the gap left by dissociative techniques *in situ* technologies provide a route to connect organs' functions to their molecular and cellular structure.

In this review, we discuss the main technologies for characterizing cells *in situ*. We additionally discuss the ways in which *in situ* measurements are contributing to our understanding of biological organization from the subcellular scale to the physiological scale. This review will not focus on the technical aspects of each technology, as previous reviews for scRNA-seq (Chen et al., 2019; Stark et al., 2019) and spatial technologies (Asp et al., 2020; Lundberg and Borner, 2019; Young et al., 2020) have thoroughly addressed these topics. Rather, we provide an overview of key approaches and how they can be used to bridge scales and connect organ cellular structure to its function.

In Situ Technologies

The fast pace of technology development in this space introduces some ambiguity related to terminology. In this review, we make a distinction between the establishment of a cell taxonomy, i.e. classification, and the creation of a cell atlas that requires spatial mapping of cell types in tissues and organs. Similarly, the term *in situ* technologies is ill-defined as *in situ* measurement technologies are as old as histology itself (Motta, 1998) and, depending on the

definition, can include a vast range of measurements. In the scope of this review, we will use a more narrow definition of *in situ* measurements to focus on highly multiplexed spatial measurements of RNA and proteins. RNA measurements are based on *in situ* hybridization, *in situ* sequencing, or RNA capture and cDNA barcoding. Protein measurements are based on antibodies that recognize a specific antigen that can be read either using many rounds of imaging or conjugation with metal ions that are read with a rastering mass spectrometer. Figure 1.1 provides an overview of current approaches for spatial *in situ* measurements.

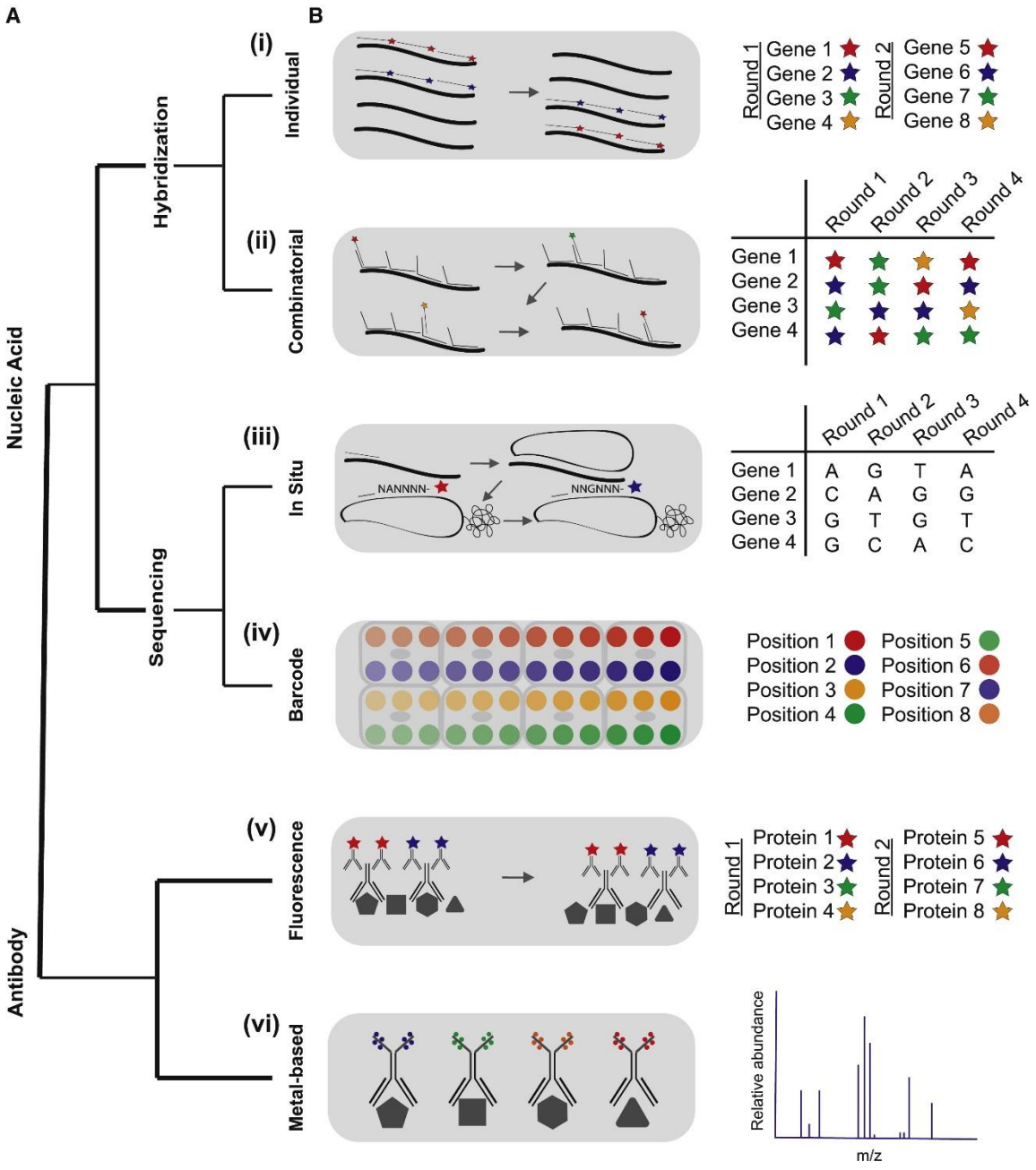

Figure 1.1: Overview of key in situ technologies.
(A) Hierarchical classification of the main approaches used for *in situ* measurements. At the top level, methods are split depending if their main targets are nucleic acids or proteins. Nucleic acid approaches are divided based on the main readout mechanism, hybridization of fluorescent probes to the transcript of interest, or use sequencing to read out the transcript identity. Hybridization approaches are further split into individual approaches or combinatorial approaches. Sequencing approaches either measure RNA in the cell directly or measure DNA barcodes. Antibodies are frequently used to measure protein *in situ* and contain either fluorophore attachments that can be read by fluorescent imaging or metals that are read out by mass cytometry. (B) Schematic representations of key technologies. (i) Individual hybridization techniques, like smFISH, employ many fluorescently-label DNA probes that bind to a transcript

of interest. Each transcript appears as a diffraction-limited fluorescent spot in an image and is identifiable by its unique color. (ii) Combinatorial hybridization techniques utilize similar principles to individual hybridization but utilize consecutive binding of probes to the same molecules and sequential imaging to create a “barcodes” of fluorescent spots across imaging rounds that are used to determine a transcript’s identity. An additional set of probes are used in combination hybridization that bind directly to an RNA transcript with overhangs for fluorescent readout probes to bind to. (iii) *In situ* sequencing involves the readout of an RNA transcript directly or of a barcoded primer used to amplify that transcript. The transcript is first reverse transcribed into cDNA, then that cDNA is amplified, frequently by rolling circle amplification. The amplified cDNA is either sequenced directly or the sequence of a specific primer that binds to the cDNA is sequenced. (iv) In *in situ* barcoding methods, a sample is applied to a slide covered with DNA-barcoded microbeads. The sample is lysed and the resulting RNA binds to the beads, which are then sequenced. The location of the RNA is mapped back to the known location of the DNA barcode sequence from the bead. (v) In each round of fluorescent-based antibody readout, proteins are bound to an antibody with a fluorescently labeled molecule attached, similar to in (i). Each protein is represented by a single-colored fluorescent spot in an image. (vi) Metal-based antibody readouts are similar to fluorescent-based antibody readouts but utilize unique metal atoms attached to antibodies instead of fluorescent molecules. These metal atoms are read out using a mass cytometer.

RNA hybridization

Single-molecule RNA fluorescence *in situ* hybridization (smFISH) (Femino et al., 1998; Raj et al., 2008) was the first widespread single-molecule *in situ* RNA measurement technology. smFISH counts the number of mRNAs transcribed from a gene of interest within a cell by using DNA probes specific to the mRNA target sequence. These DNA probes are attached to a fluorescent molecule and collectively create a single diffraction-limited spot in the position of the mRNA molecule. The number of diffraction-limited fluorescent spots in a cell is counted to determine the number of mRNA molecules present. Several techniques seek to improve upon the probe design of smFISH. A partial list of these extensions includes RNAscope ((Wang et al., 2012), which uses Z-shaped DNA probes to enhance specificity, osmFISH (Codeluppi et al., 2018) which is optimized for use in thin tissue sections such as brain slices, ExFISH (Chen et al., 2016) which uses expansion microscopy to further separate mRNA spots and make image analysis easier, and SABER-FISH which uses multi-part probes to enhance the signal from each mRNA (Kishi et al., 2019). Overall, the principle that is shared between smFISH and its many subsequent versions is that expression of pre-defined genes is measured in a targeted

manner with one measurement per gene. The high accuracy and mRNA capture rate (both >95%) have led smFISH to become the “gold standard” among validation techniques (Torre et al., 2018). However, the high accuracy comes at a price: smFISH-based approaches assign each gene a specific measurement (i.e. color), so there can only be as many genes measured in a single hybridization as there are non-overlapping fluorescent molecules available. Four rounds of hybridization with this approach using four different types of fluorescent probes can measure a maximum of 16 genes. This linear scaling limits the ability of smFISH-based approaches to provide full and detailed structural information.

Combinatorial FISH approaches address the key limitation of smFISH by increasing the number of genes that can be counted per experiment and thereby provide much more detailed information on the cellular composition in the tissue. These techniques include MERFISH (Chen et al., 2015; Moffitt et al., 2016a, 2016b; Wang et al., 2020; Xia et al., 2019a), seqFISH+ (Eng et al., 2019), and most recently split-FISH (Goh et al., 2020). These approaches, while very similar, differ in some of the details related to barcoding strategy and how they remove the fluorescently-tagged oligo probes. The core improvement over smFISH is that combinatorial FISH approaches utilize barcodes for each RNA to increase the measurement capacity. Each gene is given a ‘barcode’ that is a combination of colors, so the gene identity is uncovered by the data from every round of hybridization. This process scales exponentially, so four rounds of hybridization with four different types of fluorescent probes would allow for up to 256 genes to be analyzed, instead of 16. Using four dyes and eight rounds of hybridization ($48 = 65{,}536$), in principle an entire transcriptome can be measured. However, the use of RNA barcodes comes at a price. In the 48 scheme, any error in “calling” one of the four measurements needed to assign a gene identity to an RNA molecule will result in an incorrect assignment. Such errors have the potential to substantially reduce the accuracy of combinatorial FISH approaches. To address this limitation, the barcodes are typically chosen sparsely from a large set of possible

codes. This intentional reduction in chosen barcodes can substantially reduce the error rates of combinatorial approaches at a cost of an increase in the number of hybridization rounds. In typical combinatorial measurement, 24 rounds are used with each molecule having 4 measurements out of the possible 24 rounds. The sparsity of barcode assignment is such that 200-500 genes can be measured using 24 rounds of imaging. Both MERFISH and seqFISH+ were used to demonstrate that transcriptome scale (~10,000 genes) is possible but at a cost of a much higher number of measurements and overall reduced throughput (Chen et al., 2015; Eng et al., 2019). The flexibility of combinatorial FISH approaches is important as the complexity of the approach often requires tailoring measurements to specific samples and experiments. Optical crowding of many RNA spots per image can impede RNAs from being resolved and require integration of some smFISH rounds for highly expressed genes or a substantial increase in the number of measurements. As a result, large samples can require weeks of continuous imaging and can generate terabytes of image data. Overall, combinatorial FISH approaches provide a very powerful platform for targeted spatial RNA counting that can be tailored to the specific needs of a project.

Antibodies

Immunohistochemistry (IHC) has been used since 1942 to study the spatial location of proteins in a tissue (American Association of Immunologists, 1942). IHC involves adding labeled antibodies to a sample in order to visualize proteins and other molecules of interest. Despite the high specificity achieved by antibodies, IHC is difficult to multiplex. Only in the last two decades have a few approaches been successfully implemented to enable 30+ protein readouts in a sample. The first difficulty is the generation of validated high-quality antibodies. In practice, this is a non-trivial issue that has been partially addressed by both commercial and academic groups (Edfors et al., 2018) but is by no means a solved problem. The second difficulty relates to how the spatial distribution of these antibodies is read. To prevent cross-reactions and due to

the limited number of host animals used in antibody production, the use of primary and secondary antibodies, common in standard IHC, is difficult. This limits antibody selection to mostly primary antibodies that need to be read across multiple measurements. Here, we focus on solutions that address the readout problem. Overall the multiple attempts at “cracking” the multiplexing challenge can be divided into two types: 1) repeated imaging on a light microscope and 2) coupling antibodies to unique metal ions.

A straightforward way to increase the number of readouts is to use existing tools for fluorescence-based antibody detection and simply repeat them many times (Fig 2.1.1). For example, a set of antibodies would be added to a sample, imaged, then stripped away and replaced with a new set of antibodies. This idea is implemented in methods such as MxIF (Gerdes et al., 2013), CycIF (Lin et al., 2015, 2018), and 4i (Gut et al., 2018). The key distinction between the different variants is in how the multiple rounds of staining are achieved, i.e. are the antibodies themselves stripped from the sample, or are they simply quenched by photobleaching. An important advantage of these approaches is that since they are based on standard microscopy; they can also be coupled to live-cell imaging (Lin et al., 2015). Borrowing from the relative ease of repeated imaging after RNA hybridization a few methods, CODEX (Goltsev et al., 2018), DEI (Wang et al., 2017), and Immuno-SABER (Saka et al., 2019) use oligo-conjugated antibodies and fluidics systems almost identical to the one used by combinatorial FISH approaches.

An alternative approach for multiplexing antibody staining is based on changing the readout from a light microscope to a mass spectrometer. Mass cytometry imaging approaches have been developed to avoid some of the practical limitations encountered by attempts to multiplex IHC-based analysis. Specific implementations of imaging mass cytometry include Multiplexed Ion Beam Imaging (MIBI) (Angelo et al., 2014; Keren et al., 2019; Ptacek et al., 2020) and Imaging Mass Cytometry (IMC) (Giesen et al., 2014; Ijsselsteijn et al., 2019). Each of

these approaches uses secondary ion mass spectrometry to image antibodies tagged with isotopically pure elemental metal reporters. The main distinction between the two methods arises in sample ablation which leads to differences in image resolution and acquisition times between IMC and MIBI (Baharlou et al., 2019). Though these techniques can analyze up to 40 proteins in a sample at a given time, they are both limited by antibody availability and quality. Additionally, MIBI and IMC require specialized equipment to point-scan small fields, and therefore imaging large samples can be slow and costly.

Sequencing

The accessibility of DNA sequencing, achieved in part due to six orders of magnitude decrease in sequencing cost per base pair (Stark et al., 2017), motivated innovative approaches that leverage DNA sequencing while still preserving spatial information. The approaches that couple spatial information to RNA sequencing can be divided into three distinct categories: 1) separation of RNA based on their spatial location followed by sequencing, 2) use of spatially distinct DNA barcodes during library preparation, and 3) performing the sequencing reactions themselves *in situ*. The first two categories directly leverage existing sequencing technologies whereas the latter use many of the chemistry developed for sequencing however the readout itself is microscopy-based and shares many similarities to combinatorial FISH-based approaches.

Perhaps the most straightforward way to assign spatial information to RNA molecules is to only collect RNAs from a specific spatial domain. This concept is the basis of highly useful methods such as LCM-seq (Nichterwitz et al., 2016) and GEO-seq (Chen et al., 2017) that use laser capture microscopy to sequence a small number of cells at a time. A more systematic application of a spatial collection of RNA from distinct regions was applied using a method called Tomo-seq (Burkhard and Bakkers, 2018) that uses cryosectioning to the tissue before

sequencing. Photoactivation is another useful tool that was used to encode spatial information and capture RNAs in spatially distinct domains. Transcriptome in vivo analysis (TIVA) exposes live cells to multifunctional caged mRNA-capture molecule tags called TIVA that upon photocleavage hybridize to mRNAs within a cell allowing sequencing of RNAs from specific spatial position (Lovatt et al., 2014). A similar idea was implemented by ZipSeq (Hu et al., 2020) that used patterned light and three distinct colors to label cells according to their spatial position. Labeled cells are sorted and sequenced using standard scRNAseq tools. These spatial-specific capture approaches have been effective tools in understanding the organization of tissues. However, they suffer from an inherent tradeoff between resolution and throughput. While Tomo seq allowed sequencing entire embryos, this was done in linear sections of 18-micron thickness. On the other extreme TIVA can be used for subcellular localization of RNA molecules however it can only process one location at a time. Therefore, while the approaches that are based on spatially restricted RNA collection provide important spatial information they stop short of enabling the cellular structure of organs and tissues.

To overcome the tradeoff between spatial resolution and throughput, an alternative approach is based on localized barcoding of cDNA during library preparation prior to sequencing. The key advantage of position-based barcoding is that once each region is labeled by a specific code the entire sample can be sequenced as one and using prior knowledge of the XY position of each barcode, the spatial position of all RNA molecules is reconstructed computationally. These approaches involve capturing RNA from tissue samples on a spatially barcoded bead array which is later sequenced. Both High Definition Spatial Transcriptomics (Salmén et al., 2018; Ståhl et al., 2016; Vickovic et al., 2019) and Slide-seq (Rodriques et al., 2019; Stickels et al., 2020) use this approach. While this technique cannot define cell boundaries, High Definition Spatial Transcriptomics can achieve two-micron resolution and allows for fast, high-throughput processing (Vickovic et al., 2019). A key advantage of these

approaches is that they leverage many of the experimental and computational tools developed for scRNASeq. In fact, popular analysis tools such as Seurat were able to add the spatial capture analysis despite the scarcity of datasets that used this approach partially due to its similarity to scRNAseq (Stuart et al., 2019). Spatially resolved sequencing is a promising approach, however, presently it suffers from low RNA capture efficiency. The low capture efficiency means that the capture bin (i.e. spatial domain of a single barcode) needs to be big enough to contain a sufficient number of RNA molecules. Furthermore, even if the capture chemistry will improve, similar to other capture-based approaches there is an inherent tradeoff between resolution, i.e. the size of a single capture bin and the number of bins. To allow subcellular information, capture bins need to be <100 micrometer2 which means that a standard tissue section of 100 mm2 will need 106 distinct barcodes, a non-trivial library to sequence.

In situ sequencing leverages the conceptual advances of DNA sequencing, but not the sequencing machines themselves. In situ sequencing converts RNA in a cell to cross-linked cDNA amplicons that are sequenced within a cell on a microscope. These molecules can either be the RNAs of interest themselves, as in FISSEQ (Lee et al., 2014, 2015), or an RNA barcode specific to transcripts of interest, like in ISS (Ke et al., 2013), STARmap (Wang et al., 2018), and Baristaseq (Chen et al., 2018). FISSEQ (Lee et al., 2014, 2015) cross-links DNA amplicons to a matrix to directly sequence the amplicon inside a cell. ISS, STARmap, and Baristaseq add barcoded oligos specific to targets of interest and sequence the barcodes to determine the presence of transcripts. Similar to combinatorial FISH approaches, barcode-based *in situ* sequencing requires an oligo library that targets genes of interest. While in principle *in situ* sequencing approaches can provide an unbiased view of RNA in tissues and organs, in practice this comes at a cost associated with the need to sequence many copies of highly abundant RNA molecules. The targeted methods have shown more robustness in their implementations and have dominated over unbiased ones. Interestingly, given their targeted nature, the

distinction between them and combinatorial hybridization-based approaches diminishes. This is exemplified in a new protocol called HybISS that merges the rolling circle amplification typical to *in situ* sequencing approaches with hybridization-based multi-round readout that is common in combinatorial FISH (Gyllborg et al., 2020).

Integrative in situ measurements

Integrative spatial multi-modal *in situ* approaches combine the measurements across modalities, i.e. RNA and protein. The integrative and multi-model data will likely enable a more comprehensive understanding of single-cell processes and functions. Many recent innovations in this direction point to an exciting future with complex datasets that span different data types. Techniques like Digital Spatial Profiling (DSP) (Merritt et al., 2020), RNAscope (Kann and Krauss, 2019), smFISH-IF (Tutucci and Singer, 2020), and ImmunoFISH (Kwon et al., 2020) combine FISH and immunofluorescence methods to measure RNA and protein levels within a single cell. RNAscope has additionally been paired with mass cytometry to read RNA and protein levels (Schulz et al., 2018). SABER-FISH also allows for the *in situ* measurement of DNA or RNA transcripts and can combine protein staining for simultaneous detection of a gene's transcript and protein levels (Kishi et al., 2019). Another venture involves reading out DNA and RNA within the same cell *in situ*. ClampFISH (Rouhanifard et al., 2018) probes can be used on both DNA and RNA sequences, allowing for the measurement of DNA and RNA in the same cell in the same experiment. Additionally, live-cell imaging has been combined with *in situ* transcriptomics to allow for mapping the transcriptional state of a cell to its phenotype. CycIF tracked the translocation of a YFP-FoxO3a reporter followed by the readout of seven additional protein levels (Lin et al., 2015). A recent paper analyzed calcium signaling response and gene expression of calcium signaling-related genes in over 5,000 cells (Foreman and Wollman, 2020). New avenues of spatial multi-omics are just now being explored and could have a great impact on the construction of atlases with many maps. Furthermore, these technologies open

up new avenues to study gene perturbations (Wang et al., 2019), cell lineage tracing (Chen et al., 2018; Frieda et al., 2017), and other aspects of functional DNA and RNA biology (Cai et al., 2020; Maiser et al., 2020).

Integrative reconstructions have been developed by combining large dissociative datasets with a smaller number of spatial measurements used as a “ruler”. For example, algorithms have been developed to infer the original spatial location of cells analyzed by scRNA-seq by correlating the level of key marker genes with levels of those genes found within *in situ* datasets (Achim et al., 2015; Satija et al., 2015). Other algorithms such as trendsceek and LIGER have also been used to integrate scRNA-seq data with spatial transcriptomics information (Edsgärd et al., 2018; Welch et al., 2019). The integration across spatial and non spatial datasets enabled spatial reconstruction by combining laser capture microdissection, bulk sequencing those cells, and then reconstructing the whole tissue through spatial tissue reconstruction (Moor et al., 2018). These reconstruction-based approaches are very powerful as they merge the strengths of dissociative and spatial measurements. However, care needs to be taken in the interpretation of these reconstructions. The recovered maps are based on spatially stratified averaging of many cells. These averaging could mask additional spatial differences that are lost due to averaging. Therefore, the details of the reconstruction matter and care should be used in the interpretation of these measurements.

Challenges are truly opportunities

The ultimate *in situ* measurement technology will have sub-cellular resolution, high detection sensitivity, will be applicable to 3D volumes, compatible with multiple fixation protocols including FFPE, and provide highly multiplexed data on a wide range of molecular species. Given the inherent tradeoff between resolution, sensitivity, and throughput, none of the technologies described above should be considered a “winner”. It is likely that many different

technologies will be developed where each will be a “winner” for a specific subset of applications. As a result, *in situ* technologies contain a smorgasbord of different approaches, each with their own acronym and nuances. Despite the variety, these technologies face some similar challenges. The first challenge is the computational and data complexity. Despite the differences in methods, many computational steps, such as spot calling and cell segmentation (Littman et al., 2020), are shared across approaches. Development of standards and mature computational libraries that can allow the separation of the computational analysis from data acquisition will allow more cross-fertilization in this field. Currently, the Chan Zuckerberg Initiative (CZI) has begun building a unified data-analysis tool and file format called starfish to address this issue (Perkel, 2019). These standards will allow the use of modern machine learning methods that will invariably be key to solving many of these problems (Bannon et al., 2021; Chen et al., 2020a; Moen et al., 2019; Stringer et al., 2021). The second challenge relates to scale. MERFISH imaging of a volume comparable to a mouse brain would require more than a year of continuous imaging. Other technologies, such as spatial RNA barcoding, have a similar order of magnitude time requirements. To achieve cellular resolution for such volume requires ~1013 reads which, even on an advanced NovaSeq 6000 will take multiple years to sequence. The third challenge is the dissemination of these technologies to the scientific community. The complexity of many of these protocols makes the open-source / open hardware model challenging. Many companies are actively working on bringing these innovations to market which will help. However, whether these efforts will democratize the best technologies remain to be seen. Finally, once spatial data is collected, how to fully analyze it and maximize the insights such data provides is very much an open research question. As was the case for single-cell biology, we anticipate that increase in data availability will result in further developments in statistical and bioinformatics methodology to analyze these rich and interesting datasets. We are optimistic that these challenges will act as a catalyst for innovation and we expect further technological development in this space.

From cell biology to physiology

The technologies introduced above are paving the way for bridging the gap between intracellular, cellular, and physiological scales (Fig 2.1.2).

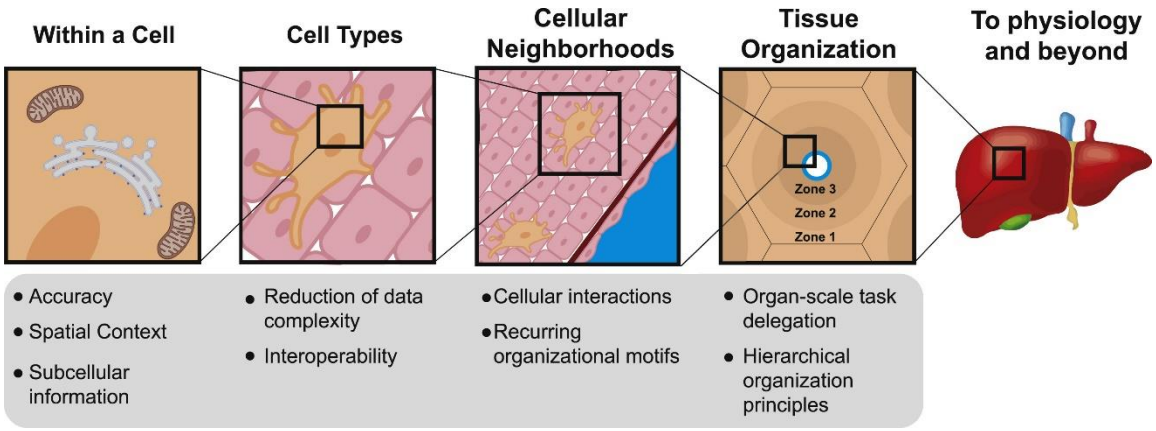

Figure 1.2: Bridging scales with in situ technologies.
In situ technologies can reveal new biology across many scales of biology, including within a cell, cell types, cellular neighborhoods, tissue organization, and physiology. By bridging these scales, *in situ* technologies can provide insights into the structure-function relationship across multiple scales.

Within a cell

In situ measurements provide three key benefits over dissociative approaches in the analysis of single cells: 1) higher accuracy, 2) spatial context, and 3) subcellular information:

Accuracy: Many *in situ* techniques like MERFISH and seqFISH are more sensitive and less biased than their dissociative counterparts like scRNAseq. Therefore, for a broad range of biological questions that require accurate transcript numbers *in situ* technologies should be used. For example, analysis of gene expression variability is non-trivial using scRNAseq data with sensitivities around 10%. Analysis of gene expression variability based on scRNAseq data requires accounting for this large measurement error with complex error models. Unfortunately, these are non-trivial and introduce a large number of additional assumptions, such as a high

degree of transcriptional bursting (Jiang et al., 2017; Larsson et al., 2019), that are not always fully substantiated (Battich et al., 2015; Foreman and Wollman, 2020).

Spatial Context: The spatial context of *in situ* technologies allow for analysis of cellular heterogeneity in a much more physiological context. To fully understand the sources of cellular heterogeneity, we need to understand what factors influence cell state. Does spatial position in a tissue affect the variance of key genes? How does a cells' gene expression predict its present and future behavior? Efforts to track cells over time have revealed that understanding a cell's gene expression is insufficient to understand the choices that cells make (Weinreb et al., 2020). More information about a cell is therefore imperative to know in order to understand how a cell makes decisions. Recent work identified more than 40 genes in the mouse hippocampus to be cell subtype-specific spatial differentially expressed genes (spDEGs) (Littman et al., 2020). These results suggest that a spatial position can explain much of the heterogeneity seen using dissociative approaches.

Subcellular Information: A subset of *in situ* techniques are capable of discerning RNA and protein localization at the subcellular level. High resolution allows for the determination of expression patterns in organelles as well as the analysis of coexpression of genes by subcellular localization. MERFISH is one such technique and has characterized the RNA enrichment in the endoplasmic reticulum and the nucleus (Xia et al., 2019b) as well as the dendrites and axons of neurons (Wang et al., 2020). On the proteomics side, 4i allows subcellular detection of protein abundances (Gut et al., 2018). 4i goes further and determines that the subcellular spatial protein distribution between single cells that experience different cell cycle states, microenvironments, or growth conditions affects the localization of EGFR upon the cell's exposure to EGF. Collectively, the accuracy, context, and resolution of many *in situ* technologies enable a more accurate picture of the biology of cells in a true physiological context.

Cell types: the building blocks of tissues

Classification of cells into (sub)types and states is an important step toward deciphering the structure/function relationship of tissues and organs. The two key advantages of classification of cells into (sub)types and state are 1) reduction of data complexity, i.e. a single cell type label can be used to replace a complex vector of transcriptome scale gene expression values. 2) interoperability between different experiments including across spatial and dissociate measurements, i.e the same nomenclature can be prescribed to cells across experiments. These benefits of cell classification systems and the existence of a large body of data from dissociative studies motivate many ongoing efforts to create robust cell classification systems (Trapnell, 2015; Yuste et al., 2020). However, the definition of a cell type and cell states is not consistent across fields, or even across researchers within a field. In addition, the appropriate criteria to use to classify cells are debated. This heterogeneity adds additional complications to cell classification, so it remains unclear if a single classification system will emerge or whether classification will have to be redefined for each analysis.

Three complementary and non-mutually exclusive views of cell types have been used as frameworks to determine cellular classification systems: landscape, microenvironment, and task (Fig 2.1.3). The first view, famously referred to as the Waddington landscape (Waddington, 1957a), suggests that intracellular biological regulatory networks are configured such that they can exist in a finite number of steady states. In the landscape point-of-view, cellular classification is molecular in origin and depends on stability analysis in high-dimensional phase space (Ferrell, 2012; Trapnell, 2015). While inputs to the cell during its developmental trajectory can influence cell fate decisions, these transitions are still encoded by the underlying regulatory network and therefore the classification is focused on a cell's internal state (Waddington, 1957b). The second view is that a cell type is defined by its microenvironment: the chemical, mechanical, and biological cues surrounding the cell. The cell is influenced and shaped by its

neighbors, its resources, and its environmental cues. The third view is that classification of cells into types has to follow the functional tasks cells are required to perform for the organism as a whole. Under this view, there are key cell archetypes, each specialized in a specific task. Each individual cell performs one or a few of these tasks and its molecular state will match the tasks it performs (Korem et al., 2015). Not only are the three views of landscape, microenvironment, and task non-mutually exclusive, they are in fact complementary and are likely different views of roles and states of cells in a multicellular organism. For example, the transition from monocyte to macrophage is guided by an internal epigenetic regulatory network (Álvarez-Errico et al., 2015). Macrophages can polarize to perform different tasks based on stimulatory cytokines (Murray, 2017) while at the same time are heavily influenced by the tissue microenvironment (Lavin et al., 2014). Combining multiple viewpoints to create one (or more) cell classification system is an important stepping stone in analyzing the cellular structure of tissues and organs.

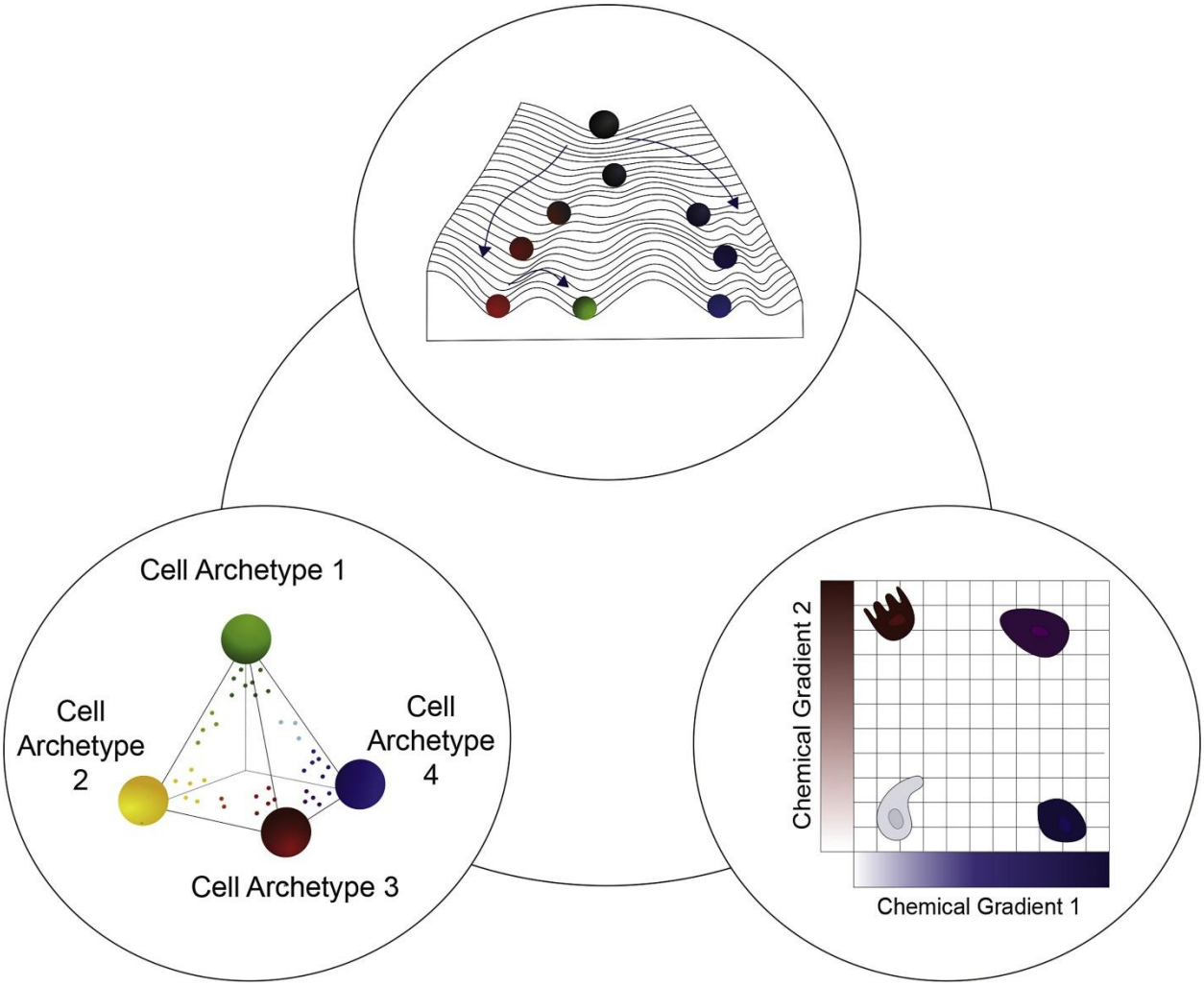

Figure 1.3: Geometrical representation of cell types.
Three complementary views of the concept of cell type. These concepts are non-mutually exclusive and represent complementary views. (Top) The Waddington landscape uses the geometrical analogy of landscape. In this view, a cell type is a specific valley in 'cell space'. As pluripotent cells differentiate they pass through the landscape to reach their final position. This view is largely focused on the intracellular epigenetic and gene regulatory networks that define the possible valleys in the landscape. (Left) The task-based view proposes that each cell performs one or a few tasks. Each task (cell archetype) is represented as a vertice on a high dimensional polyhedron. The specific tasks each cell performs will determine its position within the polyhedron. (Right) The microenvironment view proposes that cell types are defined by the chemical, mechanical, and biological cues surrounding a cell. The cartoon shows a simplified view with two signaling gradients and the position of the cell in that space will determine its type.

In situ measurement technologies are well suited to generate and utilize cell classification systems. MERFISH and seqFISH were used to categorize the organization of predefined cell types within the brain (Chen et al., 2015; Littman et al., 2020; Shah et al., 2016). Other *in situ*

technologies, such as *in situ* sequencing leverage the existing cell type taxonomies to overcome low RNA detection efficiency and still provide key cell type information (Qian et al., 2020). Rather than solely focusing on existing classification systems, work based on seqFISH in combination with scRNAseq redefined cell types based on a Hidden Random Markov Field analysis of expression domains (Zhu et al., 2018). With further improvement of cell segmentation algorithms, it is likely that morphological information could be incorporated into classification models based on *in situ* measurements. Together with existing spatial information, it is expected that *in situ* technologies will play a key role in further refinement and development of cell type and state classification.

Cellular neighborhoods and communities

Cellular neighborhoods, the local spatial distribution of different cell types on the scale of hundreds of micrometers, are poorly understood. However, such length scales likely play an important role in bridging the gap between individual cell function and complex organ function. A good analogy for cellular communities is urban planning for human residential neighborhoods. A typical neighborhood with many houses will also have a coffee shop, a grocery store, and will be served by major roads and key public transportation. Similarly, cellular communities will have many cells of a few types that are needed for the specific organ (i.e. neurons in the brain, hepatocytes in the liver), but will also have resident macrophages, mast cells, and fibroblasts and will be in proximity to blood vessels. The number and spatial distribution of these specialized cell types have major implications for the function of the organ in their ability to relay information and perform their function (Bagnall et al., 2018). A good example of these principles come from recent cell-type mapping in the brain were *in situ* multiplexed RNA FISH uncovered a high spatial self-affinity of ependymal cells as well as spatial self-avoidance of inhibitory neurons, microglia, and astrocytes (Codeluppi et al., 2018). The paper also found that endothelial cells were found within roughly 65 microns of all other cell types. Another principle

that will likely help identify cellular communities is communication between cells. Direct measurement of communication between cells is challenging, but a useful proxy is ligand-receptor interactions in neighboring cells (Browaeys et al., 2020). Work that used multiplexed RNA measurement in brain slices (Eng et al., 2019) found that endothelial cells next to microglia in the olfactory bulb express endoglin and activin A receptor mRNA while the microglia expressed TGFB ligand mRNA (Eng et al., 2019). By contrast, endothelial cells adjacent to microglia in the cortex expressed Lrp1 and Pdgfb mRNA. Collectively such studies bring an intriguing hypothesis that there are key principles that could be generalized to identify community-level 'rules' of cellular patterning. What exactly are these rules and what are the molecular mechanisms used to implement them, e.g. the chemical gradient (Lander et al., 2009) and differential adhesion (Tsai et al., 2020), are key open questions.

In situ technologies coupled with new analysis approaches are well-positioned to make valuable contributions to our understanding of cellular communities. The highly multiplexed and inherently spatial nature of *in situ* measurement technologies makes them an ideal tool to acquire the data needed to understand cellular community organization. However, data collection is only the first step in identifying the rules and principles that govern cellular community organization. New bioinformatics and statistical tools will be required to allow researchers to convert the raw data on molecular distributions of RNA and proteins into insights. The rich literature of statistical learning including ideas related to community detection in multi layer networks (Mucha et al., 2010) and concepts from topic modeling (Blei et al., 2003) will accelerate the development of these much-needed statistical analysis tools for cellular community organization. Initial implementation of ideas from topical modeling to cellular communities is very promising (Chen et al., 2020c). Overall, while the amount of work done so far to understand cellular neighborhoods remains small, the iterative development of data

collection and analysis tools presents enticing prospects for understanding the role of the microenvironment in cellular organization.

Principles of tissue organization

How do multiple cellular communities interact together to form complex organs is a fundamental question. It is unclear whether there are few fundamental principles that can explain cellular self-organization across multiple organs or whether the way that multiple cellular communities synergy is organ dependent. It is likely that the organization of complex organs such as the brain or the liver that perform many distinct functions are different from each other and from simpler tissues such as the intestine or cornea. Nonetheless, the lack of complete data on cell type, RNA, and protein distribution across entire organs makes answering such questions difficult.

The liver was the first organ to be studied in depth with *in situ* approaches. The largest internal organ in the body, it performs roughly 500 tasks, including bile production, fat metabolization, vitamin and mineral storage, and blood filtration (Ben-Moshe and Itzkovitz, 2019). These tasks are non-homogeneously carried out by different subsets of cells within the liver. While it has long been understood that the various functions of the liver are not all carried out in the same spaces, *in situ* approaches have allowed researchers to dive further into the detailed arrangement of cell types and functions throughout the liver. Halpern et al. showed that roughly half of the hepatocyte genes, the main cells of the liver, are expressed in a zonated manner (Halpern et al., 2017). A subsequent study showed that liver endothelial cells are also highly zonated, with more than 30% of their genes expressed in a zonated manner (Halpern et al., 2018). Using spatial mapping, a high-resolution, global expression map of liver zonation was created that showed tasks that are high-energy are carried out in the highly oxygenated periportal locule layers where hepatocytes can more readily generate ATP through respiration

(Halpern et al., 2017). This conclusion supports theoretical results on spatial task allocation in organs (Adler et al., 2019). While the work on the liver has shown exciting insights into its spatial organization it is still nascent and does not differentiate between the different lobes of the liver, begging the question of whether each lobe shows additional sub-specializations. Outstanding questions of the liver organization still remain, including whether specific cell types including Kupffer cells (Bykov et al., 2004) and hepatic stellate cells (Friedman, 2008) are spatially heterogeneous. As technology develops, we anticipate exciting findings on the spatial organization of the liver's 500 tasks. The example of the liver shows how much was already learned, yet at the same time how much more there is to discover on tissue spatial organization using *in situ* measurement technologies.

To physiology and beyond

The ultimate goal of biomedical research is to improve our understanding of human biology and how it is disrupted during disease. From a translational point of view, it is often an organ function that is impacted by disease. In situ measurement technologies are poised to provide new insights on normal physiology and importantly provide detailed information on what goes wrong in a disease state. In situ techniques have been applied to a subset of organ diseases. Systematic charting of the brain (Moffitt et al., 2018; Shah et al., 2017; Zhang et al., 2020), heart (Asp et al., 2019), liver (Ben-Moshe and Itzkovitz, 2019; Halpern et al., 2017, 2018), intestine (Moor et al., 2018), and bone marrow (Baccin et al., 2020) are starting to uncover the single-cell architecture of multiple tissues and organs. Spatial Transcriptomics has been used to study ALS (Maniatis et al., 2019), prostate cancer (Berglund et al., 2018), melanoma (Thrane et al., 2018), and Alzheimer's disease (Chen et al., 2020b) while MIBI-TOF and imaging mass cytometry have been applied to the study of breast cancer (Jackson et al., 2020; Keren et al., 2018). Systematic efforts to scale *in situ* mapping to tumors are ongoing (Rozenblatt-Rosen et al., 2020).

Spatial maps of prostate cancer transcriptomes have shown prostate cancer samples have high heterogeneity across the tumor and that distinct cancer expression regions can extend beyond the boundaries of annotated tumor areas (Berglund et al., 2018). These findings suggest using spatial information of tumors is important in classification schemes to rank tumor severity and could be used to predict further 'high risk' areas of potential cancer growth. Similarly, an analysis of triple-negative breast cancer by MIBI-TOF found that the spatial organization of infiltrating immune cells inside solid tumors is predictive of patient survival, where patients with more compartmentalized immune cells inside tumors fared better than patients where immune cells were well mixed within the tumor (Keren et al., 2018). Spatial transcriptomics has also been used in other diseases to track how disease progression occurs molecularly. A recent study of ALS quantified over 11,000 genes in mice and over 9,000 genes in humans to show that microglial dysfunction occurs well before ALS symptom onset and this dysfunction is mediated by the phagocytosis-related genes TREM2 and TYROBP (Maniatis et al., 2019). These technologies are pushing our frontier of understanding and even show areas where *in situ* analyses can suggest improvements in current medical practices. On a longer timescale, it is possible that *in situ* measurements will become an important diagnostic tool.

Conclusion

Spatial Biology is still an emerging field driven in large part by new *in situ* technologies. The ability of these approaches to provide rich spatially defined datasets about molecular and cellular distribution across multiple spatial scales poise these technologies to make critical contributions to our understanding of the inherent relationship between structure and function at multiple levels. The future of this field is quite bright, with applications ranging from understanding disease progression and how the structure of organs such as the brain and the liver relates to their functions. As was the case with single-cell biology, an increase in the adoption of these technologies will increase data availability and will result in innovation in data

analysis. Such iterative improvements in data acquisition and analysis will provide key insights that will allow researchers to bridge the gap from molecular and cellular biology to complex human physiology. The best is yet to come.

Acknowledgements

Results in this chapter were adapted from a manuscript published in *Cell Systems*:

Nagle, M. P., Tam, G. S., Maltz, E., Hemminger, Z. & Wollman, R. Bridging scales: From cell biology to physiology using *in situ* single-cell technologies. *Cell Syst* **12**, 388–400 (2021)

Author Contributions

MN, GT, EM, ZH & RW designed the project. MN & GT wrote the manuscript.

References

Achim, K., Pettit, J.-B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat. Biotechnol. *33*, 503–509.

Adler, M., Korem Kohanim, Y., Tendler, A., Mayo, A., and Alon, U. (2019). Continuum of Gene Expression Profiles Provides Spatial Division of Labor within a Differentiated Cell Type. Cell Syst.

Álvarez-Errico, D., Vento-Tormo, R., Sieweke, M., and Ballestar, E. (2015). Epigenetic control of myeloid cell differentiation, identity and function. Nat. Rev. Immunol. *15*, 7–17.

American Association of Immunologists (1942). The Demonstration of Pneumococcal Antigen in Tissues by the Use of Fluorescent Antibody. The Journal of Immunology *45*, 159–170.

Angelo, M., Bendall, S.C., Finck, R., Hale, M.B., Hitzman, C., Borowsky, A.D., Levenson, R.M., Lowe, J.B., Liu, S.D., Zhao, S., et al. (2014). Multiplexed ion beam imaging of human breast tumors. Nat. Med. *20*, 436–442.

Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J., Salmén, F., et al. (2019). A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. Cell *179*, 1647–1660.e19.

Asp, M., Bergenstråhle, J., and Lundeberg, J. (2020). Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. Bioessays e1900221.

Baccin, C., Al-Sabah, J., Velten, L., Helbling, P.M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L.M., Trumpp, A., and Haas, S. (2020). Combined single-cell

and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat. Cell Biol. *22*, 38–48.

Bagnall, J., Boddington, C., England, H., Brignall, R., Downton, P., Alsoufi, Z., Boyd, J., Rowe, W., Bennett, A., Walker, C., et al. (2018). Quantitative analysis of competitive cytokine signaling predicts tissue thresholds for the propagation of macrophage activation. Sci. Signal. *11*.

Baharlou, H., Canete, N.P., Cunningham, A.L., Harman, A.N., and Patrick, E. (2019). Mass Cytometry Imaging for the Study of Human Diseases-Applications and Data Analysis Strategies. Front. Immunol. *10*, 2657.

Bannon, D., Moen, E., Schwartz, M., Borba, E., Kudo, T., Greenwald, N., Vijayakumar, V., Chang, B., Pao, E., Osterman, E., et al. (2021). DeepCell Kiosk: scaling deep learning-enabled cellular image analysis with Kubernetes. Nat. Methods *18*, 43–45.

Battich, N., Stoeger, T., and Pelkmans, L. (2015). Control of Transcript Variability in Single Mammalian Cells. Cell *163*, 1596–1610.

Ben-Moshe, S., and Itzkovitz, S. (2019). Spatial heterogeneity in the mammalian liver. Nat. Rev. Gastroenterol. Hepatol. *16*, 395–410.

Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L., et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat. Commun. *9*, 2419.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.

Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. Nat. Methods *17*, 159–162.

Burkhard, S.B., and Bakkers, J. (2018). Spatially resolved RNA-sequencing of the embryonic heart identifies a role for Wnt/β-catenin signaling in autonomic control of heart rate. Elife *7*.

Bykov, I., Ylipaasto, P., Eerola, L., and Lindros, K.O. (2004). Functional Differences between Periportal and Perivenous Kupffer Cells Isolated by Digitonin-Collagenase Perfusion. Comp. Hepatol. *3 Suppl 1*, S34.

Cai, Z., Cao, C., Ji, L., Ye, R., Wang, D., Xia, C., Wang, S., Du, Z., Hu, N., Yu, X., et al. (2020). RIC-seq for global *in situ* profiling of RNA-RNA spatial interactions. Nature *582*, 432–437.

Chen, F., Wassie, A.T., Cote, A.J., Sinha, A., Alon, S., Asano, S., Daugharthy, E.R., Chang, J.-B., Marblestone, A., Church, G.M., et al. (2016). Nanoscale imaging of RNA with expansion microscopy. Nat. Methods *13*, 679–684.

Chen, G., Ning, B., and Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front. Genet. *10*, 317.

Chen, J., Suo, S., Tam, P.P., Han, J.-D.J., Peng, G., and Jing, N. (2017). Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. Nat. Protoc. *12*, 566–580.

Chen, J., Ding, L., Viana, M.P., Lee, H., Filip Sluezwski, M., Morris, B., Hendershott, M.C., Yang, R., Mueller, I.A., and Rafelski, S.M. (2020a). The Allen Cell and Structure Segmenter: a

new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090.

Chen, W.-T., Lu, A., Craessaerts, K., Pavie, B., Sala Frigerio, C., Corthout, N., Qian, X., Laláková, J., Kühnemund, M., Voytyuk, I., et al. (2020b). Spatial Transcriptomics and In situ Sequencing to Study Alzheimer's Disease. Cell *182*, 976–991.e19.

Chen, X., Sun, Y.-C., Church, G.M., Lee, J.H., and Zador, A.M. (2018). Efficient *in situ* barcode sequencing using padlock probe-based BaristaSeq. Nucleic Acids Res. *46*, e22.

Chen, Z., Soifer, I., Hilton, H., Keren, L., and Jojic, V. (2020c). Modeling Multiplexed Images with Spatial-LDA Reveals Novel Tissue Microenvironments. J. Comput. Biol.

Codeluppi, S., Borm, L.E., Zeisel, A., La Manno, G., van Lunteren, J.A., Svensson, C.I., and Linnarsson, S. (2018). Spatial organization of the somatosensory cortex revealed by osmFISH. Nat. Methods *15*, 932–935.

Edfors, F., Hober, A., Linderbäck, K., Maddalo, G., Azimi, A., Sivertsson, Å., Tegel, H., Hober, S., Szigyarto, C.A.-K., Fagerberg, L., et al. (2018). Enhanced validation of antibodies for research applications. Nat. Commun. *9*, 4130.

Edsgärd, D., Johnsson, P., and Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. Nat. Methods *15*, 339–342.

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

Femino, A.M., Fay, F.S., Fogarty, K., and Singer, R.H. (1998). Visualization of single RNA transcripts *in situ*. Science *280*, 585–590.

Ferrell, J.E., Jr (2012). Bistability, bifurcations, and Waddington's epigenetic landscape. Curr. Biol. *22*, R458–R466.

Foreman, R., and Wollman, R. (2020). Mammalian gene expression variability is explained by underlying cell state. Mol. Syst. Biol. *16*, e9146.

Frieda, K.L., Linton, J.M., Hormoz, S., Choi, J., Chow, K.-H.K., Singer, Z.S., Budde, M.W., Elowitz, M.B., and Cai, L. (2017). Synthetic recording and *in situ* readout of lineage information in single cells. Nature *541*, 107–111.

Friedman, S.L. (2008). Hepatic stellate cells: protean, multifunctional, and enigmatic cells of the liver. Physiol. Rev. *88*, 125–172.

Gerdes, M.J., Sevinsky, C.J., Sood, A., Adak, S., Bello, M.O., Bordwell, A., Can, A., Corwin, A., Dinn, S., Filkins, R.J., et al. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. Proc. Natl. Acad. Sci. U. S. A. *110*, 11982–11987.

Giesen, C., Wang, H.A.O., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P.J., Grolimund, D., Buhmann, J.M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nat. Methods *11*, 417–422.

Goh, J.J.L., Chou, N., Seow, W.Y., Ha, N., Cheng, C.P.P., Chang, Y.-C., Zhao, Z.W., and Chen, K.H. (2020). Highly specific multiplexed RNA imaging in tissues with split-FISH. Nat. Methods.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. Cell *174*, 968–981.e15.

Gut, G., Herrmann, M.D., and Pelkmans, L. (2018). Multiplexed protein maps link subcellular organization to cellular states. Science *361*.

Gyllborg, D., Langseth, C.M., Qian, X., Choi, E., Salas, S.M., Hilscher, M.M., Lein, E.S., and Nilsson, M. (2020). Hybridization-based *in situ* sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. Nucleic Acids Res.

Halpern, K.B., Shenhav, R., Matcovitch-Natan, O., Toth, B., Lemze, D., Golan, M., Massasa, E.E., Baydatch, S., Landen, S., Moor, A.E., et al. (2017). Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature *542*, 352–356.

Halpern, K.B., Shenhav, R., Massalha, H., Toth, B., Egozi, A., Massasa, E.E., Medgalia, C., David, E., Giladi, A., Moor, A.E., et al. (2018). Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. Nat. Biotechnol. *36*, 962–970.

Hu, K.H., Eichorst, J.P., McGinnis, C.S., Patterson, D.M., Chow, E.D., Kersten, K., Jameson, S.C., Gartner, Z.J., Rao, A.A., and Krummel, M.F. (2020). ZipSeq: barcoding for real-time mapping of single cell transcriptomes. Nat. Methods *17*, 833–843.

Ijsselsteijn, M.E., van der Breggen, R., Farina Sarasqueta, A., Koning, F., and de Miranda, N.F.C.C. (2019). A 40-Marker Panel for High Dimensional Characterization of Cancer Immune Microenvironments by Imaging Mass Cytometry. Front. Immunol. *10*, 2534.

Jackson, H.W., Fischer, J.R., Zanotelli, V.R.T., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., et al. (2020). The single-cell pathology landscape of breast cancer. Nature *578*, 615–620.

Jiang, Y., Zhang, N.R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. Genome Biol. *18*, 74.

Kann, A.P., and Krauss, R.S. (2019). Multiplexed RNAscope and immunofluorescence on whole-mount skeletal myofibers and their associated stem cells. Development *146*.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. Nat. Methods *10*, 857–860.

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S.-R., Kurian, A., Van Valen, D., West, R., et al. (2018). A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. Cell *174*, 1373–1387.e19.

Keren, L., Bosse, M., Thompson, S., Risom, T., Vijayaragavan, K., McCaffrey, E., Marquez, D., Angoshtari, R., Greenwald, N.F., Fienberg, H., et al. (2019). MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. Sci Adv *5*, eaax5851.

Kishi, J.Y., Lapan, S.W., Beliveau, B.J., West, E.R., Zhu, A., Sasaki, H.M., Saka, S.K., Wang, Y., Cepko, C.L., and Yin, P. (2019). SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nat. Methods *16*, 533–544.

Korem, Y., Szekely, P., Hart, Y., Sheftel, H., Hausser, J., Mayo, A., Rothenberg, M.E., Kalisky, T., and Alon, U. (2015). Geometry of the Gene Expression Space of Individual Cells. PLoS Comput. Biol. *11*, e1004224.

Kwon, S., Chin, K., and Nederlof, M. (2020). Simultaneous Detection of RNAs and Proteins with Subcellular Resolution. In RNA-Chromatin Interactions: Methods and Protocols, U.A.V. Ørom, ed. (New York, NY: Springer US), pp. 59–73.

Lander, A.D., Lo, W.-C., Nie, Q., and Wan, F.Y.M. (2009). The measure of success: constraints, objectives, and tradeoffs in morphogen-mediated patterning. Cold Spring Harb. Perspect. Biol. *1*, a002022.

Larsson, A.J.M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O.R., Reinius, B., Segerstolpe, Å., Rivera, C.M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. Nature.

Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., Jung, S., and Amit, I. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. Cell *159*, 1312–1326.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., et al. (2014). Highly multiplexed subcellular RNA sequencing *in situ*. Science *343*, 1360–1363.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R., Turczyk, B.M., Yang, J.L., Lee, H.S., Aach, J., et al. (2015). Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat. Protoc. *10*, 442–458.

Lin, J.-R., Fallahi-Sichani, M., and Sorger, P.K. (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. Nat. Commun. *6*, 8390.

Lin, J.-R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P.M., Santagata, S., and Sorger, P.K. (2018). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t CyCIF and conventional optical microscopes. Elife *7*.

Littman, R., Hemminger, Z., Foreman, R., Arneson, D., Zhang, G., Gómez-Pinilla, F., Yang, X., and Wollman, R. (2020). JSTA: joint cell segmentation and cell type annotation for spatial transcriptomics.

Lovatt, D., Ruble, B.K., Lee, J., Dueck, H., Kim, T.K., Fisher, S., Francis, C., Spaethling, J.M., Wolf, J.A., Grady, M.S., et al. (2014). Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. Nat. Methods *11*, 190–196.

Lundberg, E., and Borner, G.H.H. (2019). Spatial proteomics: a powerful discovery tool for cell biology. Nat. Rev. Mol. Cell Biol. *20*, 285–302.

Maiser, A., Dillinger, S., Längst, G., Schermelleh, L., Leonhardt, H., and Németh, A. (2020). Super-resolution *in situ* analysis of active ribosomal DNA chromatin organization in the nucleolus. Sci. Rep. *10*, 7462.

Maniatis, S., Äijö, T., Vickovic, S., Braine, C., Kang, K., Mollbrink, A., Fagegaltier, D., Andrusivová, Ž., Saarenpää, S., Saiz-Castro, G., et al. (2019). Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science *364*, 89–93.

Merritt, C.R., Ong, G.T., Church, S.E., Barker, K., Danaher, P., Geiss, G., Hoang, M., Jung, J., Liang, Y., McKay-Fleisch, J., et al. (2020). Multiplex digital spatial profiling of proteins and RNA in fixed tissue. Nat. Biotechnol. *38*, 586–599.

Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M., and Van Valen, D. (2019). Deep learning for cellular image analysis. Nat. Methods.

Moffitt, J.R., Hao, J., Wang, G., Chen, K.H., Babcock, H.P., and Zhuang, X. (2016a). High throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence *in situ* hybridization. Proc. Natl. Acad. Sci. U. S. A. *113*, 11046–11051.

Moffitt, J.R., Hao, J., Bambah-Mukku, D., Lu, T., Dulac, C., and Zhuang, X. (2016b). High performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. Proc. Natl. Acad. Sci. U. S. A. *113*, 14456–14461.

Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science *362*.

Moor, A.E., Harnik, Y., Ben-Moshe, S., Massasa, E.E., Rozenberg, M., Eilam, R., Bahar Halpern, K., and Itzkovitz, S. (2018). Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis. Cell *175*, 1156–1167.e15.

Motta, P.M. (1998). Marcello Malpighi and the foundations of functional microanatomy. Anat. Rec. *253*, 10–12.

Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. Science *328*, 876–878.

Murray, P.J. (2017). Macrophage Polarization. Annu. Rev. Physiol. *79*, 541–566.

Nichterwitz, S., Chen, G., Aguila Benitez, J., Yilmaz, M., Storvall, H., Cao, M., Sandberg, R., Deng, Q., and Hedlund, E. (2016). Laser capture microscopy coupled with Smart-seq2 for precise spatial transcriptomic profiling. Nat. Commun. *7*, 12139.

Palade, G.E., and Porter, K.R. (1954). Studies on the endoplasmic reticulum. I. Its identification in cells *in situ*. J. Exp. Med. *100*, 641–656.

Perkel, J.M. (2019). Starfish enterprise: finding RNA patterns in single cells. Nature *572*, 549–551.

Ptacek, J., Locke, D., Finck, R., Cvijic, M.-E., Li, Z., Tarolli, J.G., Aksoy, M., Sigal, Y., Zhang, Y., Newgren, M., et al. (2020). Multiplexed ion beam imaging (MIBI) for characterization of the tumor microenvironment across tumor types. Lab. Invest.

Qian, X., Harris, K.D., Hauling, T., Nicoloutsopoulos, D., Muñoz-Manchado, A.B., Skene, N., Hjerling-Leffler, J., and Nilsson, M. (2020). Probabilistic cell typing enables fine mapping of closely related cell types *in situ*. Nat. Methods *17*, 101–106.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods *5*, 877–879.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467.

Rouhanifard, S.H., Mellis, I.A., Dunagin, M., Bayatpour, S., Jiang, C.L., Dardani, I., Symmons, O., Emert, B., Torre, E., Cote, A., et al. (2018). ClampFISH detects individual nucleic acid molecules using click chemistry-based amplification. Nat. Biotechnol.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al. (2020). The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell *181*, 236–249.

Saka, S.K., Wang, Y., Kishi, J.Y., Zhu, A., Zeng, Y., Xie, W., Kirli, K., Yapp, C., Cicconet, M., Beliveau, B.J., et al. (2019). Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. Nat. Biotechnol. *37*, 1080–1090.

Salmén, F., Ståhl, P.L., Mollbrink, A., Navarro, J.F., Vickovic, S., Frisén, J., and Lundeberg, J. (2018). Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. Nat. Protoc. *13*, 2501–2534.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Schulz, D., Zanotelli, V.R.T., Fischer, J.R., Schapiro, D., Engler, S., Lun, X.-K., Jackson, H.W., and Bodenmiller, B. (2018). Simultaneous Multiplexed Imaging of mRNA and Proteins with Subcellular Resolution in Breast Cancer Tissue Samples by Mass Cytometry. Cell Syst *6*, 25–36.e5.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. Neuron *92*, 342–357.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2017). seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus. Neuron *94*, 752–758.e1.

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet. *20*, 631–656.

Stark, Z., Schofield, D., Alam, K., Wilson, W., Mupfeki, N., Macciocca, I., Shrestha, R., White, S.M., and Gaff, C. (2017). Prospective comparison of the cost-effectiveness of clinical whole exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. Genet. Med. *19*, 867–874.

Stickels, R.R., Murray, E., Kumar, P., Li, J., Marshall, J.L., Di Bella, D.J., Arlotta, P., Macosko, E.Z., and Chen, F. (2020). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. Nat. Biotechnol.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nat. Methods *18*, 100–106.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

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

Thrane, K., Eriksson, H., Maaskola, J., Hansson, J., and Lundeberg, J. (2018). Spatially Resolved Transcriptomics Enables Dissection of Genetic Heterogeneity in Stage III Cutaneous Malignant Melanoma. Cancer Res. *78*, 5970–5979.

Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single Molecule RNA FISH. Cell Syst *6*, 171–179.e5.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.

Tsai, T.Y.-C., Sikora, M., Xia, P., Colak-Champollion, T., Knaut, H., Heisenberg, C.-P., and Megason, S.G. (2020). An adhesion code ensures robust pattern formation during tissue morphogenesis. Science *370*, 113–116.

Tutucci, E., and Singer, R.H. (2020). Simultaneous Detection of mRNA and Protein in S. cerevisiae by Single-Molecule FISH and Immunofluorescence. In RNA Tagging: Methods and Protocols, M. Heinlein, ed. (New York, NY: Springer US), pp. 51–69.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for *in situ* tissue profiling. Nat. Methods *16*, 987–990.

Waddington, C.H. (1957a). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology. With an Appendix by H. Kacser.

Waddington, C.H. (1957b). The Strategy of the Genes. Allen.

Wang, C., Lu, T., Emanuel, G., Babcock, H.P., and Zhuang, X. (2019). Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization. Proc. Natl. Acad. Sci. U. S. A. *116*, 10842–10851.

Wang, F., Flanagan, J., Su, N., Wang, L.-C., Bui, S., Nielson, A., Wu, X., Vo, H.-T., Ma, X.-J., and Luo, Y. (2012). RNAscope: a novel *in situ* RNA analysis platform for formalin-fixed, paraffin embedded tissues. J. Mol. Diagn. *14*, 22–29.

Wang, G., Ang, C.-E., Fan, J., Wang, A., Moffitt, J.R., and Zhuang, X. (2020). Spatial organization of the transcriptome in individual neurons.

Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science *361*.

Wang, Y., Woehrstein, J.B., Donoghue, N., Dai, M., Avendaño, M.S., Schackmann, R.C.J., Zoeller, J.J., Wang, S.S.H., Tillberg, P.W., Park, D., et al. (2017). Rapid Sequential *in situ* Multiplexing with DNA Exchange Imaging in Neuronal Cells and Tissues. Nano Lett. *17*, 6131–6139.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*.

Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell *177*, 1873–1887.e17.

Xia, C., Babcock, H.P., Moffitt, J.R., and Zhuang, X. (2019a). Multiplexed detection of RNA using MERFISH and branched DNA amplification. Sci. Rep. *9*, 7721.

Xia, C., Fan, J., Emanuel, G., Hao, J., and Zhuang, X. (2019b). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. Proc. Natl. Acad. Sci. U. S. A.

Young, A.P., Jackson, D.J., and Wyeth, R.C. (2020). A technical review and guide to RNA fluorescence *in situ* hybridization. PeerJ *8*, e8806.

Yuste, R., Hawrylycz, M., Aalling, N., Aguilar-Valles, A., Arendt, D., Arnedillo, R.A., Ascoli, G.A., Bielza, C., Bokharaie, V., Bergmann, T.B., et al. (2020). A community-based transcriptomics classification and nomenclature of neocortical cell types. Nat. Neurosci.

Zhang, M., Eichhorn, S.W., Zingg, B., Yao, Z., Zeng, H., Dong, H., and Zhuang, X. (2020). Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by *in situ* single-cell transcriptomics.

Zhu, Q., Shah, S., Dries, R., Cai, L., and Yuan, G.-C. (2018). Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence *in situ* hybridization data. Nat. Biotechnol.

(2014). Method of the year 2013. Nat. Methods *11*, 1.

Chapter 2

Joint cell segmentation and cell type annotation for spatial transcriptomics

Littman, Russell; Hemminger, Zachary; Foreman, Robert; Arneson, Douglas; Zhang, Guanglin; Gómez-Pinilla, Fernando; Yang, Xia; Wollman, Roy

Abstract

RNA hybridization-based spatial transcriptomics provides unparalleled detection sensitivity. However, inaccuracies in segmentation of image volumes into cells cause misassignment of mRNAs which is a major source of errors. Here, we develop JSTA, a computational framework for joint cell segmentation and cell type annotation that utilizes prior knowledge of cell type-specific gene expression. Simulation results show that leveraging existing cell type taxonomy increases RNA assignment accuracy by more than 45%. Using JSTA, we were able to classify cells in the mouse hippocampus into 133 (sub)types revealing the spatial organization of CA1, CA3, and Sst neuron subtypes. Analysis of within cell subtype spatial differential gene expression of 80 candidate genes identified 63 with statistically significant spatial differential gene expression across 61 (sub)types. Overall, our work demonstrates that known cell type expression patterns can be leveraged to improve the accuracy of RNA hybridization-based spatial transcriptomics while providing highly granular cell (sub)type information. The large number of newly discovered spatial gene expression patterns substantiates the need for accurate spatial transcriptomic measurements that can provide information beyond cell (sub)type labels.

Introduction

Spatial transcriptomics has been employed to explore the spatial and cell type-specific gene expression to better understand physiology and disease (Asp et al, 2019; Burgess, 2019).

Compared to other spatial transcriptomics methods, RNA hybridization-based approaches provided the highest RNA detection accuracies with capture rates > 95% (Lubeck et al, 2014). With the development of combinatorial approaches for RNA hybridization, the ability to measure the expression of hundreds to thousands of genes makes hybridization-based methods an attractive platform for spatial transcriptomics (Beucher, 1979; Najman & Schmitt, 1994; Al-Kofahi et al, 2010; Lubeck et al, 2014; Chen et al, 2015; Eng et al, 2019; preprint: Park et al, 2019; Vu et al, 2019; preprint: Petukhov et al, 2020; Qian et al, 2020; Yuste et al, 2020). Nonetheless, unlike dissociative approaches, such as single-cell RNA sequencing (scRNAseq) where cells are captured individually, RNA hybridization-based approaches have no a priori information of which cell a measured RNA molecule belongs to. Segmentation of image volumes into cells is therefore required to convert RNA detection into spatial single-cell data. Assigning mRNA to cells remains a challenging problem that can substantially compromise the overall accuracy of combinatorial FISH approaches.

Generation of spatial single-cell data from imaging-based spatial transcriptomics relies on algorithmic segmentation of images into cells. Current combinatorial FISH work uses watershed-based algorithms with nuclei as seeds, and the total mRNA density to establish cell borders (Najman & Schmitt, 1994; Chen et al, 2015; Eng et al, 2019). Watershed algorithm was proposed more than 40 years ago (Beucher, 1979), and newer segmentation algorithms that utilize state of the art machine learning approaches have been shown to improve upon classical watershed approach (Al-Kofahi et al, 2010; Vu et al, 2019). However, their performance is inherently bounded by the quality of the “ground truth” dataset used for training. In tissue regions with dense cell distributions, there is simply not enough information in the images to perform accurate manual labeling and create a sufficiently accurate ground truth training datasets. Therefore, there is an urgent need for new approaches that can combine image

information with external datasets to improve image segmentation and thereby the overall accuracy of spatial transcriptomics.

Due to the deficiency in existing image segmentation algorithms, a few segmentation-free spatial transcriptomic approaches were proposed. pciSeq's primary goal is to assign cell types to nuclei by using proximity to mRNA, and an initialized segmentation map to compute the likelihood of each cell type (Qian et al, 2020). Similarly, SSAM creates cell type maps based on RNA distributions, without creating a cell segmentation map because it ignores cellular boundaries (preprint: Park et al, 2019). Therefore, while both pciSeq and SSAM leverage cell type catalogs to provide insights into the spatial distribution of different cell types, they do not produce a high-quality cell segmentation map. More recently, an approach for updating cell boundaries in spatial transcriptomics data has been developed (preprint: Petukhov et al, 2020). Baysor uses neighborhood composition vectors and Markov random fields to segment spatial transcriptomics data and identify cell type clusters.

Here, we present JSTA, a computational framework for jointly determining cell (sub)types and assigning mRNAs to cells by leveraging previously defined cell types through scRNAseq. Our approach relies on maximizing the internal consistency of pixel assignment into cells to match known expression patterns. We compared JSTA to watershed in assigning mRNAs to cells through simulation studies to evaluate their accuracy. Application of JSTA to MERFISH measurements of gene expression in the mouse hippocampus together with Neocortical Cell Type Taxonomy (NCTT) (Yuste et al, 2020) provides a highly granular map of cell (sub)type spatial organization and identified many spatially differentially expressed genes (spDEGs) within these (sub)types (Lein et al, 2007).

Results

Our computational framework of JSTA is based on improving initial watershed segmentation by incorporating cell (sub)type probabilities for each pixel and iteratively adjusting the assignment of boundary pixels based on those probabilities (Fig 2.1A).

To evaluate JSTA, we chose to use the mouse hippocampus for two reasons: (i) The mouse hippocampus has high cell (sub)type diversity as it includes more than 35% of all cell (sub)types defined by the NCTT. (ii) The mouse hippocampus has areas of high and low cell density. These two reasons make the mouse hippocampus a good test case for the hypothesis that external cell (sub)type-specific expression data could be leveraged to increase the accuracy of spatial transcriptomics, as implemented in our approach. We performed multiplexed error robust fluorescent in situ hybridization (MERFISH) of 163 genes which include 83 selected cell marker genes, which show distinct expression between cell types and are used for cell classification and segmentation and 80 genes previously implicated with biological importance in traumatic brain injury (Fig 2.1B). Combining this MERFISH dataset, DAPI stained nuclei, and the NCTT reference dataset using JSTA, we created a segmentation map that assigns all mRNAs to cells while simultaneously classifying all cells into granular (sub)types based on NCTT.

In JSTA, we leverage the NCTT information to infer probabilities at the pixel level. However, learning these probabilities from NCTT is challenging for two reasons. (i) NCTT data were acquired with scRNAseq technology that has higher sparsity due to low capture rates and needs to be harmonized. (ii) NCTT data provide expression patterns at the cell level and not the pixel level. We expect the mean expression among all pixels in a cell to be the same as that of the whole cell. Yet, variance and potentially higher distribution moments of the pixel-level distribution are likely different from those of the cell-level distribution due to sampling and

biological factors such as variability in subcellular localization of mRNA molecules (Eng et al, 2019). To address these issues, JSTA learns the pixel-level cell (sub)type probabilities using two distinct deep neural network (DNN) classifiers, a cell-level type classifier, and a pixel type classifier. Overall, JSTA learns three distinct layers of information: segmentation map, pixel-level classifier, and cell-level classifier.

Learning of model parameters is done using a combination of NCTT and the MERFISH data. The cell type classifier is learned directly from NCTT data after harmonization. The other two layers are learned iteratively using expectation maximization (EM) approach (Chen et al, 2015). Given the current cell type assignment to cells, we train a pixel-level DNN classifier to output the cell (sub)type probability of each pixel. JSTA can be applied on any user-selected subset of the genes; the local mRNA density of these selected genes around each pixel is used as the input for the pixel-level classifier. The selection of genes drives how well the cell type classifier can distinguish between distinct cell types. The updated pixel classifier is used to assign probabilities to all border pixels. The new probabilities are then used to “flip” border pixels' assignment based on their type probabilities. The updating of the segmentation map requires an update of the cell-level type classification which triggers a need for an update of pixel-level classifier training. This process is then repeated until convergence. Analysis of the mean pixel-level cell (sub)type classification accuracy shows an increase in the algorithm's classification confidence over time demonstrating that the NCTT external information gets iteratively incorporated into the tasks of cell segmentation and type annotation (Fig 2.1.1). For computational efficiency, we iterate between training, reassignment, and reclassification in variable rates. As this approach uses cell type information to improve border assignment between neighboring cells, in cases where two neighboring cells are of the same type, the border between them will stay the same as the initial watershed segmentation. The final result is a cell type segmentation map that is initialized based on watershed and adjusted to allow pixels

to be assigned to cells to maximize consistency between local RNA density and cell type expression priors.

Performance evaluations

Performance evaluation using simulated hippocampus data

To test the performance of our approach, we utilized synthetic data generated based on the NCTT (Lein et al, 2007) (Fig 2.2A and B). Details on the synthetic generation of cell position, morphologies, type, and expression profiles are available in the Materials and Methods section. Using this synthetic data, we evaluated the performance of JSTA in comparison with watershed at different cell type granularities. For example, two cells next to each other that are of subtypes CA1sp1 and CA1sp4 would add to the error in segmentation, but if the cell type resolution decreases to CA1 cells, these would be considered the same type, and misassignment of mRNA between these cells is no longer penalized. Evaluating the methods in this manner allows us to explore the trade-off between cell type granularity and mRNA assignment accuracy. Our analysis shows that JSTA consistently outperforms watershed at assigning spots to cells (Fig 2.2C). Interestingly, the benefit of JSTA was evident even with a small number of genes (Fig 2.2D). With just 12 genes, the performance jumps to 0.50 at the highest cell type granularity, which is already higher than watershed's accuracy; at a granularity of 16 cell types, the accuracy reached 0.62 (Fig 2.2C and D). Overall the synthetic data showed that JSTA outperforms watershed approach, and at physiologically relevant parameters, can increase mRNA assignment accuracy by > 45%. We additionally compared JSTA to pciSeq (Qian et al, 2020), in the assignment of mRNA molecules to cells. We note that pciSeq is mainly designed to assign cell types to nuclei based on surrounding mRNA and therefore is not primarily focused on assigning most mRNA molecules to cells as JSTA does. Furthermore, since pciSeq is not designed to operate on 3D data, we simulated 2D data and applied both JSTA and pciSeq. We

found that JSTA was more accurate at assigning mRNA molecules to cells than pciSeq (Fig 2.1.1A). pciSeq tends to incorrectly assign many spots to background, as segmentation is not its primary goal. However, when ignoring mRNAs assigned to background in a true-positive calculation, pciSeq performs well as it primarily assigns mRNAs close to the nuclei, which is an easier task. In this case, JSTA has comparable performance (Fig 2.1.1B).

Time requirements of JSTA

We simulated data of different sizes and ran JSTA to determine how the run time scales with larger datasets. We simulated three replicates of data with a width and height of 100, 200, 300, 400, 500, and 1,000 μm. The run time of JSTA scales linearly with both the area and number of cells in the section (Fig 2.1A and B).

Performance evaluation using empirical spatial transcriptomics of mouse hippocampus

We next tested the performance of JSTA using empirical data and evaluated its ability to recover the known spatial distribution of coarse neuron types across the hippocampus (Fig 2.3). First, we subset the NCTT scRNAseq data to the shared genes we have in our MERFISH data and harmonized the MERFISH and scRNAseq datasets (Moffitt et al, 2018). Using the cell type annotations from the single-cell data, we trained a DNN to classify cell types. As expected, our classifier derived a cell type mapping agreeing with known spatial patterns in the hippocampus (Fig 2.3A). For example, CA1, CA3, and DG cells were found with high specificity to their known subregions (Fig 2.3B). We found that the gene expression of the segmented cells in MERFISH data highly correlated with their scRNAseq counterparts, and displayed similar correlation patterns between different cell types (Fig 2.3C) as seen in scRNAseq data (Fig 2.3D). These results show that our data and JSTA algorithm can recover existing knowledge on the spatial distribution of cell types and their gene expression patterns in the mouse hippocampus.

JSTA performs high-resolution cell type mapping in the mouse hypothalamic preoptic region

We applied JSTA to a MERFISH dataset from a previously published mouse hypothalamic preoptic region with 134 genes provided (Moffitt et al, 2018). Using the provided scRNAseq reference dataset, we accurately mapped 87 high-resolution cell types in this region (Fig 2.3.1A). The mapped cell types follow spatial distributions of high-resolution cell types of this region previously annotated through clustering and marker gene annotation. We find the gene expression profiles of the cell types from the MERFISH data are highly correlated with their scRNAseq counterparts (Fig 2.3.1B).

JSTA performs high-resolution cell type mapping in the mouse somatosensory cortex

Next, we applied JSTA to an osmFISH dataset from the mouse somatosensory cortex with the 35 genes provided (Codeluppi et al, 2018). Using the NCTT reference, we mapped 142 high-resolution cell types in this region. We found that the glutamatergic neuronal populations follow known spatial organization (Fig 2.3.2A) and that the gene expression patterns of high-resolution cell types in the osmFISH data are highly correlated with their NCTT counterparts (Fig 2.3.2B).

Applications of JSTA for biological discovery

JSTA identifies spatial distribution of highly granular cell (sub)types in the hippocampus

A key benefit of JSTA is its ability to jointly segment cells in images and classify them into highly granular cell (sub)types. Our analysis of mouse hippocampus MERFISH data found that these subtypes, defined only based on their gene expression patterns, have high spatial localization in the hippocampus. From lateral to medial hippocampus, the subtypes transitioned spatially from CA1sp10 to CA1sp6 (Fig 2.4A). Likewise, JSTA revealed a non-uniform distribution of subtypes in the CA3 region. From lateral to medial hippocampus, the subtypes transitioned from CA3sp4 to CA3sp6 (Fig 2.4B). This gradient of subtypes reveals a high level of spatial organization and points to potentially differential roles for these subtypes.

JSTA shows that spatially proximal cell subtypes are transcriptionally similar

Next, we tested whether across different cell types spatial patterns match their expression patterns by evaluating the colocalization of cell subtypes and their transcriptional similarity. Indeed, spatially proximal CA1 subtypes showed high transcriptional similarity (Figs 2.5A and 2.5.1A and B). For example, cells in the subtypes CA1sp3, CA1sp1, and CA1sp6 are proximal to each other and show a high transcriptional correlation. Interestingly, this relationship was not bidirectional, and transcriptional similarity by itself is not necessarily predictive of spatial proximity. For example, subtypes CA1sp10, CA1sp7, and CA1sp4 show > 0.95 correlation but are not proximal to each other. Similar findings were seen in the CA3 region as well (Figs 2.5B and, 2.5.1A and B).

To test whether this principle goes beyond subtypes of the same type, we compared CA1 neurons and the Sst interneurons. We found that many Sst subtypes have high specificity in their localization and are transcriptionally related to their non-Sst neighbors. Using permutation tests, we found that subtypes Sst12, Sst19, Sst20, Sst28 are significantly colocalized with these same subtypes and are specific to the CA1 region (Fig 2.5C and D, Materials and Methods). Analysis of their transcriptional similarity showed that these subtypes are highly correlated in their gene expression to all CA1 subtypes (Fig 2.5E) but not to CA3 subtypes. These results show that both within a cell type and across cell types spatial proximity indicate similarity in expression patterns.

JSTA identifies spatial differential gene expression

Given our results on the relationship between spatial localization and gene expression patterns across cell subtypes, we next tested whether spDEGs within the highly granular cell subtypes can be identified. We focused our analysis on the 80 genes in our dataset that were not genes used to classify cells into cell (sub)types. We identified spDEGs by determining if the

spatial expression pattern of a given gene was statistically different from a null distribution by permuting the gene expression values. Importantly, the null model was restricted to the permutation of only the cells within that subtype. As a result, our spDEG analysis specifically identifies genes whose expression within a specific subtype has a spatial distribution that is different than random. We found that within hippocampal cell subtypes, many genes were differentially expressed based on their location (Fig 2.6). For example, Tox in CA1sp1 shows higher expression on the medial side of the hippocampus and decreases to the lateral side. Leng8 in subtype CA3sp3 is highly expressed closer to the CA1 region and lower in the medial CA3. Hecw1 in the DG2 subtype has varying spatial distribution in the DG region. The lower portion of the DG has clusters of higher expression, while the upper portion has lower expression. These spatial differences in gene expression are not limited to neuronal subtypes. Astrocyte subtype “Astro1” shows spatial heterogeneity in expression of Thra, with large patches of high expression levels and other patches of little to no expression (Fig 2.2.6A). Overall, we tested for spDEGs in 61 (sub)types with more than 40 cells. We found that all 61 of the tested hippocampal cell subtypes have spDEGs (Figs 2.6B and 2.6.1B), with more than 50% (63 of 80) of the tested genes showing non-random spatial pattern (Figs 2.6C and 2.6.1C). Certain genes also show spatial patterns in many subtypes (e.g., Thra 2.6.1ac), while others are more specific to a one or a few subtypes (e.g., Farp1, 2.6.1ac). Identification of spDEGs highlights an interesting application of highly accurate cell type and mRNA assignment in spatial transcriptomic data.

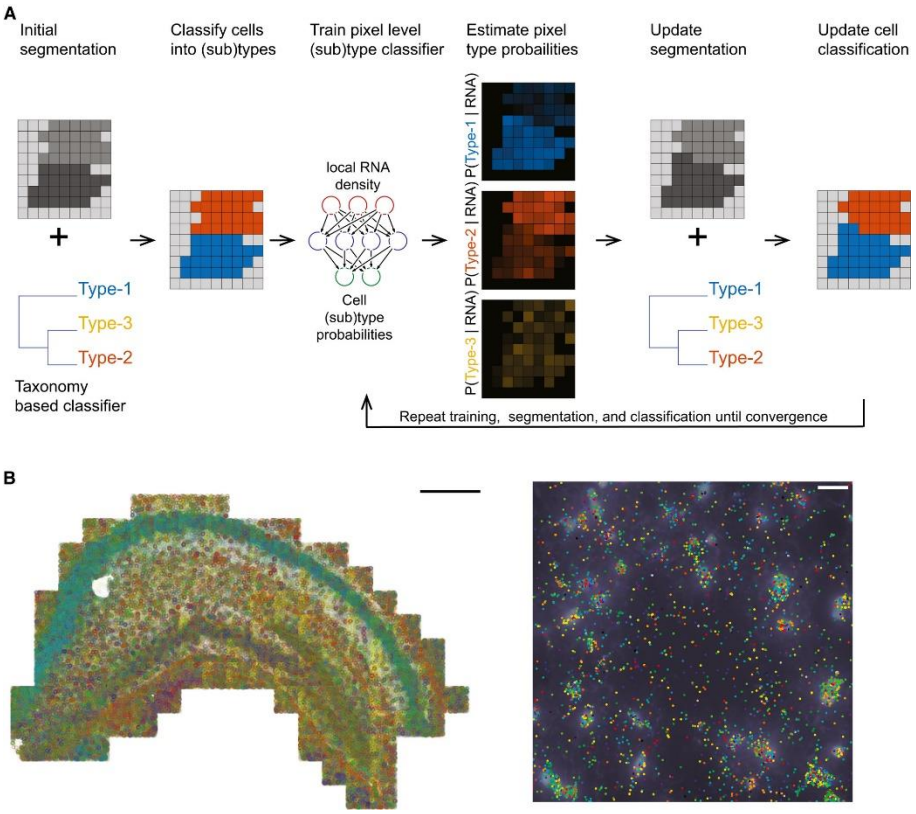

Figure 2.1: Overview of JSTA and the spatial transcriptomic data used for performance evaluation

A. Joint cell segmentation and cell type annotation (JSTA) overview. Initially, watershed-based segmentation is performed and a cell-level type classifier is trained based on the Neocortical Cell Type Taxonomy (NCTT) data. The deep neural network (DNN) parameterized cell-level classifier then assigns cell (sub)types (red and blue in this cartoon example). Based on the current assignment of pixels to cell (sub)types, a new DNN is trained to estimate the probabilities that each pixel comes from each of the possible (sub)types given the local RNA density at each pixel. In this example, two pixels that were initially assigned to the “red” cells got higher probability to be of a blue type. Since the neighbor cell is of type “blue”, they were reassigned to that cell during segmentation update. Using the updated segmentation and the cell type classifier cell types are reassigned. The tasks of training, segmentation, and classification are repeated over many iterations until convergence.B. Multiplexed error robust fluorescent in situ hybridization (MERFISH) and DAPI stained nuclei in the mouse hippocampus. Each gene is represented by a different color. For the entire hippocampus (left), only the mRNA spots are shown with a scale bar of 500 μm. On the zoomed-in section (right), each gene is represented by a different color dot, and the DAPI intensity is displayed in white. The scale bar is 20 μm.

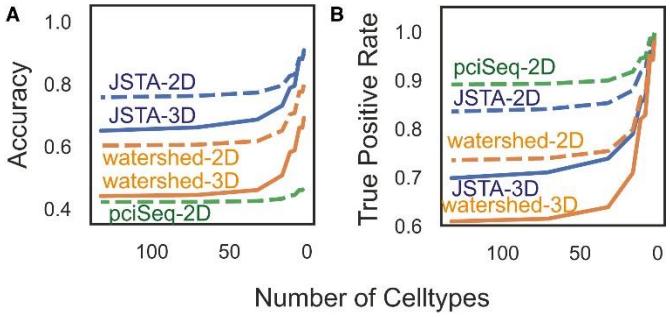

Figure 2.1.1: Performance evaluation of JSTA, pciSeq, and watershed.
A, B.pciSeq is unable to run on 3D data (solid line), so we simulated additional 2D data (dotted line). We evaluated these methods on the performance of accuracy of assigning mRNA to the correct cell (A). JSTA is more accurate than pciSeq on the accuracy metric. pciSeq is not very accurate here, because many mRNA are incorrectly assigned to background. We additionally tested these methods on their performance of assigning mRNA to the correct cell while ignoring mRNA assigned to background (B). pciSeq is highlighted here, because it mainly assigns spots close to the nucleus; JSTA is comparable.

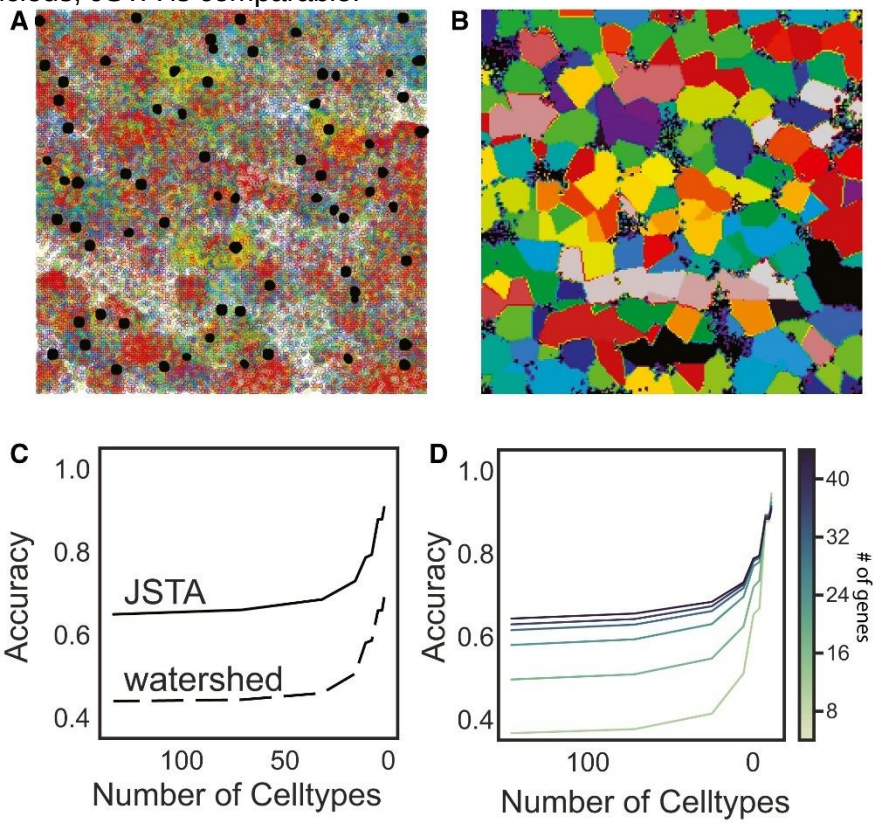

Figure 2.2: Performance evaluation of JSTA using simulated data.
A. Representative synthetic dataset of nuclei (black) and mRNAs, where each color represents a different gene. B. Ground truth segmentation map of the cells in the representative synthetic dataset. Each color represents a different cell. C. Average Accuracy of calling mRNA spots to cells at different cell type resolutions using 83 genes across 10 replicates. Accuracy was determined by the assignment of each mRNA molecule to the correct cell type. JSTA (solid line) is more accurate than watershed (dashed line) at assigning mRNA molecules to the correct cells (FDR < 0.05). Statistical significance was determined with a Mann–Whitney test and false discovery rate correction. D. Accuracy (as described in (C)) of

calling mRNA spots to cells when using JSTA to segment cells with a lower selection of cell type marker genes (8–44 genes tested). The color of the line gets progressively darker as the number of genes used increases.

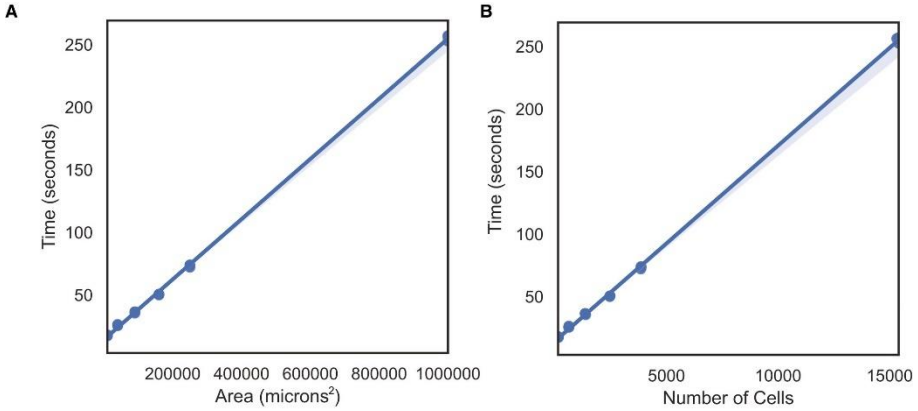

Figure 2.2.1: Application of JSTA to osmFISH data from the mouse somatosensory cortex.
A. Glutamatergic neurons are consistent with previously identified spatial patterns of the somatosensory cortex. B. JSTA-mapped high-resolution (sub)types are correlated with their NCTT counterparts in terms of gene expression patterns (Table 2.3.2). Cell types with at least five cells were kept.

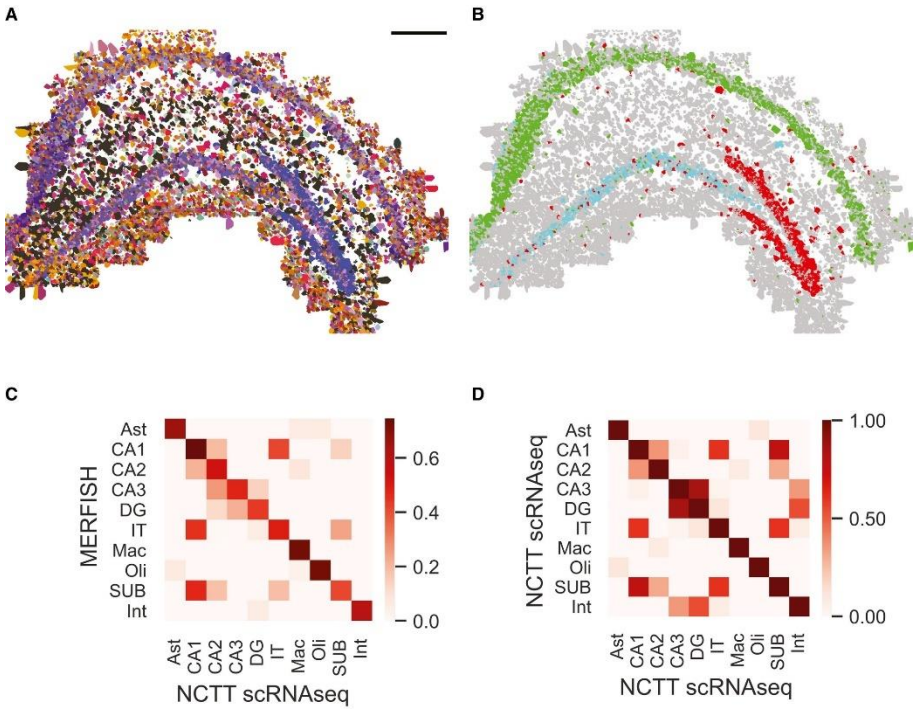

Figure 2.3: Segmentation of MERFISH data from the hippocampus using JSTA.
A. High-resolution cell type map of 133 cell (sub)types segmented and annotated by JSTA. Colors match those defined by Neocortical Cell Type Taxonomy (NCTT). Scale bar is 500 μm. B. JSTA-based classification of CA1 (green), CA3 (cyan), and DG (red) neurons matches their known domains. C. Correlation of the average expression of 163 genes across major cell types between MERFISH measurements to scRNAseq data from NCTT. D. Correlation of the average

expression of the same genes as in (C) between expression of types in scRNAseq data from NCTT. The correlation structure in panel (C) closely mirrors the structure in panel (D).

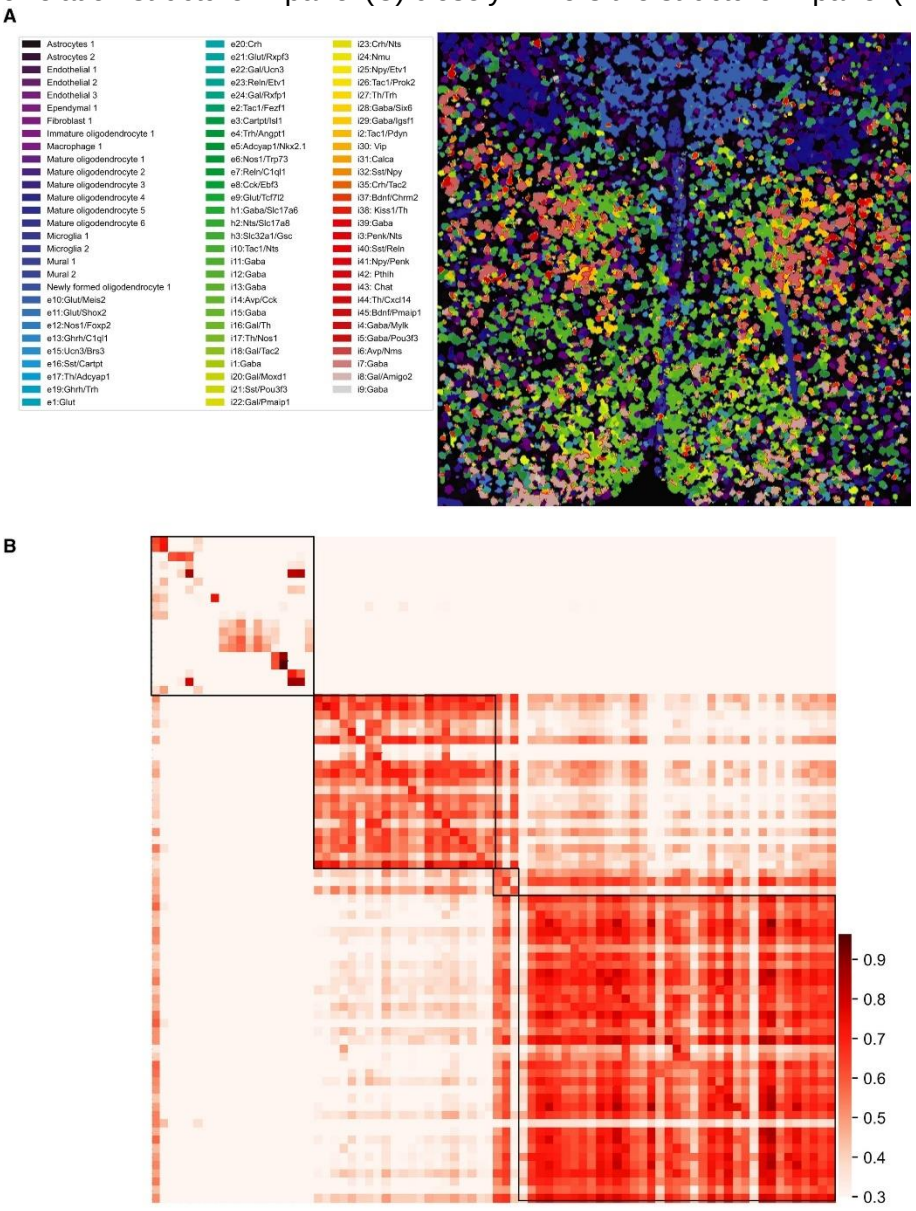

Figure 2.3.1: Run time evaluation of JSTA on simulated data.
A. B. We ran JSTA on data simulated with a width and height of 100, 200, 300, 400, 500, and 1,000 μm, with three replicates each. We evaluated the time taken to run JSTA by the area of the section (A), and the number of cells in each section (B).

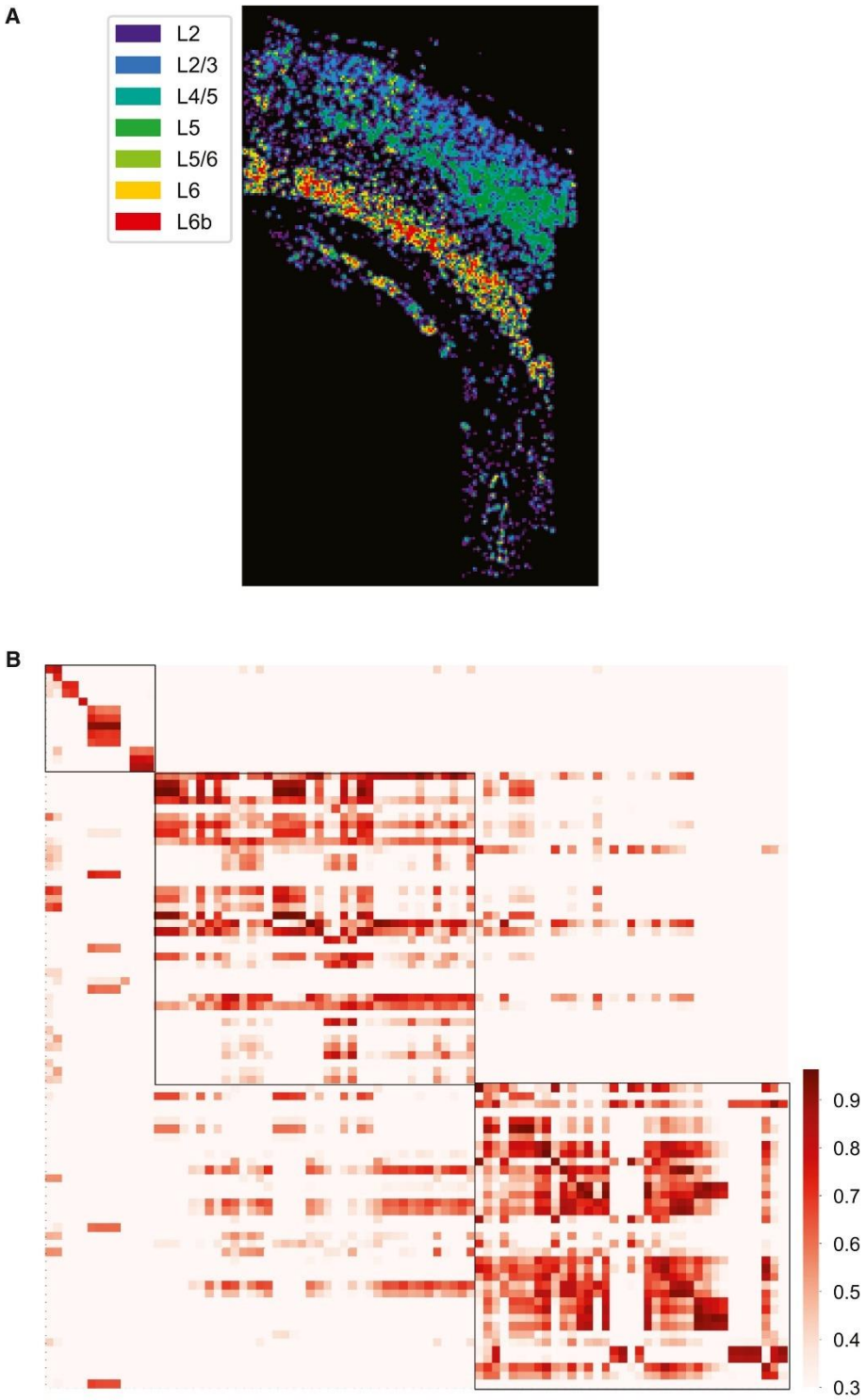

Figure 2.3.2: Application of JSTA to MERFISH data from the mouse hypothalamic preoptic region.
A. High-resolution cell types identified by JSTA. The spatial mappings of these high-resolution cell types are consistent with the manually annotated data from Moffit et al (2018). B. JSTA-mapped high-resolution (sub)types are highly correlated with their scRNAseq reference counterparts in terms of gene expression patterns (Table 2.3.1). Cell types with at least five cells were kept.

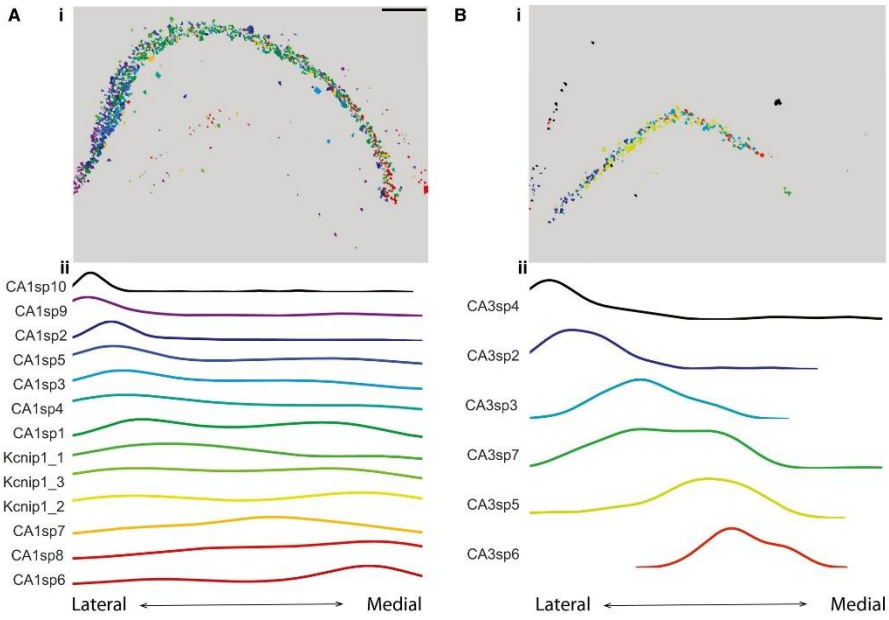

Figure 2.4: Spatial distribution of neuronal subtypes in the hippocampus.
A. (i) Cell subtype map of CA1 neurons in the hippocampus as annotated by JSTA. Scale bar is 500 μm. Distribution of CA1 subtypes in the hippocampus, computed by projecting cell centers to the lateral to medial axis. CA1 neuronal subtypes show a non-uniform distribution across the whole CA1 region. (ii) Smoothed histogram highlighting the density of CA1 subtypes across the CA1 region. B. (i) Cell subtype map of CA3 neurons in the hippocampus as annotated by JSTA. Distribution of CA3 subtypes in the hippocampus, computed by projecting the cell centers to the lateral to medial axis. CA3 neuronal subtypes show a non-uniform distribution across the whole CA3 region. (ii) Smoothed histogram highlighting the density of CA3 subtypes across the CA3 region.

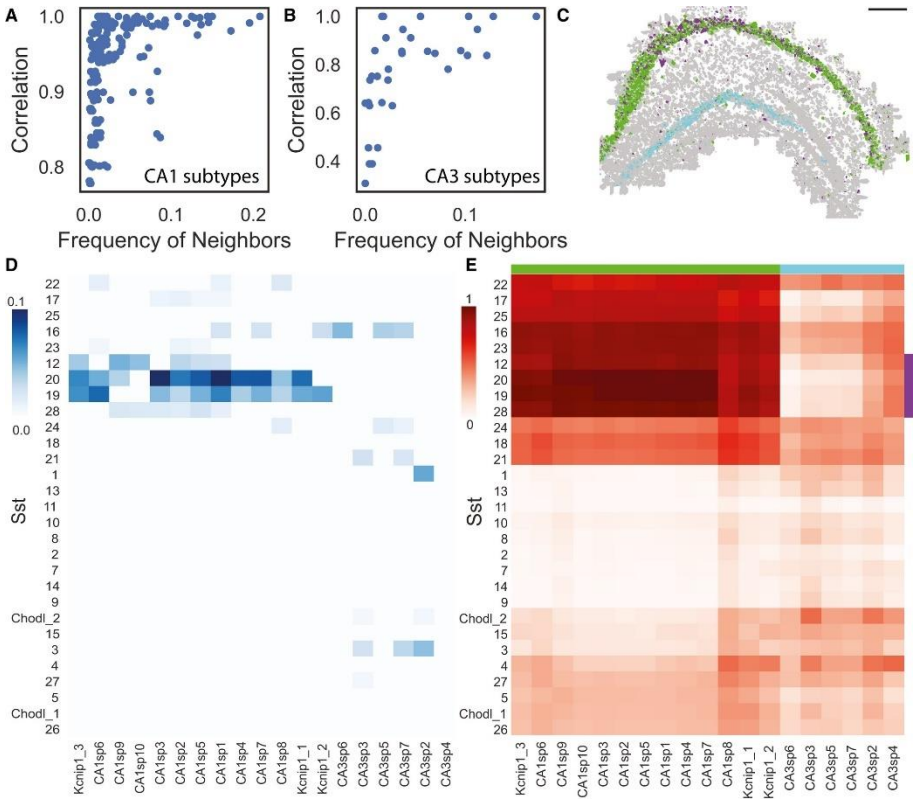

Figure 2.5: Agreement between spatial proximity and gene coexpression in highly granular cell subtypes in the hippocampus.

A, B. Relationship between the frequency of a (sub)type's neighbors and its transcriptional Pearson correlation between CA1 subtypes (A) and between CA3 subtypes (B). C. Cell type map in the hippocampus shows specific colocalization patterns between a subset of Sst subtypes (purple) and CA1 neurons (green); these Sst subtypes do not colocalize with CA3 neurons (cyan). Scale bar is 500 μm. D. Colocalization patterns of Sst subtypes with CA1 and CA3 subtypes. Sst subtypes that colocalize with the CA1 subtypes have high transcriptional similarity. Colocalization was defined as the percent of neighbors that are of that subtype (Materials and Methods). E. Transcriptional correlation patterns between Sst subtypes and CA1 and CA3 neurons. Green, purple and cyan sidebars highlight the subset of Sst colocalized with CA1 (purple), CA1 (green), and CA3 (cyan).

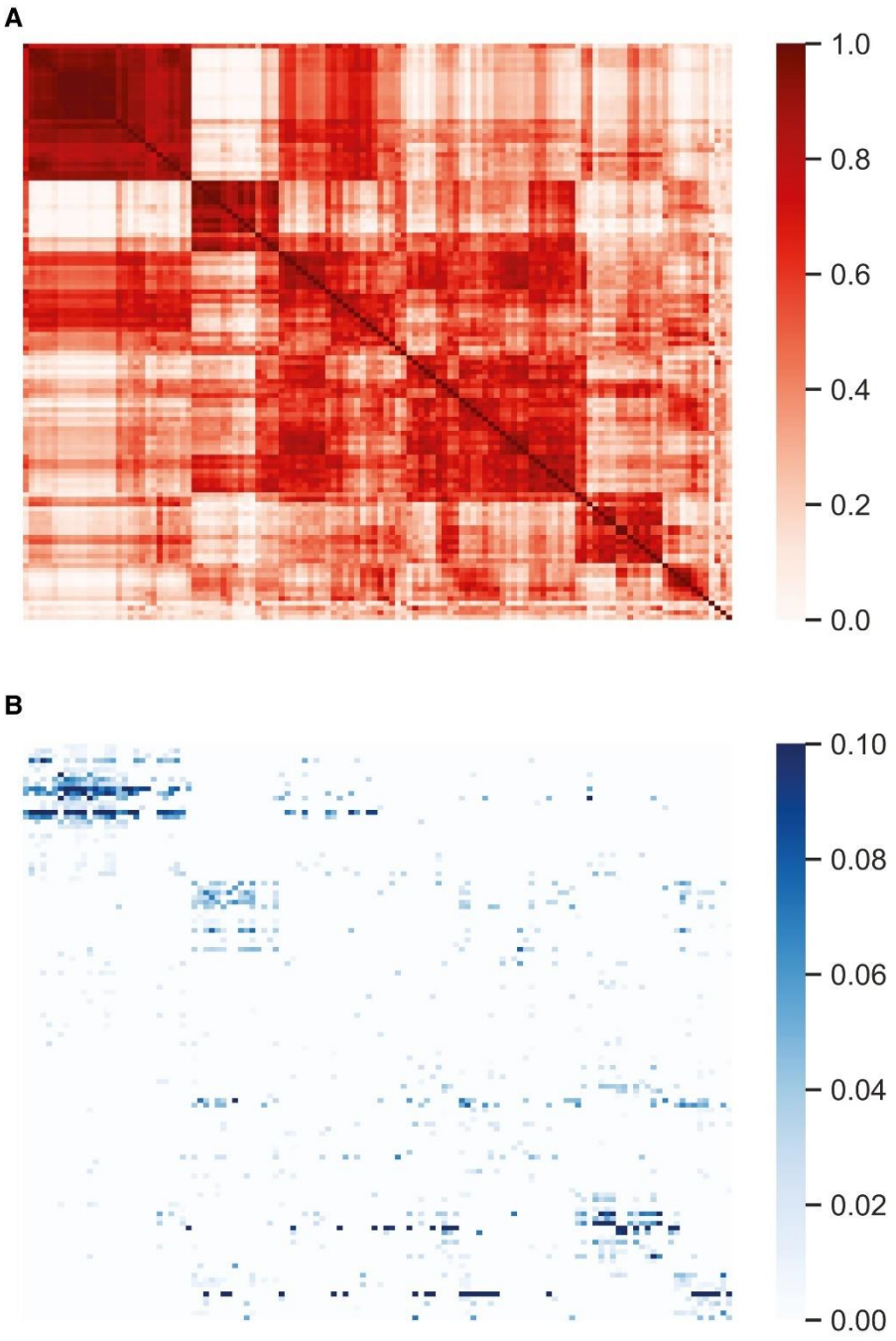

Figure 2.5.1: Correlation structure of cell types compared with their colocalization
Neuronal subtypes that are highly colocalized are often correlated in their gene expression. Cell types with more than 10 cells were included. A. Pearson correlation of 122 (sub)types across 83 selected genes. B. Frequency of neighbors between each of 122 (sub)types. Only significant (FDR < 0.05) colocalizations are shown. Labels and values are detailed in Tables 2.5.1 and 2.6.1.

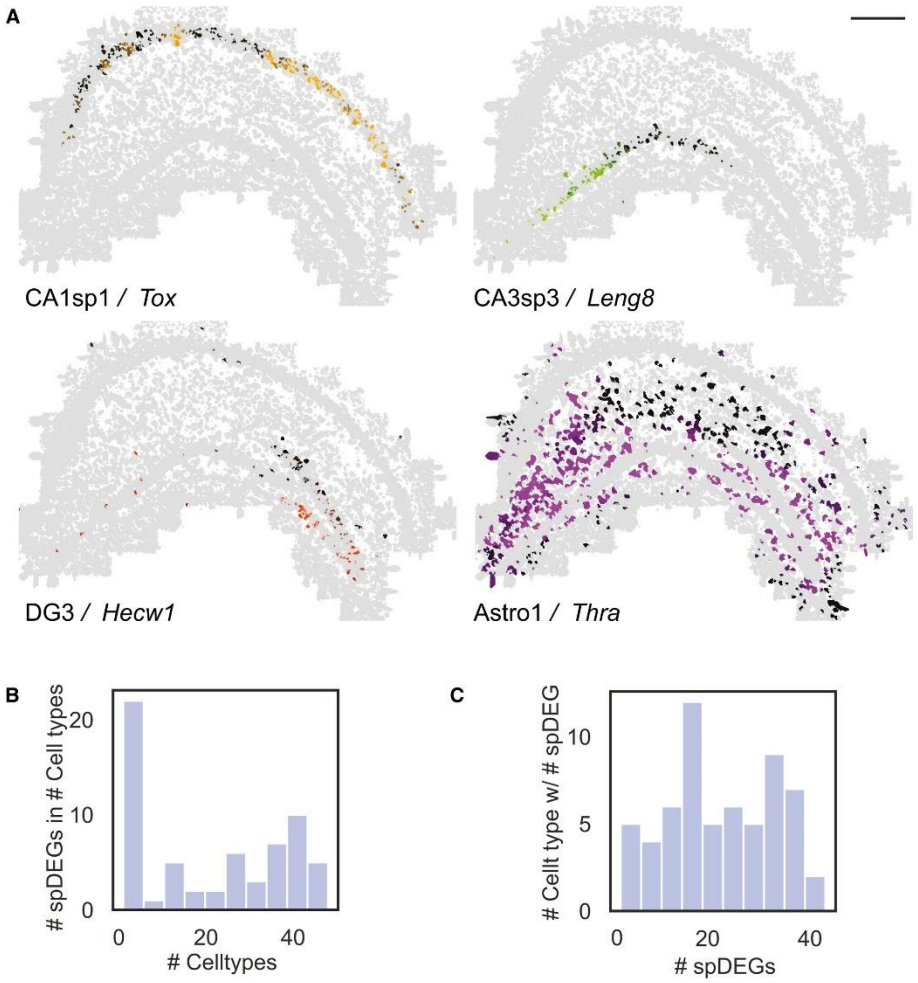

Figure 2.6: Identification of spatial differential gene expression (spDEGs).
A. Normalized expression of Tox in CA1sp1, Leng8 in CA3sp3, Hecw1 in DG3, and Thra in Astro1 shows variable expression throughout the hippocampus. Scale bar is 500 μm. spDEGs were computed by comparing the true variance in gene expression between cell subtype neighborhoods to that of randomly permuted cell (sub)type neighborhoods. B. Histogram of the number of statistically significant spDEGs (Benjamini–Hochberg-corrected FDR < 0.05) in each subtype. C. Histogram of the number of subtypes that have an spDEG for each gene.

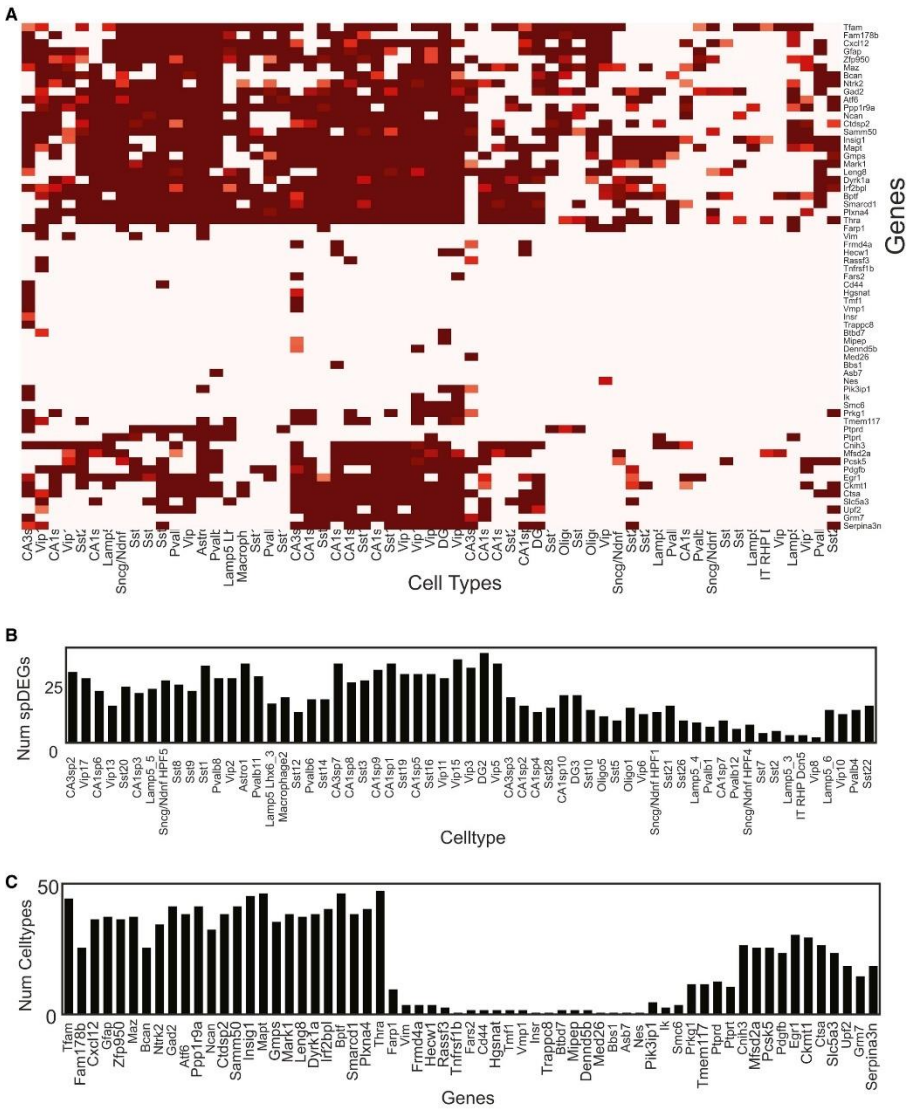

Figure 2.6.1: Identification of spatial differentially expressed genes (spDEGs).
A. spDEGs were computed by comparing the true variance in gene expression between cell subtype neighborhoods to that of randomly permuted cell (sub)type neighborhoods. B. 63 genes across 61 cell types show significant spDEGs. Heatmap values correspond to −log2(P-value). C. Number of spDEGs in each of the 61 cell types. D. Number of cell types with each of the 63 spDEGs.

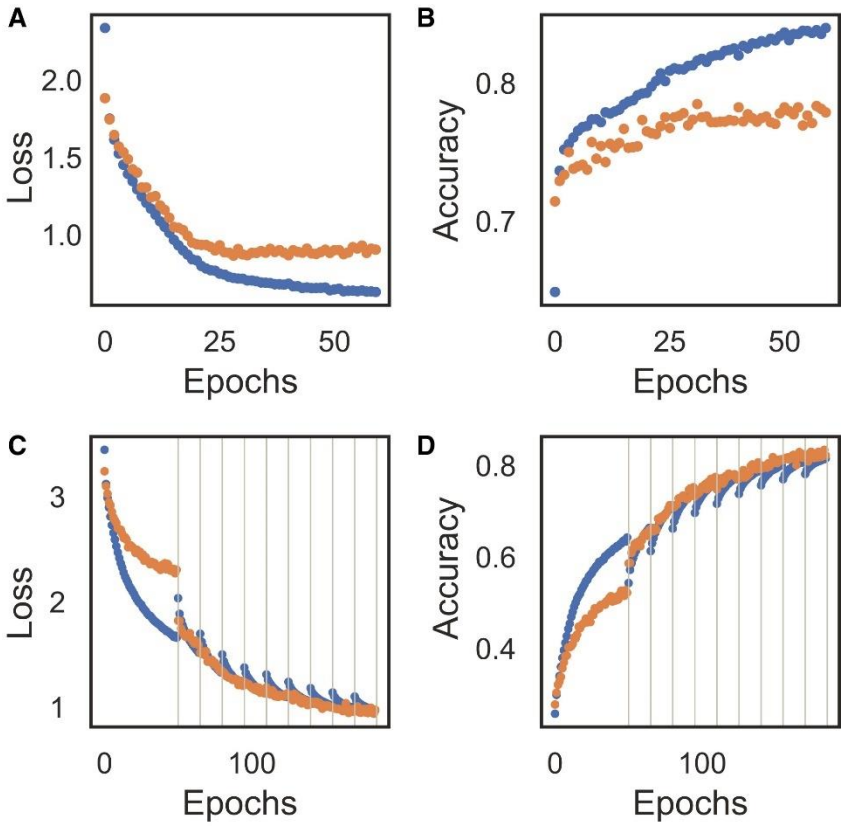

Figure 2.6.2: Cross-entropy loss and accuracy of cell type (A, B) and pixel (C, D) classifier during training for the train (blue) and validation (orange) datasets.
A, B. Cross-entropy (A) loss and accuracy (B) during training cell type classifier. The cell type classifier overfits the training data and is mitigated by stopping training after 40 epochs. C, D. Cross-entropy loss (C) and accuracy (D) during training of the pixel classifier. Black lines indicate new training iteration after pixel reassignment.

Discussion

Spatial transcriptomics provides the coordinates of each transcript without any information on the transcript cell of origin (Lee, 2017). Here, we present JSTA, a new method to convert raw measurements of transcripts and their coordinates into spatial single-cell expression maps. The key distinguishing aspect of our approach is its ability to leverage existing scRNAseq-based reference cell type taxonomies to simultaneously segment cells, classify cells into (sub)types, and assign mRNAs to cells. The unique integration of spatial transcriptomics with existing scRNAseq information to improve the accuracy of image segmentation and enhance the biological applications of spatial transcriptomics, distinguishes our approach from

other efforts that regardless of their algorithmic ingenuity are bounded by the available information in the images themselves. As such, JSTA is not a generalist image segmentation algorithm rather a tool specifically designed to convert raw spatial transcriptomic data into single cell-level spatial expression maps. We show the benefits of using a dedicated analysis tool through the insights it provides into spatial organization of distinct (sub)types in the mouse hippocampus and the hundreds of newly discovered cell (sub)type-specific spDEGs. These insights into the molecular- and cellular-level structural architecture of the hippocampus demonstrate the types of biological insights provided by highly accurate spatial transcriptomics.

The promise of single cell and spatial biology lends itself to intense focus on technological and computational development and large-scale data collection efforts. We anticipate that JSTA will benefit these efforts while at the same time benefit from them. On the technology side, we have demonstrated the performance of JSTA for two variants of spatial transcriptomics, MERFISH and osmFISH. However, the algorithm is extendable and could be applied to other spatial transcriptomic approaches that are based on in situ sequencing (Lee et al, 2014; Lee et al, 2015; Turczyk et al, 2020), subcellular spatial barcoding (Ståhl et al, 2016; Salmén et al, 2018), and potentially any other spatial “omics” platforms (Gerdes et al, 2013; Lin et al, 2015; Goltsev et al, 2018; Keren et al, 2018; Lin et al, 2018; Lundberg & Borner, 2019). Additionally, cell segmentation results from JSTA can be used as input for other tools such as GIOTTO (Dries et al, 2021) and TANGRAM (preprint: Biancalani et al, 2020) to facilitate single cell and spatial transcriptomic data analysis. The benefits of JSTA are evident even with a small number of measured genes. This indicates that it is applicable to a broad range of platforms across all multiplexing capabilities. JSTA is limited by its ability to harmonize technical differences between spatial transcriptomic data modalities and the scRNAseq reference. Harmonization between datasets is an active area of research, and JSTA will benefit from these advances (preprint: Lopez et al, 2019; Stuart et al, 2019; Welch et al, 2019; Abdelaal et al,

2020; Tran et al, 2020). JSTA relies on initial seed identification (nuclei or cell centers), and incorrect identification can lead to split or merged cells. JSTA currently does not split or merge cells, but this postprocessing step could be added to further improve segmentation (Chaudhuri & Agrawal, 2010; Surut & Phukpattaranont, 2010; Correa-Tome & Sanchez-Yanez, 2015; Gamarra et al, 2019). On the data side, as JSTA leverages external reference data, it will naturally increase in its performance as both the quality and quantity of reference cell type taxonomies improve (HuBMAP Consortium, 2019). We see JSTA as a dynamic analysis tool that could be reapplied multiple times to the same dataset each time external reference data is updated to always provide highest accuracy segmentation, cell (sub)type classification, spDEG identification.

Due to the nascent status of spatial transcriptomics, there are many fundamental questions related to the interplay between cell (sub)types and other information gleaned from dissociative technologies and tissue and organ architecture (Trapnell, 2015; Mukamel & Ngai, 2019). Our results show that strong codependency between spatial position and transcriptional state of a cell in the hippocampus, these results mirror findings from other organs (Halpern et al, 2017; Moor et al, 2018; Egozi et al, 2020). This codependency supports the usefulness of the reference taxonomies that were developed without the use of spatial information. Agreements between cell type taxonomies developed solely based on scRNAseq and other measurement modalities, i.e., spatial position, corroborate the relevance of the taxonomical definitions created for mouse brain (Yuste et al, 2020). At the same time, the spatial measurements demonstrate the limitation of scRNAseq. We discovered many spatial expression patterns within most cell (sub)types that prior to these spatial measurements would have been considered biological heterogeneity or even noise but in fact they represent key structural features of brain organization. High accuracy mapping at the molecular and cellular level will allow us to bridge

cell biology with organ anatomy and physiology pointing toward a highly promising future for spatial biology.

Materials and Methods

Tissue preparation

All experiments were performed in accordance with the United States National Institutes of Health Guide for the Care and Use of Laboratory Animals and were approved by the University of California at Los Angeles Chancellor's Animal Research Committee. B6 mouse was euthanized using carbon dioxide with cervical dislocation. Its brain was harvested and flash-frozen in Optimal Cutting Temperature Compound (OCT) using liquid nitrogen. 15 μm sections were prepared and placed on pretreated coverslips.

Coverslip functionalization

Coverslips were functionalized to improve tissue adhesion and promote gel attachment (Moffitt & Zhuang, 2016). Briefly, 40 mm No.1 coverslips were cleaned with a 50:50 mixture of concentrated 37% hydrochloric acid and methanol under sonication for 30 min. Coverslips were silanized to improve gel adhesion with 0.1% triethylamine and 0.2% allyltrichlorosiloxane in chloroform under sonication for 30 min then rinsed once with chloroform then twice with ethanol. Silanization was cured at 70°C for 1 h. An additional coating of 2% aminopropyltriethoxysilane to improve tissue adhesion was applied in acetone under sonication for 2 min then washed twice with water and once with ethanol. Coverslips were dried at 70°C for 1 h then stored in a desiccator for less than 1 month.

Probe design and synthesis

A total of 18 readout probes were used to encode the identity of each gene. Each gene was assigned four of the possible 18 probes such that each combination was a minimum

hamming distance of 4 away from any other gene. This provides classification that is robust up to 2-bit errors. 80–120 encoder probes were designed for each target gene. Encoder probes contained a 30 bp region complementary to the transcript of interest with a melting point of 65°C and less than 17 bp homology to off-target transcripts including highly expressed ncRNA and rRNA. Probes also contained three of four readout sequences assigned to each gene. Sequences are available in supplementary material. Probes were designed using modified MATLAB code developed by the Zhuang Lab (Moffitt & Zhuang, 2016). Probes were ordered from custom arrays as a single strand pool. A T7 promoter was primed into each sequence with a limited cycle qPCR to allow amplification through in vitro transcription and reverse transcription (Moffitt & Zhuang, 2016).

Hybridization

Hybridization was performed using a modified MERFISH protocol (Moffitt & Zhuang, 2016). Briefly, tissue sections were fixed in 4% PFA in 1xPBS for 15 min and washed three times with 1×PBS for 5 min each. Tissue was permeabilized with 1% Triton X-100 in 1×PBS for 30 min and washed three times with 1×PBS. Tissue was incubated in 30% formamide in 2×TBS at 37°C for 10 min. Encoding probes were hybridized at 5 nM per probe in 30% formamide 10% dextran sulfate 1 mg/ml tRNA 1 µM poly-T acridite anchor probed and 1% murine RNAse inhibitor in 2xTBS. A 30 µl drop of this encoding hybridization solution was placed directly on the coverslip, and a piece of parafilm was placed on the coverslip to prevent evaporation. Probes were hybridized for 30–40 h at 37°C in a humidity chamber. Tissue was washed twice with 30% formamide in 2×TBS for 30 min each at 45°C. Tissue was washed three times with 2×TBS. Tissue was embedded in a 4% polyacrylamide hydrogel with 0.5 µl/ml TEMED 5 µl 10% APS and 200 nm blue beads for 2 h. Tissue was cleared with 1% SDS, 0.5% Triton x-100, 1 mM EDTA, 0.8 M guanidine HCl 1% proteinase K in 2×TBS for 48 h at 37°C replacing clearing solution every 24 h. Sample was washed with 2×TBS and mounted for imaging. Readout

hybridization was automated using a custom fluidics system. Sample was rinsed with 2×TBS and buffer exchanged into 10% dextran sulfate in 2×TBS for hybridization. Hybridization was performed in 10% dextran sulfate in 2×TBS with a probe concentration of 3 nM per probe. Sample was washed with 10% dextran sulfate then 2×TBS. Sample chamber was filled with a 2 mM pca 0.1& rPCO 2 mM VRC 2 mM Trolox in 2×TBS Imaging Buffer. Sample was imaged at 63× using a custom epifluorescent microscope. After imaging, fluorophores were stripped using 50 mM TCEP in 2×TBS and the next round of readout probes was hybridized.

Image analysis

Image analysis was performed using custom python code (Wollman lab). To register multiple rounds of imaging together with subpixel resolution, fiduciary markers were found and a rigid body transformation was performed. Images were preprocessed using hot pixel correction, background subtraction, chromatic aberration correction, and deconvolution. An 18-bit vector was generated for each pixel where each bit represented a different round and fluorophore. Each bit was normalized so that background approached 0 and spots approached 1. An L2 normalization was applied to the vector, and the Euclidean distance was calculated to the 18-bit gene barcode vectors. Pixels were classified if their Euclidean distance was less than a 2-bit error away from the nearest gene barcode. Individual pixels that were physically connected were merged into a spot. Dim spots and spots that contained 1 pixel were removed.

Nuclei segmentation

Nuclei were stained using dapi and imaged after MERFISH acquisition. Each 2D image was segmented using cellpose with a flow threshold of 1 and a cell probability threshold of 0 (preprint: Stringer et al, 2020). 2D masks of at least 10 μm2 area were merged if there was at least 30 percent overlap between frames. 3D masks that were present in < 5 z frames (2 μm) were removed.

Simulation

scRNAseq reference preparation

The NCTT was subset to the cells found in the hippocampus and to the genes from our MERFISH data. Expression levels of simulated genes were taken from scRNAseq reference and were harmonized to qualitatively match the variance observed in measured in MERFISH data. These were then rounded to create a scaled count matrix. For each of the 133 hippocampal cell types from the NCTT, we computed a mean vector and covariance matrix of gene expression. We additionally computed the cell type proportions in the single-cell data for later use in cell type assignment.

Creating the cell map

Initially, the cell centers were placed in a 200 × 200 × 30 μm grid, equidistant from one another, with an average distance between cell centers of 4 μm. The cell centers were then moved around in each direction (x, y, z) based on a Gaussian function with mean 0 and standard deviation 0.6. Pixels were then assigned to their closest center with a minimum distance of 5 μm and maximum distance of 7 μm. Cells with less than 30 pixels were removed due to small unrealistic sizes. To create more realistic and non-round cells, we merged neighboring, touching cells twice. Each cell was assigned a (sub)type uniformly across all 133 types in our dataset. Nuclei were randomly placed within each cell with 20 pixels. Nuclei pixels placed on the border were removed. We simulated 10 independent replicates in each simulation study.

Generating cell transcriptional profiles and placing spots

Each cell's gene expression profile was drawn from a multivariate Gaussian using the mean vector, and covariance matrix computed from the scRNAseq reference. This vector and

matrix are cell type specific, and each cell's gene expression profile is sampled from these cell type-specific distributions. The mRNA spots were then placed inside of each cell, slightly centered around the nucleus, but mostly uniform throughout.

Simulated data on limited genes

To perform feature selection and extract a limited number of important genes (4, 12, 20, 28, 36, 44), we used a random forest classifier with 100 trees to predict cell types in the reference dataset. The top n important features for classifying cell types were used. Other simulation parameters were the same as above.

K-nearest neighbor-based density estimation method

We used a K-nearest neighbor approach to estimate density for many genes at each point (Wasserman, 2006). The volume required to reach the 5th spot was computed and used to compute the density estimation (equation 1). Where r is the radius to the 5th closest spot of that gene, we repeated this process for all genes.

JSTA overview

Expectation maximization can be used to jointly classify the identity of an observation of interest, while learning the parameters that describe the class distributions. In EM, the object classes are initialized with a best guess. The parameters of the classifying function are learned from this distribution of initialized classes (M-step). The objects are reclassified according to the updated function parameters (E-step). These steps are repeated until the function parameters converge. JSTA is designed with an EM approach for reclassifying border pixels in the 3-dimensional grid of pixels based on their estimated transcriptional densities. First, we initialize the spatial map with watershed, in Euclidean space with a maximum radius. Next, we classify cell types of the segmented cells based on the computed count matrix. We then randomly

sample a fraction of the pixels' gene expression vectors, and train a pixel classifier (M-step). The pixel classifier is used to reclassify the cell identity of pixels that are at the border between different cell types, or between a cell and empty space (E-step).

Cell type classification

Data preparation

To match the distributions of both scRNAseq and MERFISH, we centered and scaled each cell across all genes. We then subsequently centered and scaled each gene across all cells. We note that other harmonization approaches could be applied here.

Cell type classifier

We parameterized the cell type classifier as a neural network, with three intermediate layers with three times the number of input genes as nodes. We used a tanh activation function with L1 regularization (1e-4) allowing for the influence of negative numbers in the scaled values and parameter space sparsity (preprint: Bach et al, 2011). Batch normalization was used on each layer (preprint: Ioffe & Szegedy, 2015), and a softmax activation was used for the output layer (Goodfellow et al, 2016) (Table 1.1).

Training the classifier

The network parameters were initialized with Xavier initialization (Glorot & Bengio, 2010). The neural network was trained with two steps with learning rates of 5e-3 and 5e-4 for 20 epochs each, with batch size of 64, and the Adam optimizer was used (preprint: Kingma & Ba, 2014). A 75/25 train validation split was used to tune the L1 regularization parameter and reduce overfitting. We used 75/25 to increase the representation of lower frequency cell classes. Cross-entropy loss was used to penalize the model and update parameters accordingly (Fig 2.2.6.2A and B).

Pixel classification

Pixel classifier

We parameterized the pixel classifier as a neural network with three intermediate layers. Each layer was twice the size of the last to increase the modeling power of this network and indirectly model the other genes not in the MERFISH dataset. Each layer used the tanh activation function and used an l2 regularizer (1e-3). Each layer was centered and scaled with batch normalization, and the output activation was an l2 regularized softmax function (Table 2.1).

Training the classifier

Each time cell types are reclassified, a new network was reinitialized with Xavier initialization. The network was initially trained with learning rates or 1e-3 and 1e-4 for 25 epochs. After the first round of classifying and flipping the assignment of pixels, the network was retrained on a new sample of pixels starting from the previous parameter values. This was then trained with a learning rate of 1e-4 for 15 epochs. All training was performed with the Adam optimizer and a batch size of 64. We used an 80/20 train validation split to help monitor any overfitting that might be occurring, and adjust the hyperparameter selection accordingly. We used cross-entropy loss (Fig 2.2.6.2C and D).

Identifying border pixels

Border pixels are defined as pixels that are between two cells of different types, or between a cell and empty space. To enhance the smoothness of cells' borders, we require a border pixel to have 5 of its surroundings be from a different cell, and 2 of its surroundings be from the same cell.

Classifying pixels

The trained classifier was then used to estimate the cell type class of border pixels. The pixel classifier outputs a probability vector for each cell type, and the probabilities are scaled by a distance metric based on the distance to the cells' nuclei that it could flip to. Probabilities less than 0.05 are set to 0. The classification is sampled from that probability vector subset to cell types of its neighbors, and renormalized to 1. If the subset probability vector only contains 0, the pixel identity is set to background. To balance the exploration and exploitation of pixel classification map, we anneal the probability of selecting a non-maximum probability cell type by multiplying the maximum probability by (1 + number of iterations run*0.05). If this is selected as 0, complete stochasticity presides, and if it is large, the maximum probability will be selected.

JSTA formalization

Definitions and background

The gene expression level of nc cells and np pixels is described by the matrices Ec (cells) and Ep (pixels) which are nc × m and np × m matrices, respectively, where m is the number of genes. Likewise, cell type probability distributions of all cells or pixels can be described by matrices. These distributions for cells and pixels are Pc and Pp, respectively, represented as nc × k and np × k matrices, where k is the number of cell types. We aim to learn θ and ϕ, such that fθ and gϕ, accurately map from Ec to Pc and Ep to Pp. We used the cross-entropy loss function for penalizing our models.

Cell type classification

First, we learn the parameters of fθ by:

where Eref is an nref × m gene expression matrix representing the harmonized NCTT data and Tref is an nref vector of cell type labels provided by NCTT. We then use the newly learned mapping to infer the cell type probability distributions in the initialized dataset Ec with:

We classify each cell as the highest classification probability for that cell:

where Tc are the predicted cell types for each of the cells in the matrix Ec.

Joint pixel and parameter updates

We initialize the labels Tp for all pixels based on the current segmentation map that assigns pixels to cells. We then learn the parameters of the mapping function gϕ (maximization). Learning is performed by updating the parameters of the mapping function gϕ with:

The updated mapping function is then used to infer the probability of observing a type Tp given expression Ep in all pixels:

The next step is to update Pp based on spatial proximity to cells of each type. Using the notation q for the vector of probabilities of a single pixel (q = Ppj=[q0…, qi,… qk]), we next update the elements in the vector q based on neighborhood information. We scaled the values of qi based on its distance from the nuclei and its neighbors. q′ is intermediate in the calculation that does not represent true probabilities.

where r is the distance from the nucleus of the closest cell of cell type i, d is the distance threshold for which a pixel should automatically be assigned to that nucleus. The values 10 and 5 were determined empirically to modify the sharpness of probability decline based on distance. 10 was chosen to be much bigger than probabilities produced by gϕ, and 5 was chosen to allow the probability to decay to half over 5d.

We then only kept probabilities for cell types of neighboring cells:

We then used the intermediate q' to recalculate the pixel type probabilities:

The updated values per cell (qj) are then used to update the probability matrix Pp. The type per pixel (Tp). The assignment of pixel to cells is then stochastically assigned according to the inferred probability Pp per pixel basis.

We then repeat updating gϕ and Tp until convergence.

Segmentation

Density estimation

The 3-dimensional space was broken into a grid of pixels with the edge of each pixel 2 μm in length (1 μm in simulation). The density was estimated at the center of each pixel, for each gene. The volume required to reach five mRNA molecules was used as the denominator of the density estimation.

Segmentation with JSTA

The cell assignment map was initialized with watershed on the distance transform with a maximum distance from the nucleus of 2 μm. The cells were only classified once. The pixel classifier was trained six times (5 in simulation) on 10% of the pixels excluding pixels without assignment. After each training step, we reassigned pixels for 10 iterations (5 in simulation). The lowest probability kept in the predicted pixel assignment vector was 0.05 (0.01 in simulation).

Segmentation with watershed

The overall gene density was the sum of each gene in a given pixel. To smooth the range of the density, we log2 transformed the density values. Log-transformed density values less than 1 were masked. The segmentation used the nuclei locations as seeds and watershed from the skimage python package, with compactness of 10. Using compactness of 10 was the

highest performing value for watershed. A watershed line was used to separate cells from one another.

Evaluation of segmentation in simulated data

mRNA spot call accuracy was evaluated at different taxonomic levels. For a given cell, the accuracy was defined as the number of mRNA spots correctly assigned to that cell divided by the total number of mRNA spots assigned to that cell. To match the algorithm's ability to segment based on cell type information, RNAs that were assigned to a neighboring cell of the same (sub)type were also considered correct assignment. The overall segmentation accuracy was the mean accuracy across all cells in a given sample. To evaluate accuracy at different levels, we utilized the NCTT dendrogram. We used dendrogram heights at 0 through 0.8 with a step size of 0.05 (133, 71, 32, 16, 11, 8, 5, 4, 3, 2 cell types).

Correlation of segmented MERFISH with scRNAseq

The NCTT scRNAseq data were subset to the genes from our MERFISH data. Cells in the segmented MERFISH dataset were assigned to canonical hippocampus cell types (Astrocyte, CA1 pyramidal neuron, CA2 Pyramidal neuron, CA3 Pyramidal, Dentate Gyrus, Inferior temporal cortex, Macrophage, Oligodendrocyte, Subiculum, Interneuron) based on their high-resolution cell type classification. In each cell type, the average expression in each gene was calculated. Only genes were kept that had an average expression of at least five counts in one of the cell types. Values were centered and scaled across all cell types. The Pearson correlation was computed for each gene for the matching cell types between scRNAseq and MERFISH.

Distribution of high-resolution cell types in the hippocampus

CA1 and CA3 subtypes were projected onto the lateral medial axis. The smoothed density across this dimension was plotted for each of the subtypes.

Colocalization of high-resolution cell types

Significant colocalization of subtypes was determined through a permutation test. First, the 20 nearest cell types around each cell were determined. We counted the number of cells from each type that surround each cell type and computed the fraction of neighbors coming from each subtype. This created a matrix with the fraction of colocalizations per cell between each cell type combination. We then permuted the labels of the cell types 1,000 times and recomputed this interaction matrix to create a null distribution. For each cell type colocalization, we determined the percentage of colocalizations in the null distribution that is higher than the true colocalization number to create a P-value for each colocalization. We corrected for multiple testing with the Benjamini–Hochberg procedure and determined significance using FDR < 0.05.

Identification of spatial differential gene expression

spDEGs were calculated in cell types with more than 40 cells. Within each cell type, we computed a local expression of each gene for each cell. The local expression was the mean expression of a gene in the cell and its nine nearest neighbors. We then built a null distribution by permuting gene expression values within the cell type, and repeating the local expression process for 100 permutations. Determining if a gene was spatially differentially expressed, we compared the variance of the null distribution within a cell type with the variance of the true distribution of local expression to get a P-value. We corrected for multiple testing with Benjamini–Hochberg procedure and determined significance using FDR < 0.05.

Python packages used

python (3.8.3), numpy (1.18.5), pandas (1.0.5), matplotlib (3.2.2), scipy (1.5.0), scikit-learn (0.23.1), scikit-image (0.16.2), tensorflow (2.2.0). seaborn (0.10.1).

Data availability

Source code: GitHub (https://github.com/wollmanlab/JSTA; https://github.com/wollmanlab/PySpots).

Raw images: Figshare (https://doi.org/10.6084/m9.figshare.14531553).

Acknowledgements

The work was funded by NIH grant R01NS117148 and T32CA201160.

Results in this chapter were adapted from a manuscript published in *Molecular Systems Biology*:

Littman, R. **et al.** Joint cell segmentation and cell type annotation for spatial transcriptomics. Mol. Syst. Biol. 17, e10108 (2021)

Author Contributions

ZH&RF performed sample preparation and data acquisition. ZH performed data processing. RL performed data analysis. RF&DL designed probe set. RL&RW wrote the manuscript. RW,RL&ZH designed the project.

References

Abdelaal T, Mourragui S, Mahfouz A, Reinders MJT (2020) SpaGE: spatial gene enhancement using scRNA-seq.Nucleic Acids Res48:e107

Al-Kofahi Y, Lassoued W, Lee W, Roysam B (2010) Improved automatic detection and segmentation of cell nuclei in histopathology images.IEEETrans Biomed Eng57:841–852

Asp M, Giacomello S, Larsson L, Wu C, F€urth D, Qian X, W€ardell E, Custodio J, Reimegard J, Salmen Fet al(2019) A Spatiotemporal organ-wide gene expression and cell atlas of the developing human heart.Cell179:1647–1660.e19

Bach F, Jenatton R, Mairal J, Obozinski G (2011) Optimization with sparsity-inducing penalties.arXivhttps://arxiv.org/abs/1108.0775v2[PREPRINT] [csLG]

Beucher SLC (1979) Use of watersheds in contour detection.International Workshop on Image Processing: Real-time Edge and Motion Detection/estimation, Rennes, France

Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z (2020) Deep learning andalignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram.bioRxivhttps://doi.org/10.1101/2020.08.29.272831v1[PREPRINT]

Burgess DJ (2019) Spatial transcriptomics coming of age.Nat Rev Genet20:317

Chaudhuri D, Agrawal A (2010) Split-and-merge procedure for image segmentation using bimodality detection approach.Def Sci J60:290–301

Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X (2015) RNA imaging.Spatially resolved, highly multiplexed RNA profiling in single cells.Science348: aaa6090

Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI,Linnarsson S (2018) Spatial organization of the somatosensory cortex revealed by osmFISH.Nat Methods15:932–935

Correa-Tome FE, Sanchez-Yanez RE (2015) Integral split-and-merge methodology for real-time image segmentation.J Electron Imaging24:013007

Dries R, Zhu Q, Dong R, Eng C-HL, Li H, Liu K, Fu Y, Zhao T, Sarkar A, Bao Fet al(2021) Giotto: a toolbox for integrative analysis and visualization ofspatial expression data.Genome Biology22:78

Egozi A, Bahar Halpern K, Farack L, Rotem H, Itzkovitz S (2020) Zonation of pancreatic acinar cells in diabetic mice.Cell Rep32:108043

Eng C-H, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, KarpC, Yuan G-Cet al(2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH.Nature568:235–239

Gamarra M, Zurek E, Escalante HJ, Hurtado L, San-Juan-Vergara H (2019) Splitand merge watershed: a two-step method for cell segmentation in fluorescence microscopy images.Biomed Signal Process Control 53:101575

Gerdes MJ, Sevinsky CJ, Sood A, Adak S, Bello MO, Bordwell A, Can A, CorwinA, Dinn S, Filkins RJet al(2013) Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue.Proc Natl Acad Sci USA 110:11982–11987

Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. InProceedings of the Thirteenth InternationalConference on Artificial Intelligence and Statistics, Teh YW, Titterington M(eds), pp 249–256. Sardinia: PMLR

Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, BlackS, Nolan GP (2018) Deep profiling of mouse splenic architecture withCODEX multiplexed imaging.Cell174:968–981.e15

Goodfellow I, Bengio Y, Courville A, Bengio Y (2016)Deep learning. Cambridge,MA: MIT press

Halpern KB, Shenhav R, Matcovitch-Natan O, Toth B, Lemze D, Golan M,Massasa EE, Baydatch S, Landen S, Moor AEet al(2017) Single-cell spatial reconstruction reveals global division of labour in the mammalian liver.Nature542:352–356

HuBMAP Consortium (2019) The human body at cellular resolution: the NIHHuman Biomolecular Atlas Program.Nature574:187–192

Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift.arXivhttps://arxiv.org/abs/1502.03167[PREPRINT]

Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, Yang S-R, KurianA, Van Valen D, West Ret al(2018) A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging.Cell174:1373–1387.e19

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization.arXivhttps://arxiv.org/abs/1412.6980v9[PREPRINT]

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang Jl, Ferrante TC, Terry R,Jeanty SSF, Li C, Amamoto Ret al(2014) Highly multiplexed subcellularRNA sequencing in situ.Science343:1360–1363

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, TurczykBM, Yang JL, Lee HS, Aach Jet al(2015) Fluorescent in situ sequencing(FISSEQ) of RNA for gene expression profiling in intact cells and tissues.Nat Protoc10:442–458

Lee JH (2017) Quantitatcontext of gene expression.Wiley Interdiscip Rev SystBiol Med9:e1369Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF,Boguski MS, Brockway KS, Byrnes EJet al(2007) Genome-wide atlas of gene expression in the adult mouse brain.Nature445:168–176

Lin J-R, Fallahi-Sichani M, Sorger PK (2015) Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method.Nat Commun6:8390

Lin J-R, Izar B, Wang S, Yapp C, Mei S, Shah PM, Santagata S, Sorger PK (2018)Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes.Elife7:e31657

Lopez R, Nazaret A, Langevin M, Samaran J, Regier J, Jordan MI, Yosef N(2019) A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements.arXivhttps://arxiv.org/abs/1905.02269[PREPRINT]

Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L (2014) Single-cell in situRNA profiling by sequential hybridization.Nat Methods11:360–361

Lundberg E, Borner GHH (2019) Spatial proteomics: a powerful discovery tool for cell biology.Nat Rev Mol Cell Biol20:285–302

Moffitt JR, Zhuang X (2016) RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH).Methods Enzymol572:1–49

Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac Cet al(2018) Molecular, spatial, andfunctional single-cell profiling of the hypothalamic preoptic region.Science362: eaau5324

Moor AE, Harnik Y, Ben-Moshe S, Massasa EE, Rozenberg M, Eilam R, BaharHalpern K, Itzkovitz S (2018) Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis.Cell175:1156–1167.e15

Mukamel EA, Ngai J (2019) Perspectives on defining cell types in the brain.Curr Opin Neurobiol56:61–68Najman L, Schmitt M (1994) Watershed of a continuous function.SignalProcess38:99–112

Park J, Choi W, Tiesmeyer S, Long B, Borm LE, Garren E, Nguyen TN,Codeluppi S, Schlesner M, Tasic Bet al(2019) Segmentation-free inference of cell types from in situ transcriptomics data.bioRxivhttps://doi.org/10.1101/800748[PREPRINT]

Petukhov V, Soldatov RA, Khodosevich K, Kharchenko PV (2020) Bayesian Segmentation of spatially resolved transcriptomics data.bioRxivhttps://doi.org/10.1101/2020.10.05.326777[PREPRINT]

Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Mu~noz-Manchado AB,Skene N, Hjerling-Leffler J, Nilsson M (2020) Probabilistic cell type enables fine mapping of closely related cell types in situ.Nat Methods17:101–106

Salmen F, Stahl PL, Mollbrink A, Navarro JF, Vickovic S, Frisen J, Lundeberg J(2018) Barcoded solid-phase RNA capture for spatial transcriptomics profiling in mammalian tissue sections.Nat Protoc13:2501–2534

Stahl PL, Salmen F, Vickovic S, Lundmark A, Navarro JF, Magnusson J,Giacomello S, Asp M, Westholm JO, Huss Met al(2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.Science353:78–82

Stringer C, Wang T, Michaelos M, Pachitariu M (2020) Cellpose: a generalist algorithm for cellular segmentation.bioRxivhttps://doi.org/10.1101/2020.02.02.931238[PREPRINT]

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck3rd WM, HaoY, Stoeckius M, Smibert P, Satija R (2019) Comprehensive Integration of single-cell data.Cell177:1888–1902.e21

Surut Y, Phukpattaranont P (2010) Overlapping cell image segmentation using surface splitting and surface merging algorithms.InSecond APSIPA A nnual Summit and Conference, Singapore,pp662–666

Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J (2020)A benchmark of batch-effect correction methods for single-cell RNAsequencing data.Genome Biol21:12

Trapnell C (2015) Defining cell types and states with single-cell genomics.Genome Res25:1491–1498

Turczyk BM, Busby M, Martin AL, Daugharthy ER, Myung D, Terry RC, InversoSA, Kohman RE, Church GM (2020) Spatial sequencing: a perspective.J Biomol Tech31:44

Vu QD, Graham S, Kurc T, To MNN, Shaban M, Qaiser T, Koohbanani NA,Khurram SA, Kalpathy-Cramer J, Zhao Tet al(2019) Methods for segmentation and classification of digital microscopy tissue images.FrontBioeng Biotechnol7:53

Wasserman L (2006)All of nonparametric statistics. Secaucus, NJ: SpringerScience & Business Media

Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ (2019)Single-cell multi-omic integration compares and contrasts features of brain cell identity.Cell177:1873–1887.e17

Yuste R, Hawrylycz M, Aalling N, Aguilar-Valles A, Arendt D, Arnedillo RA,Ascoli GA, Bielza C, Bokharaie V, Bergmann TBet al(2020) A community-based transcriptomics classification and nomenclature of neocortical celltypes.Nat Neurosci23:1456–146

Chapter 3

Spatial Tissue Perturbation Profiling At The Molecular Level Using Multiplexed Error Robust Fluorescence In Situ Hybridization

Hemminger, Zachary; Tam, Gabriella; Aamodt, Caitlin; White, Stephanie; Wollman, Roy

Abstract

Image-based spatial single cell technologies have developed to a point where the molecular profiles of tens of thousands of single cells can be measured spatially with subcellular precision. The vast majority of the published work has been focused on technological development and demonstration. Few if any works exist where image-based spatial transcriptomics is used to compare biological conditions. Despite this, the high-quality data generated from these technologies is ideal for detailed molecular and spatial comparison across biological conditions. Here we apply an image-based spatial transcriptomic technique Multiplexed Error Robust Fluorescence In Situ Hybridization or MERFISH to three increasingly difficult biological systems to demonstrate the comparative utility of these technologies. We first perform MERFISH on cell culture to show TNF-a induced differential gene expression. We then investigate inflammation response in the homogeneous epithelium of the mouse cornea during wound healing. Lasty, we perform MERFISH on Zebra Finch brains to interrogate the role of mir128 on Area X and its role in vocal learning. Together we show that MERFISH is ideal for comparative studies but that important consideration is needed to minimize sample to sample bias.

Introduction

Understanding a system's function often starts with determining the structure of the system, perturbing that system, and investigating how that system responds. Investigation into

the function (physiology) of biological systems like tissues and organs has always been connected to their structure (anatomy). Advances in single-cell transcriptomics have allowed measurements of hundreds of thousands to millions of cells in single datasets (Svensson et al 2020). Despite deep profiling of the building blocks of biology, the lack of spatial resolution has limited their usefulness in advancing our understanding of anatomy and physiology. Historically histology has dominated the structural profiling of tissues at the cellular and molecular levels. Recent advances in our ability to profile biological tissues have allowed us to quantify detailed anatomy at levels that were previously impossible (Moses & Patcher 2022). Detailed molecular maps of tissues provide an immediate and transformative impact on our interpretation of how those tissues function.

Spatial profiling technology exists at the proteomic as well as the transcriptomic level (Nagle et al 2021). The proteomic level relies heavily on antibodies which need to be developed as well as be highly specific. This severely limits the number of proteomic measurements that can be performed on a single sample. The transcriptomic level also contains a division between sequencing and imaging-based approaches. Sequencing approaches label individual RNAs with a molecular barcode that encodes for their spatial positioning. Low capture efficiencies as well as the limited spatial resolution limit the biological processes that can be learned from this data. Image-based approaches rely on labeling targeted individual RNAs in situ.

Multiplexed Error Robust Fluorescence In Situ Hybridization or MERFISH is a gold standard for image-based spatial transcriptomics (Chen et al 2015). This approach works by designing DNA probes that tag RNA with a combination of readout sequences that is unique to the specific gene of interest. A fluorescent readout probe is then used to read out each sequence and the specific location of all of the transcripts that have been tagged with that sequence. This is accomplished with iterative rounds of imaging and results in diffraction limited spots that represent transcripts that were assigned that molecular readout. This allows sub cellular

resolution and high capture efficiency but the technology is limited predominantly to hundreds of genes at a time and relatively small areas of interest. Despite the limitations of scale present in image-based approaches, the high-quality measurement allows in-depth profiling of tissues with accuracies that far exceed other approaches.

Although a substantial amount of work has been done to develop as well as commercialize MERFISH, most work has been technical and focuses primarily on method development (Xia et al 2019, Moffit & Zhuang 2016, Wang et al 2018, Xia et al 2019, Littman et al 2021) .The biological applications of the work have overwhelmingly been on profiling the spatial composition of wild type tissues (Zhang et al 2021,Fang et al 2022). While a few exceptions exist, they are limited to single condition experiments (Foreman & Wollman 2020, Maltz & Wollman 2022) . Work including multiple conditions or experimental perturbations are essentially non-existent despite the benefit that they may provide.

Understanding the spatial and temporal organization of various genetic and non genetic perturbations as well as complex biological processes requires the collection of multiple samples. Image based spatial transcriptomics while powerful have issues with throughput, signal to noise and robustness. Sensitivity to RNase contamination, fluidics, and imaging failures decrease the likelihood that samples successfully complete the long experimental protocol to completion with sufficient signal to be processed. The complex biological makeup of tissues can lead to high levels of background requiring improved clearing protocols. Even when signals are sufficient and background is low robust fluidics protocols are needed to ensure reproducibility across multiple samples. Image based transcriptomics relies on high magnification imaging of single molecules which has a small field of view and requires high exposure times to get usable signal to noise ratios. Successful acquisition can generate terabytes of image data that need to be processed in a reasonable timescale to iterate experimental conditions. Together these features make image based spatial transcriptomics a

difficult method to implement and limits the total number of samples that can be generated (Moses & Patcher 2022).

Given the lack of datasets, it is not clear how technical and biological variation across samples will impact MERFISH. If this impact is large enough, even robust biological phenotypes can be difficult to observe. It is important to understand the limitations of any technology as well as designing ways to overcome these limitations. To better understand the limitations of MERFISH, we performed MERFISH on three different systems of ranging complexity. With the simplest system of cell culture we show that MERFISH can be used to quantify differences in inflammation response to various concentrations of TNF-a. In the homogeneous transparent epithelium of the mouse corneas, we show that MERFISH can detect inflammation response in epithelial cells in response to wounds as a function of distance from the wound and time since the wound. In a much more complex system of Zebra Finch brains, we identify cell type composition differences and differential expression involved in vocal learning and the microRNAs that are involved. Here we show that even through technical limitations, MERFISH can be used to investigate biological processes across samples.

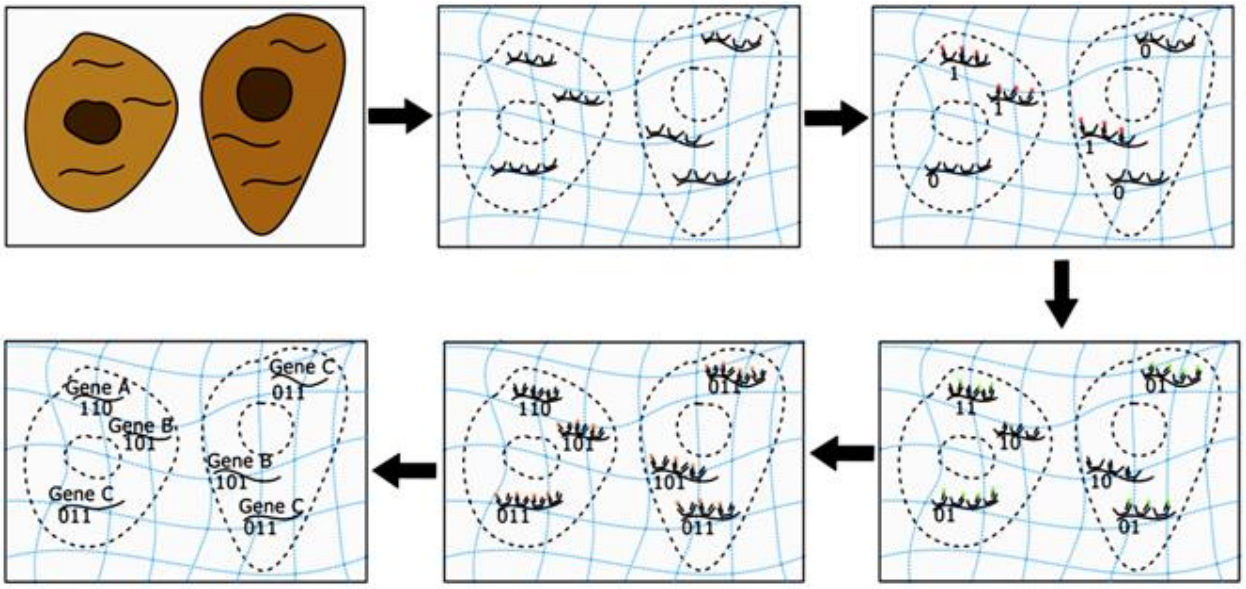

Figure 3.1: MERFISH Methodology
MERFISH is performed by hybridizing DNA encoding probes to targeted RNAs, embedding the sample into a hydrogel and clearing away proteins and lipids. Through iterative rounds, readout probes are hybridized to these encoding probes. The presence or absence of signal in each round encodes the identity of the transcript of interest. These observed barcodes are compared to the barcodes that were encoded into the encoding probes to annotate detected transcripts with their identities.

Reproducible biological measurements across samples are essential to profile the spatial gene expression changes that occur during a perturbation. The experimental, instrumental and computational infrastructure necessary for this is nontrivial and requires expertise in a broad range of technical skills from molecular biology and optical engineering to data science. This is likely why there are few examples of spatial transcriptomic measurements of perturbations in literature.

MERFISH Experimental Outline

The experimental procedure for MERFISH is a multi day protocol with the goal to localize and visualize single mRNA molecules with sufficient signal to noise ratio across multiple rounds of imaging. MERFISH experimental design typically begins with the determination of

which species, tissue type and perturbations or lack thereof will be used. Previously collected gene expression data for the system is then used to determine which genes should be measured. This reference gene expression data is pivotal to ensure effective use of the relatively few gene targets that MERFISH can perform. Genes that are not expressed are poor candidates for target while genes with high expression can also be poor candidates due to the optical crowding limitations of the technology. On top of the gene expression levels, genes are not all equally informative. Their usefulness is highly dependent on the biological system and what biological processes the experiment is hoping to uncover. Even highly informative genes with appropriate expression levels may not be ideal especially if the gene is short and few probes could be designed. Significant work has been done on the design of these probes but the decisions of which genes to target are often performed manually using literature to guide the process. This process is significantly easier in homogeneous samples where the gene expression of multiple cell types is not needed to be considered.

Once these probes are designed and ordered they need to be amplified to sufficient purity and concentration for staining. This is predominantly done by using PCR to add a T7 promoter sequence to the probes, T7 polymerase generates many ssRNA copies which are converted back to DNA with a reverse transcriptase (Moffit & Zhuang 2016). Given the short nature of these probes, standard protocols for PCR, IVT and Reverse Transcription give poor yield and increased template concentrations often improve yields. Cleaning up intermediates with a phenol chloroform extraction and dialysis columns also increase the yield of these reactions. Clean concentrated probes are essential for reproducibility across samples and can lead to differences in hybridization efficiency and non-specific binding.

A key aspect of MERFISH and other image based spatial transcriptomic techniques is hydrogel embedding. In order to ensure that the hydrogel sample remains adherent to the coverslip across many protocol steps these coverslips need to be functionalized. Coverslips

after being cleaned in a mixture of concentrated hydrochloric acid and methanol are silanized in chloroform to add allyl groups to the surface of the glass. These allyl groups are then incorporated into the hydrogel to provide a covalent attachment to the glass surface. Other functionalizations can also be applied such as amination to improve tissue adherence prior to hydrogel embedding. These coverslip functionalizations are often made within a few weeks of use and stored in a dessicator. Ineffective functionalization can lead to tissue loss and morphology changes across the experimental protocol which may be inconsistent across datasets.

Preservation of RNA during sample collection and sample preparation is essential for reproducible MERFISH. Many sample preparation protocols include a flash freeze in liquid nitrogen to stop any enzymatic activity including RNases. Other protocols include prefixing tissue to prevent biological changes during handling. Fixation, paraffin embedding and other related tissue preservation protocols can dramatically impact MERFISH results and often require additional protocol development to overcome the complications added by these approaches. RNase inhibitors are also in common use for experimental steps. While protein based RNAse inhibitors are expensive and often limited to the hybridization solutions, chemical based RNAse inhibitors are cheap enough to be used in all solutions to minimize the effects of RNAses on gene expression. Failure to account for differences in tissue preparation can lead to changes in the amount of RNA that is lost, the hybridization efficiency, and the clearing across samples.

Without the use of optical sectioning MERFISH and most other image based spatial transcriptomics techniques are limited to relatively thin specimens. For cell culture this is not an issue but for tissues this requires sectioning to about the width of a single cell. Placing the thin section flat on the coverslip without wrinkles is also an unexpected difficulty of working with these thin sections. Given a complex tissue such as the mouse brain, accurate sectioning is

essential to ensure comparability across biological samples. The gene expression pattern in one ~10 um section may vary from another ~10 um section that is 10's to 100's of micrometers further into the tissue. Comparing gene expression patterns across datasets requires accurate sectioning and biological replicates. Without these the changes across samples may be due to sectioning accuracy and not the difference between samples.

After coverslip mounting the tissue is fixed to ensure that the RNA remains in the sample and to eliminate RNAses. While various fixation protocols exist, crosslinking protocols including paraformaldehyde are common. A caveat of these protocols is that they are sensitive to temperature and time. Variations in these factors can lead to some samples receiving more or less fixation. Under Fixation can lead to RNA and morphology loss. Overfixation can lead to autofluorescence and can reduce the diffusion of probes into the sample as well as decrease clearing efficiency. After fixation samples are often stored for later use. A common approach is to place the samples in 70% ethanol at -20C. This can also permeabilize the sample to allow probe penetration. Samples may be stored for shorter or longer times resulting in differences in permeabilization. One approach to account for this is to perform a secondary permeabilization with a detergent to ensure even permeabilization across samples. Differences in permeabilization can likely lead to differences in detected gene expression.

In order to perform hydrogel embedding and clearing, RNA must be anchored to the hydrogel. Without this RNA would be cleared away with lipids and proteins. Multiple approaches exist for this step but they can be simplified to covalent or non covalent. Non-covalent approaches include using a poly T DNA probe with an acridite modification which will be incorporated into the gel Moffit & Zhuang 2016). By hybridizing this to the mRNA within the sample the RNA cannot diffuse out of the sample as long as the probe is hybridized. Covalent methods often rely on modifying the RNA to add an acridite or allyl functionalization. The most common uses of this approach are Label-X and Melpha_X (Eng et al 2019, Wang et al 2021).

Covalent anchoring of RNA to hydrogels is ideal as it can allow for the preservation of non messenger RNAs and is robust to changes in temperature and salinity which do vary across the experimental protocol. Robust capture of RNAs is essential for consistent MERFISH across samples.

Hybridization of probes to the RNA is also essential for sufficient signal and reproducibility across datasets. Probes are typically hybridized at 2-5 nm per probe. While protocols exist for faster hybridization using higher concentrations, this may be cost prohibitive for MERFISH with 10's of thousands of probes. Hybridization conditions often include temperatures ranging from room temperature to 37C with formamide concentrations ranging from 10 to 50%. The addition of salts to stabilize charge interactions and dextran sulfate to satisfy hydrogen bonding interaction are also often added. Given the viscous nature of the hybridization solution, accurate consistency across samples can be difficult. Probes are typically hybridized anywhere from 12 to 36 hours although exceptions exist. Variations to these parameters between samples can lead to changes in hybridization efficiency as well as changes in non-specific binding.

Post hybridization, unbound and nonspecifically bound probes are removed through a series of washes. Often these steps occur at higher temperatures approaching the melting temperature for the probes. While this is optimal for ensuring specificity, the higher temperatures can cause correctly bound probes to fall off leading to a decrease in signal. Given the proximity to melting temperature minor variations in time or temperature can lead to vastly different signal across samples. By washing at the temperature that the probes were hybridized at this can be minimized leading to more consistent staining across samples.

Samples are then embedded within a hydrogel; often polyacrylamide or similar analogs. Efficient penetration of monomers within the sample is required for an even capturing of the

RNA within the sample. In addition to catalysts, these reactions are often temperature and oxygen sensitive. Changes in any of these are likely to lead to differences in hydrogel composition and RNA capture efficiency.

The purpose of the hydrogel embedding is to allow the removal of lipids and proteins without significant loss of RNA. This leads to improved optical clarity and can remove probes that were nonspecifically bound to lipids and proteins. Together the goal is to maximize signal light collection and decrease background. Detergents and proteases are often used together to accomplish this. In many of these digestion buffers EDTA is added to eliminate the activity of DNAses. Given that most DNAses will be inactivated with the paraformaldehyde, the benefit of the EDTA is limited compared to the decreased activity of proteinase k. Rather than using EDTA, CaCl2 can be added to increase the proteinase activity and clear proteins more efficiently [CITE FROM GABY]. Robust and consistent clearing is necessary to reproducibility and quantitatively perform MERFISH across multiple samples.

The hybridization of a fluorescent probe to the initial probe is necessary for measurement and localization of the RNA molecules. Between rounds of imaging the fluorophores are stripped off using a reducing agent which cleaves the disulfide between the DNA and fluorophore. These steps are predominantly done with an automated fluidics system and a closed chamber around the sample. These fluidics systems are sensitive to clogs as well as bubbles that can reduce the laminar flow of the system and lead to inconsistent hybridization and removal of fluorophores.

During Imaging the sample is often put into an imaging buffer which reduces the photobleaching of the fluorophore. These buffers typically rely on an enzyme to catalyze the sequestration of molecular oxygen from the solution. These enzymes can go bad over time and the substrates can run out leading to more or less photobleaching depending on when the

imaging buffer was created. This adds another inconsistency that can lead to difficulties generating reproducible gene expression measurements across samples.

Individual RNA molecules will at most have about 100 fluorophores bound to it at any time. Given that the hybridization efficiency is much lower than perfect, this number is likely smaller. In order to have sufficient signal over background, a bright excitation light is necessary and longer exposure times may be required. It can be expensive to optimize a system for multiple excitation wavelengths. One option is to focus on maximizing the excitation of a single fluorophore. While using fewer fluorophores means more rounds of hybridization, the optimized signal can lead to more consistent signal across rounds of imaging.

High resolution imaging is necessary to identify and localize individual RNA molecules in situ with voxel sizes of about 100 nm x 100 nm x 400 nm. This leads to small field of views and many zindexes that need to be imaged in order to capture all of the RNA within the cells of a sample. The combination of this and the time it takes to perform hybridization on the microscope with a fluidics system makes acquisition of MERFISH data a time consuming step. Imaging over long time periods like days to weeks introduces its own set of challenges to reproducibility. Throughput for MERFISH experiments and collecting enough biological replicates to statistically quantify a difference across samples is difficult to say the least.

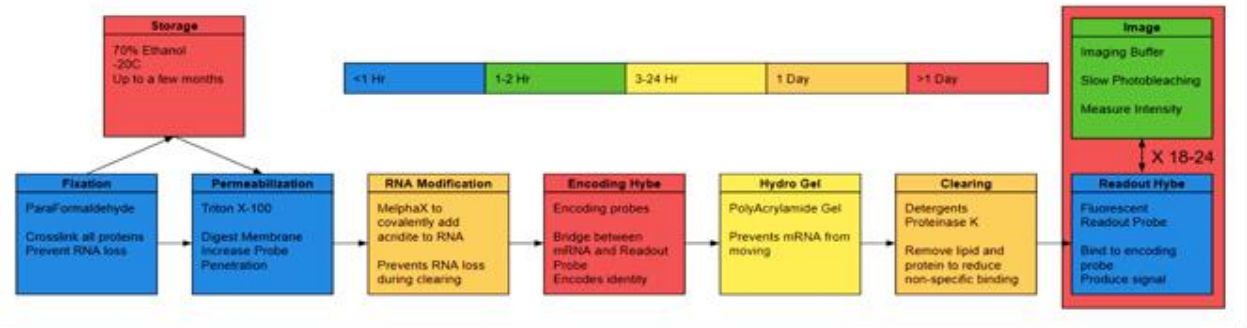

Figure 3.2: MERFISH Experimental Workflow
Samples undergo various experimental procedures during the MERFISH protocol. These procedures each take hours to days and each can affect the resulting MERFISH data quality.

MERFISH Computational Outline

MERFISH experiments can generate Terabytes of data per sample. In short images must be processed so that the signal from the RNA is preserved while competing signals are minimized. Images must be registered to ~100 nm accuracy to allow the pairing of the same RNA spot across multiple rounds. Individually measured spots must be correctly paired to localize and identify transcripts. These transcripts must then be assigned to their respective cell. Computationally processing and analyzing this data in a scalable and reproducible manner is its own challenge. Completing this in a way that the detected RNA transcripts are comparable across samples adds significant complexity to the problem.

There are multiple image artifacts that should be corrected to improve the signal to noise ratio of the RNA spots prior to detection. The first artifact is hot/dead pixels, these are pixels that don't change with the change in photons within the sample. They can be detected by looking at how different pixels are from the median of their neighbors. Given that pixels are ~100 nm apart they should have similar measurements. Pixels that are consistently different can be replaced with the median of their neighbors.

Another artifact is that across the field of view fluorophores can be excited and their emissions can be collected unevenly. This is a multiplicative effect that can be measured and then applied to each image. Correction ensures that RNA spots with the same number of fluorophores have the same signal no matter where they are located in the field of view.

Light travels differently through optics depending on the wavelength. In order to use multiple fluorophores for MERFISH the divergence should be corrected. This is done by using multicolor beads. The centers of these beads can be measured across the different emissions bands and the chromatic aberration can be measured. Individual images can then be interpolated so that all wavelengths have the same chromatic aberration.

Many photons detected in the image will not be coming from the MERFISH signal. This background signal can be uneven across the image but is often of lower frequency than the MERFISH RNA spots. One approach to correcting this is to apply a gaussian blur with a sigma larger than the RNA spots across the image. This calculated background can then be subtracted from the image leaving just the high frequency signal. Not all of the high frequency signal is RNA spots as some can be due to poisson noise in the image. This high frequency noise can also be smoothed out with a gaussian blur with a sigma between the size of an RNA spot and the high frequency noise. These steps are essential to ensure that sample to sample background is removed and that the RNA are detected accurately across samples.

Light does not travel directly between the fluorophore and the camera pixel. Given that it is a wave there is a pattern of constructive and destructive interferences that generate what is known as a point spread function. This can lead to photons from a fluorophore appearing in the wrong place within the image. One method to correct this is deconvolution. By computationally returning the signal back to where it is most likely to have originated from the image can be made sharper and the blur induced by the interference patterns can be removed. This is most useful for out of focus light in the z axis where the point spread function is widest. Performing this step after background subtraction has minimal benefits since the out of focus light typically is of lower frequency and is removed.

Images across multiple rounds may not necessarily be aligned perfectly. Inaccuracies in the stage as well as shifts induced by the fluidics can lead to images across different rounds not being aligned. This alignment can be measured and corrected with the use of fiduciary markers. These are fluorescent markers that were added within the hydrogel and do not move independently of the sample. By imaging these each round the distance that the sample has moved in xy and z can be calculated. This registration is done by localizing and pairing beads across different rounds of imaging and then measuring the differences in their localizations.

These are often rigid transformations which can be applied to the processed images to ensure that the same RNA molecule imaged across multiple rounds of imaging are located at the same pixel regardless of which round they had signal in. Inaccuracies in this step can lead to misidentification of RNA transcripts and lower detection efficiency which may not be consistent across all genes and especially across fields of views and samples.

While image processing generates images that by eye are visually cleaner the purpose is actually to allow quantitative measurement of the remaining RNA signal. Given that the goal of MERFISH is to identify and localize individual RNA molecules, the RNA's present in the images need to be detected and then matched to the genes that were targeted by the probes. Multiple methods exist to convert images to gene expression each with their own caveats and sensitivities. Accurate and reproducible detection and quantification is ideal in order to compare gene expression patterns across samples.

A pixel based approach exists where a vector is generated for each pixel across the different rounds of imaging. This vector is then normalized so that rounds in which an RNA had signal will be bright and the rounds without signal will be dim. By comparing this vector to each of the designed MERFISH barcodes the gene can be decoded. The localization of this gene will be the average of the neighboring pixels which are also classified as the same gene. This approach works well for optically crowded samples as you only need a few pixels to accurately decode but is sensitive to variations across rounds of imaging, especially registration errors. Given that the vector is normalized prior to decoding, this approach is also sensitive to high frequency noise.

A spot based approach also exists where spots are detected in each image either using a feature matching approach or looking for peaks of maximal intensity. The localization of these spots is determined by fitting a gaussian to their intensity. These spots are paired with spots in

the same xyz location within a defined radius across different rounds of imaging. This approach is less sensitive to high frequency noise and round to round variations in intensity as well as registration inaccuracies. Optical crowding does complicate this approach as nearby spots from other transcripts within the defined radius could be inaccurately paired leading to inaccurate gene assignment and lower detection efficiency.

Detected and identified transcripts must be accurately assigned to the correct cell. This is done by image segmentation. Nuclear stains and occasionally cytoplasmic stains are acquired during the imaging to allow quantification as to which pixels belong to each cell. Even with these stains it is possible to misassign RNAs to the wrong cell, especially in densely populated tissues. Computational approaches can be applied to decide which cell the gene expression pattern within image voxels correlate more with. This can be done purely based on the gene expression of the nucleus or by using reference single cell RNA sequencing data to generate expected gene expression patterns for cell types. Inaccurate RNA transcript assignment can lead to inaccurate cell type identification and complicate comparative analysis across samples.

Despite the best efforts to control experimental bias, some will exist. On top of that experimental bias will be biological variation that is independent of the experimental conditions. Together these can affect the gene expression detection efficiency as well as the false positive rate. Systematic changes across datasets can be corrected with batch correction. One common approach is Harmony which projects the data from multiple samples into their principal components and then iteratively corrects the batch effects until the datasets align. This correction can either be correcting one dataset to another or finding a middle ground between multiple datasets. While technical bias and unwanted biological bias can be corrected this way there may be unexpected artifacts that are created and experimental gene expression differences may be artificially minimized. Generating high quality data is essential to minimize

the amount of batch correction that is needed to ensure meaningful results are not lost across samples.

Experimentally there are multiple steps in which minor variations to protocol steps can lead to differences in measured gene expression. Minimizing these differences and collecting sufficient data to ensure that the measured differences are non trivial. Developing the infrastructure to automate data collection and image processing in a robust way also requires significant technical knowledge. It is not surprising that few spatial gene expression datasets exist that compare across biological conditions in a statistically quantitative manner.

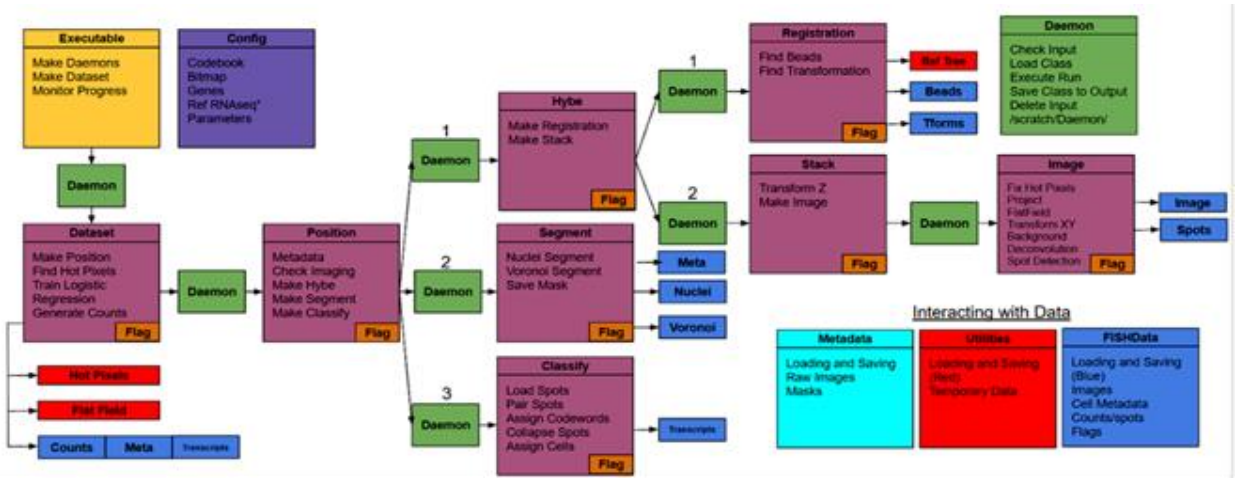

Figure 3.3: MERFISH Computational Workflow
Collected data must be processed to detect individual molecules across the rounds of imaging as well as decode the identities of these molecules and assign these decoded transcripts to cells. This processing consists of multiple image processing steps and inconsistencies in processing can limit the quality of MERFISH data.

Mouse 3T3 Cells TNF Stimulation

One of the simplest samples that can be performed with MERFISH is cell culture. Relatively flat cells can be grown directly on coverslips with minimal autofluorescence. Their flatness ensures that there is minimal sample above or below the plane of focus resulting in minimal out of focus light. The gene expression profiles of these cells are often well understood so MERFISH gene target decisions are easier to make. Segmentation of cells in cell culture is also trivial as they are rarely dense and well separated from the background. Cell culture allows

you to place multiple experimental conditions on the same coverslip. This significantly reduced the amount of technical variation between samples.

Here we show mouse 3T3 cells that were stimulated with varying amounts of TNF-a. MERFISH was performed on these cells with an inflammation gene probe set. With low background signal levels and high spot intensity over 600 transcripts were detected per cell on average with a pearson correlation to reference RNAseq data of 0.72 (Wang et al 2022) . Inflammation genes were shown to vary with TNF concentrations in a dose dependent manner for some but not all genes.

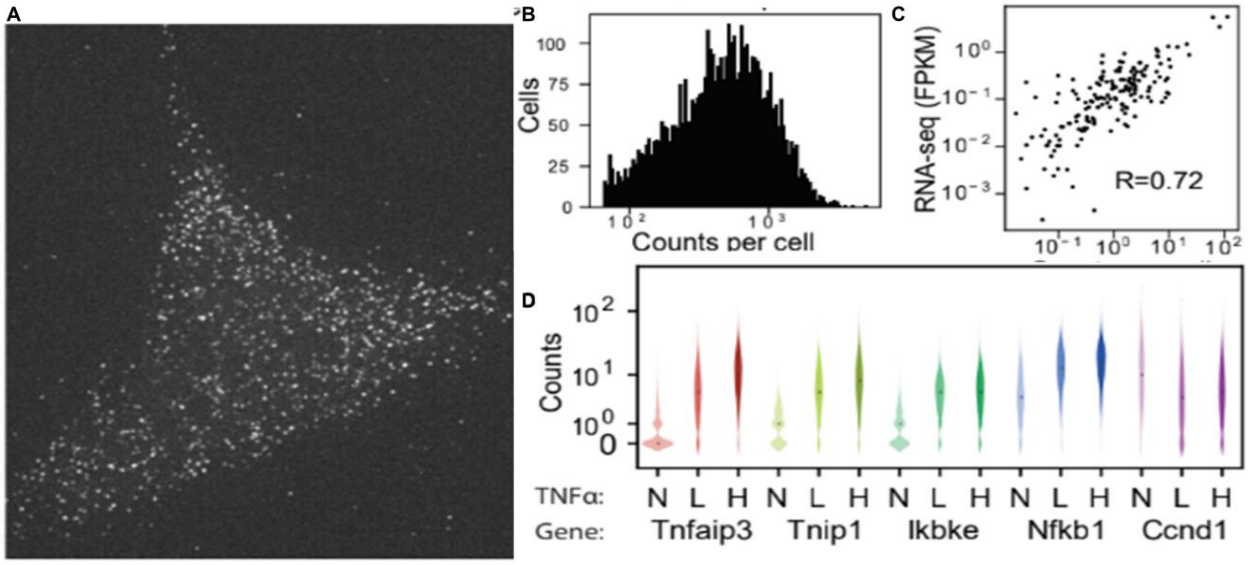

Figure 3.4: Cell Culture MERFISH.
A. Processed Images of a single mouse 3T3 cell and the diffraction limited spots that are observed for each transcript that has been tagged with encoding probes for this specific round of imaging. B. Histogram of the Number of transcripts detected per cell average of ~600 transcripts per cell. C . Correlation of 0.72 between bulk RNAseq and MERFISH. D . Violin plots of expression distribution for cells exposed to 0, 1 or 10 ng/ uL of TNF-A for 3 hours showing a dose-dependent increase in gene expression for some genes but not all.

Mouse Cornea: Wound Healing

Cell culture is significantly different from tissue in the context of MERFISH. Tissues contain orders of magnitude more extracellular matrix which can limit probe diffusion and lead to light scatter. One middle ground is the use of the mouse cornea. The mouse cornea is composed of an epithelial sheet on top of a thick sparsely populated collagen rich stroma and

then a thin layer of endothelium. The epithelial layers are relatively homogeneous making gene selection simple. The optical transparent nature of the cornea means minimal light scatter despite the extracellular matrix. Segmentation within a dense epithelium is notoriously difficult as all cells are in contact with each other but primarily the same shape and size.

Corneas have biological interest as they are notoriously good at healing (Stepp et al 2014). After an epithelial debridement the epithelium slides in to fill the wound while immune cells migrate through the stroma (Oyler-Yaniv et al 2021). This process is highly reproducible and scars rarely form. While the global cell type migrations of the cornea wound healing process were known the gene expression patterns present in the epithelial layers were not. It was clear that some cells migrate or apoptose while others proliferate and possibly differentiate.

Using a MERFISH gene panel consisting of Inflammation markers the gene expression profiles of the whole mount epithelium were measured 2 and 15 hours after epithelial debridement and compared to an unwounded control. The dense epithelium limited our ability to segment due to too much out of focus light from cells above and below each other. This limited our ability to look at single cell gene expression. Despite this we could look for overall patterns of gene expression. Flatness issues of the whole mount also limited our ability to quantify differences across timepoints. Minimal evidence suggested that cells proximal to the wound may have upregulated expression of the inflammation markers as well as epithelial cells after they migrate into the wound bed although without biological replicates we were unable to statistically prove this. Given that the majority of cornea wound healing is performed in the stroma and that inflammation in the epithelium can lead to scarring it is not surprising that we didn't see much gene expression changes beyond the wound bed (Stepp et al 2014).

Our work in corneas highlights the importance of biological replicates as well as tissue flatness in performing MERFISH. This also highlights the differences between tissues and cell

culture. Even in a homogenous transparent epithelium technical artifacts limit our ability to quantify differences across experimental conditions. In addition to artifacts, the MERFISH datasets generated took significant time to generate which limited the number of biological replicates that could reasonably be expected.

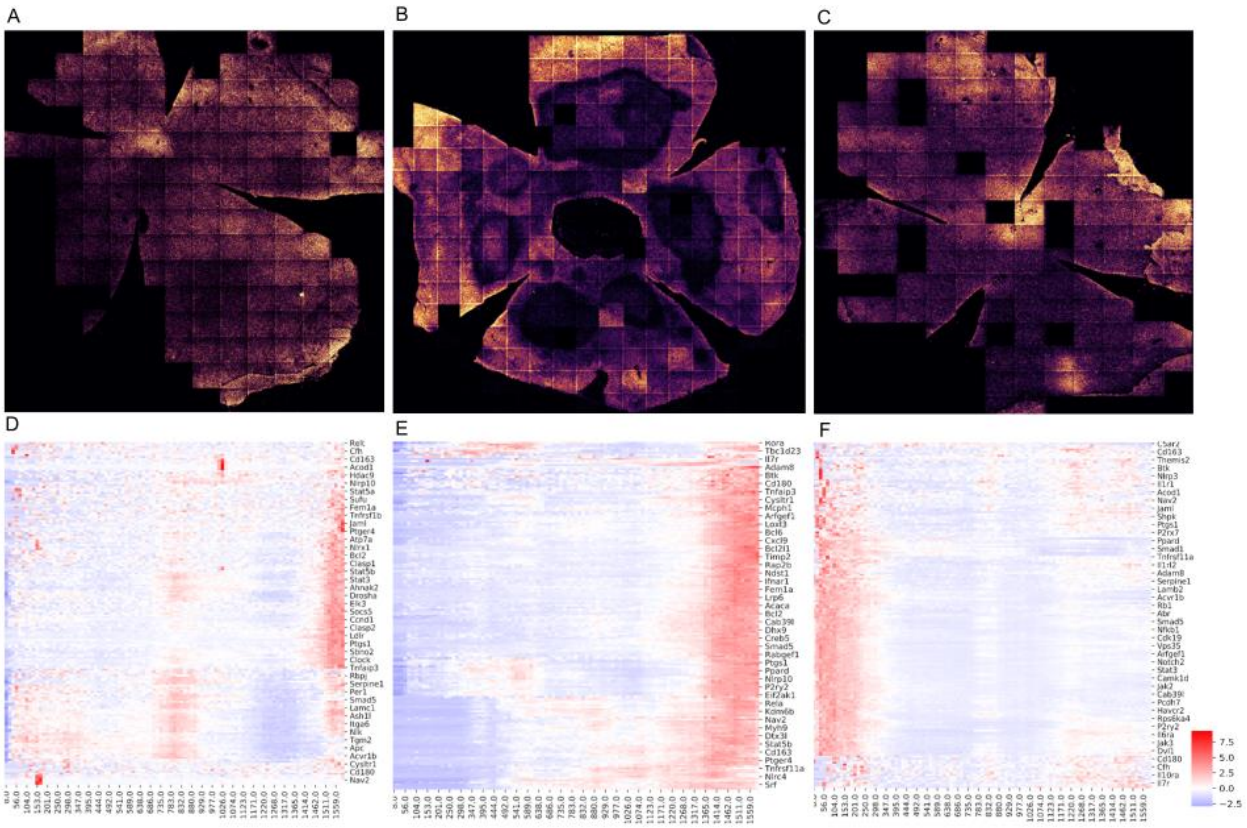

Figure 3.5: Mouse Cornea MERFISH
A,B,C. 2D Histograms of detected transcripts for (A: Unwounded, B: 2 Hours post Wound, C: 15 Hours post wound). D,E,F. Heatmaps for each gene binned into ~50 um bins as a function of distance from the center of the cornea or wound (Left is Center, Right is Edge). Colormap is Zscore for individual genes from -2.5 (Blue) to 7.5 (Red) standard deviations from mean.

Zebra Finch Vocal Learning Regulation

Moving from a relatively homogenous tissue like the mouse corneal epithelium to a more complex tissue like the Zebra Finch Brain possess greater challenges to MERFISH reproducibility. In a homogeneous tissue bulk RNAseq can be used to decide gene targets. When multiple cell types are present scRNAseq or extensive literature search is necessary to

ensure that genes targets are expressed in at least some of the cell types but not overexpressed in any of the cell types. In an ideal MERFISH experiment there will be many but not too many spots present in each cell during each round of imaging. Since higher expressed genes lead to optical crowding, MERFISH gene target decisions must be made with highly noisy genes. This can lead to non ideal staining where some cell types receive more spots in a single round while others receive less. The closer you get to optical crowding the higher false positive rate and the lower detection efficiency. This applies not only to the highly expressed genes but also to the other genes that share bits with them. This bias will not be consistent across cell types. This also proposes an issue for reproducibility across biological conditions. If a gene is upregulated compared to control it will increase the optical crowding leading to bias in those bits compared to the control. While capturing as many transcripts as possible is enticing, designing probe sets for fewer transcripts per cell can lead to more reproducible decoding across biological conditions.

Zebra finches are a model organism of vocal learning which correlates with patterns of human vocal learning. A specific region within the zebra finch brain Area X has been shown to be heavily involved in this vocal learning which mirrors regions of the human brain. A microRNA mir128 known for regulating genes associated with vocal learning in zebra finch and analogs have also shown to be dysregulated in certain autism spectrum disorder data (Aamodt & White 2022). Knock down experiments were performed in order to understand how dysregulation of this microRNA affects the structure and gene expression patterns within the zebra finch Area X and MERFISH was performed on these samples.

The zebra finch transcriptome is less well studied compared to mammalian transcriptomes. To compensate for this an extensive literature search was performed to identify cell type marker genes as well as genes known to be regulated by mir128.

Given the small size of the Zebra Finch Brains, different experimental samples could be placed on the same coverslip to minimize batch effects between conditions and allow us to elucidate the effect of this regulating microRNA on the circuitry involved in vocal learning. A marker for Area X was used to ensure that the areas measured were indeed Area X. Expression of this marker did not vary significantly between control and knock down. Expression of the mir128 targets were shown to be enriched in the knock down. Given that the knockdown was performed when the zebra finch was an adolescent and then the samples were collected when the zebra finches were adults, it wasn't guaranteed that the knockdown would still be active. Given that mir128 expression decreases the expression of these targets, the higher expression suggested that the knockdown was generating lasting effects.

Using markers for known cell types as well as neurogenesis we saw relatively consistent compositions of cell types between the knock down and control with the exception of astrocytes and neurons. Compositions of astrocytes seem to be a technical artifact present only in one hemisphere of the control brain. Cells expressing higher amounts of neuronal markers were enriched in the knock down. This matches well with literature that shows a connection between mir128 and neuronal apoptosis, migration and proliferation (Zhang 2016).

For the cell types that did not show composition differences, we looked at the expression levels of the mir128 targets. We noted that all cell types show similar upregulation of mir128 targets suggesting that the knockdown was affecting all cells rather than a specific cell type. We next looked at individual genes to see how the mir128 targets were upregulated in these cells. Rather than higher expression levels within single cells for each gene, it seems like there are more cells expressing these genes in general. Given that the overall expression of the mir128 targets was upregulated suggests that knocking down mir128 increases the number of target genes that are expressed, not the expression levels of each gene. This is true for some mir128 targets but not all.

These preliminary results support the role of mir128 in suppressing certain gene expression programs associated with vocal learning and that at least some of these programs are involved with recruiting neurons to area X within the zebra finch brain. Having both samples on the same coverslip ensured that minor variations in experimental and instrumental conditions were consistent between samples allowing for clearer interpretation of the results with no batch correction. Having multiple samples on a coverslip does significantly increase the acquisition time for the dataset. This significantly decreases the number of biological replicates that can be performed and increases the risk of experimental failure at any point during acquisition.

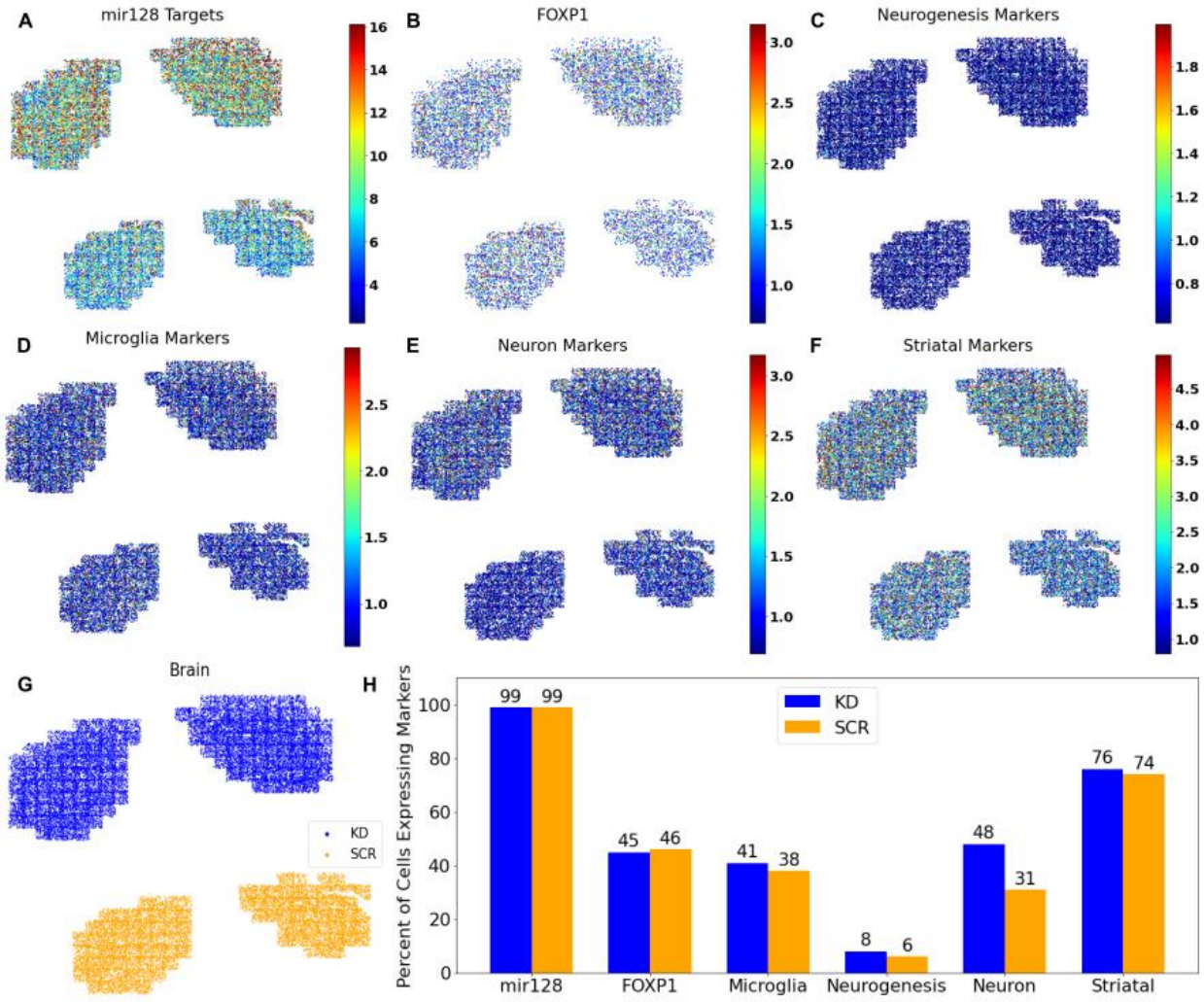

Figure 3.6: Zebra Finch Area X MERFISH
A. dots represent individual cells within Knock Down or Scramble Brain. Coordinates represent spatial position within Area X. B. Percent of Cells expressing Markers. C-H Spatial map of cells colored by the number of transcripts for each marker set present in each cell.

Discussion

While there is still a wealth of information to be gained from understanding the spatial gene expression of individual samples, the ability to compare those spatial gene expression profiles across different experimental conditions is essential for understanding non wild type biological processes. Here we have shown that performing MERFISH across biological conditions is non trivial but worth the effort.

Experimental variation can be minimized by developing robust protocols and placing multiple samples on the same coverslip but is not always feasible when imaging large samples. Failure to minimize the experimental variation can lead to differences in signal as well as noise which can contribute to inconsistent processing across samples. Designing and implementing efficient computational pipelines for processing terabytes of this data in ways that minimize technical artifacts is essential. Tradeoffs exist between maximizing the decoding efficiency of a single field of views versus generating more consistent decoding and assignment to cells.

Using these approaches, we have shown spatial gene expression differences across conditions in three increasingly difficult samples. In cell culture we show high performance MERFISH that aligns well across technologies and can identify dose dependent gene expression changes to TNF-a. In a wound healing model ,mouse cornea. We have shown upregulated inflammation response in epithelial cells as they migrate into the wound bed but were unable to generate the biological replicates needed to statistically quantify. In the vocal learning model zebra finch brains we show the role of mir128 in regulating the activity of gene expression programs some of which contribute to migration of neurons into area x and may lead to better understanding and potential treatments in autism spectrum disorder.

Generating and processing these datasets took years of experimental, instrumentation, computational and design development and required technical expertise in a number of areas. Given this it is not surprising that there are few examples of Image based spatial gene expression datasets that compare gene expression patterns across different experimental conditions. With the commercialization of MERFISH and other related technologies, the design of systems that can perform various steps of the experimental protocol is promising to help reduce the batch effects in each dataset. This should lead to more and more high quality comparative studies but that is yet to be seen. High spatial resolution single cell spatial gene expression data should provide unique insights into biological processes that currently cannot be elucidated through either low throughput or noisy sequencing approaches. Significant work is needed to approach the reproducibility needed to produce this data.

Materials And Methods

Encoding Probe Design and Synthesis

Encoding probes were designed using existing software to generate 30 base pair specific homology to RNA targets with a gc content of 45 to 65% and a melting temperature of 65 to 72C. Readout probe bonding sequences were concatenated to the encoding regions and amplification primers were designed and appended to both ends.

Coverslip functionalization

40mm round type 1.5 coverslips were cleaned in a 50:50 mixture of 37% concentrated HCl and Methanol for 30 minutes with sonication. Coverslips were rinsed with deionized water 3 times for 5 minutes each, once in Ethanol, and dried at 70C. Coverslips were modified with 0.2% allyltrichlorosiloxane in chloroform with 0.1% triethylamine for 30 minutes with sonication to facilitate hydrogel adhesion. Coverslips were rinsed once with chloroform, twice with ethanol and dried for 1 hour at 70C. In cases where additional sample adhesion is necessary,

Coverslips were modified with 2% aminopropyltriethoxysilane in acetone for 2 minutes. Coverslips were rinsed with deionized water twice, ethanol once, and dried at 70C.

Fixation

Samples were placed on functionalized 40mm round coverslips and fixed in cold 4% paraformaldehyde in 1xPBS for 5 minutes for cells and 15 minutes for tissues with agitation and washed three times in 1xPBS with 3 mg/mL poly vinylsulfonic acid (PVSA) and 0.1% triton x-100 for 5 minutes each with agitation. Samples were buffer exchanged into 70% ethanol and stored at -20C.

Permeabilization

Samples were rinsed with 1xPBS with PVSA and 0.1% triton x-100 three times for five minutes each with agitation. Samples were permeabilized with 1% triton in 1xPBS with PVSA for 30 minutes at 37 C with agitation. Samples were rinsed with 1xPBS with 0.1% triton x-100 and PVSA three times for five minutes each with agitation.

RNA Modification with MelphaX

Samples were rinsed with 30 mM MOPS ph 7.7 + 0.1% triton x-100 + 3mg/mL PVSA three times for five minutes each with agitation. To the same 50 uL of 1 mg/ml MelphaX in MOPS was added and a parafilm square was placed on top to prevent evaporation. Sample was reacted at 37 C overnight in a humidity chamber. Sample was washed with 1xPBS + 0.1% triton x-100 + PVSA three times for 5 minutes each with agitation.

Encoding Hybridization

Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation. Sample was rinsed at 37 C with 30% formamide in 1xTBS + 0.1%

tween20 + 3mg/ml PVSA for ten minutes with agitation. To the sample 30uL of 2-5nM each encoding probe in 30% formamide + 1xTBS +0.1% tween20 + 3mg/ml PVSA + 10% dextran sulfate + 1mg/ml yeast tRNA + 1% murine RNase Inhibitor was added and a parafilm square was placed on top to prevent evaporation. 1 uM polyT acridite probe was added to hybe unless MelphaX was used. Sample was hybridized at 37 C for 36 hours in a humidity chamber. Sample was rinsed at 37 C with 30% formamide in 1xTBS + 0.1% tween20 + 3mg/ml PVSA four times for fifteen minutes with agitation. Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation.

Hydrogel Embedding

Sample was embedded in 50 uL of degassed 4% 19:1 acrylamide:bis-acrylamide in 1xTBS + 0.1% tween20 + 3mg/ml PVSA + 0.1% temed + 1% APS by inverting coverslip onto 50 uL of gel solution on a gel slick treated glass plate for 3 hours. Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation.

Clearing

Samples were digested in 1% proteinase k + 1xTBS + 0.1% triton x-100 + 3 mg/ml PVSA + 2mM CaCl2 800 mM Guanidine HCl ph 8 for 24-48 hours at 37 C with agitation. Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation.

Readout hybridization

Samples were hybridized in a custom built fluidics system. Samples were rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA and stripped of previous fluorophores in 25mM TCEP in 1xTBS + 0.1% tween20 + 3mg/ml PVSA for 10 minutes. Samples were rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA and 10% ethylene carbonate in 1xTBS + 0.1% tween20 +

3mg/ml PVSA. Readout probes were hybridized at 3nm in 10% ethylene carbonate in 1xTBS + 0.1% tween20 + 3mg/ml PVSA for 10 minutes. Samples were rinsed in 10% ethylene carbonate in 1xTBS + 0.1% tween20 + 3 mg/ml PVSA and 1xTBS + 0.1% tween20 + 3mg/ml PVSA. Samples were imaged in 0.1% rPCO + 2mM PCA + 2mM Trolox + 1xTBS + 0.1% Tween 20 + 3 mg/mL PVSA.

Imaging

Samples were imaged with a epifluorescent microscope at 63x with a ~100 nm pixel size and a flir camera.

Image Registration

Fiduciary markers embedded in the hydrogel were imaged for each round of hybridization. These markers were localized and paired to the first round of imaging. A rigid transformation was calculated in xyz to align all rounds of imaging.

Image Processing

Hot pixels were detected and corrected for each image. Chromatic aberrations between fluorophores were corrected. Backgrounds were calculated and subtracted using a high pass gaussian filter. High frequency noise was smoothed with a gaussian filter.

Spot based Image Decoding

Spots were detected and localized in the processed images and paired across rounds of imaging to form candidate transcripts. Candidates were matched to designed barcodes and candidates with more than a 1 bit error were removed. Additional transcripts were filtered based on signal to noise ratio by a logistic regressor using blank barcodes.

Pixel based Image Decoding

Processed images for each round of imaging were zscored and stacked. Vectors across the rounds of imaging for each individual pixel in xy were pulled. These vectors were l2 normalized and their Euclidean distance to l2 normalized codebook was generated. Neighboring pixels with the same decoded gene were collapsed into candidate transcripts. These transcripts were further filtered based on the number of pixels per transcript, their codeword distances(equivalent to a 1 but error), and their signal to noise ratios.

Image Segmentation

Nuclear images were background subtracted using a high pass gaussian filter. Cellpose was used to generate nuclear masks and a 5 um diameter voronoi dilation was used to generate cytoplasm masks. Transcripts within these masks were assigned to their respective cells.

3T3 Cell TNF Stimulation

Mouse 3T3 cells were plated onto coverslips within PDMS wells and allowed to grow overnight to a density of 60 to 80% confluency in DMEM. Cells were stimulated with 0, 1 or 10 ng of TNF-a for 3 hours prior to fixation.

Cornea Ex Vivo Wound

Mice were euthanized with carbon dioxide and cervical dislocation and their eyes were harvested with forceps. Eyes were washed and stored in 1xPBS at 4C for no more than a few hours. Epithelial debridement was induced using a 0.5mm rotating burr until a visible wound was formed. Eyes were cultured in DMEM at 37C for 2 or 15 hours. Corneas were removed from eyes with spring scissors under a dissecting microscope. Corneas were incubated at 37C in 0.5M EDTA to detach epithelium from stroma. Forceps were used to further separate the

epithelial whole mount from stroma. Epithelium was placed basal side down onto treated coverslips prior to fixation.

Zebra Finchmir128 Knock Down

Subjects were juvenile male zebra finches (Taeniopygia guttata), beginning at 30 days post-hatch. (30d) and raised to 75d. A total of 20 birds underwent stereotaxic neurosurgeries targeting Area X bilaterally. Ten birds from seven breeding pairs were treated by focal injection of an AAV bearing a miR-128 sponge sequence and 10 siblings were treated with a scrambled sequence as a control. Birds were primarily housed in home cages with parents and siblings, unless being recorded individually in a sound attenuation chamber. The vivarium and recording chambers are humidity- and temperature-controlled (22°C) and on a 12 hr light: dark cycle with half hour ‘dusks’ and ‘dawns’. Birdseed, water, millet, cuttlebone, and grit were provided ad libitum. Baths, hard boiled egg, and vegetables were provided weekly. Animal use was in accordance with the Institutional Animal Care and Use Committee at the University of California, Los Angeles and complied with the American Veterinary Medical Association Guidelines. Birds were isolated from the tutor at 10d and housed in a recording chamber with the mother and another female for care support. Birds were isolated in a recording chamber at 35d. At 120d surgeries were performed and the birds were allowed to recover for two weeks, then returned to their home cage with both parents. Birds were housed in their home cage for four weeks, with a break at two weeks to record changes in song. At ~165d birds were returned to recording chambers for behavioral experiments. On the final day birds were allowed to sing for two hours and then tissue was collected.

Author Contributions

ZH performed sample preparation, data acquisition for culture and cornea MERFISH. GT&ZH performed sample preparation and data acquisition for Zebra Finch MERFISH. ZH performed

data processing and data analysis for all MERFISH datasets. CA performed Zebra Finch mir128

Knockdown. ZH&RW designed the project. ZH wrote the manuscript.

References

Sartaj, R. et al. Characterization of slow cycling corneal limbal epithelial cells identifies putative stem cell markers. Sci. Rep. 7, 3793 (2017)

Sagga, N., Kuffová, L., Vargesson, N., Erskine, L. & Collinson, J. M. Limbal epithelial stem cell activity and corneal epithelial cell cycle parameters in adult and aging mice. Stem Cell Res. 33, 185–198 (2018)

Shaheen, B. S., Bakir, M. & Jain, S. Corneal nerves in health and disease. Surv. Ophthalmol. 59, 263–285 (2014)

Foulsham, W., Coco, G., Amouzegar, A., Chauhan, S. K. & Dana, R. When Clarity Is Crucial: Regulating Ocular Surface Immunity. Trends Immunol. 39, 288–301 (2018)

Liu, Q., Smith, C. W., Zhang, W., Burns, A. R. & Li, Z. NK cells modulate the inflammatory response to corneal epithelial abrasion and thereby support wound healing. Am. J. Pathol. 181, 452–462 (2012)

Yoon, J. J., Ismail, S. & Sherwin, T. Limbal stem cells: Central concepts of corneal epithelial homeostasis. World J. Stem Cells 6, 391–403 (2014)

Jeon, K.-I. et al. Corneal myofibroblasts inhibit regenerating nerves during wound healing. Sci. Rep. 8, 12945 (2018)

Espana, E. M. et al. The heterogeneous murine corneal stromal cell populations in vitro. Invest. Ophthalmol. Vis. Sci. 46, 4528–4535 (2005)

Li, J. et al. Identification for Differential Localization of Putative Corneal Epithelial Stem Cells in Mouse and Human. Sci. Rep. 7, 5169 (2017)

Chen, Z. et al. Characterization of putative stem cell phenotype in human limbal epithelia. Stem Cells 22, 355–366 (2004)

Notara, M., Lentzsch, A., Coroneo, M. & Cursiefen, C. The Role of Limbal Epithelial Stem Cells in Regulating Corneal (Lymph)angiogenic Privilege and the Micromilieu of the Limbal Niche following UV Exposure. Stem Cells Int. 2018, 8620172 (2018)

Ecoiffier, T., Yuen, D. & Chen, L. Differential distribution of blood and lymphatic vessels in the murine cornea. Invest. Ophthalmol. Vis. Sci. 51, 2436–2440 (2010)

Lee, E. J., Rosenbaum, J. T. & Planck, S. R. Epifluorescence intravital microscopy of murine corneal dendritic cells. Invest. Ophthalmol. Vis. Sci. 51, 2101–2108 (2010)

Kiesewetter, A., Cursiefen, C., Eming, S. A. & Hos, D. Phase-specific functions of macrophages determine injury-mediated corneal hem- and lymphangiogenesis. Sci. Rep. 9, 308 (2019)

Liu, J. et al. CCR2- and CCR2+ corneal macrophages exhibit distinct characteristics and balance inflammatory responses after epithelial abrasion. Mucosal Immunol. 10, 1145–1159 (2017)

Chinnery, H. R., McMenamin, P. G. & Dando, S. J. Macrophage physiology in the eye. Pflugers Arch. 469, 501–515 (2017)

Seyed-Razavi, Y., Chinnery, H. R. & McMenamin, P. G. A novel association between resident tissue macrophages and nerves in the peripheral stroma of the murine cornea. Invest. Ophthalmol. Vis. Sci. 55, 1313–1320 (2014)

He, J. & Bazan, H. E. P. Neuroanatomy and Neurochemistry of Mouse Cornea. Invest. Ophthalmol. Vis. Sci. 57, 664–674 (2016)

Li, Z., Burns, A. R. & Smith, C. W. Two waves of neutrophil emigration in response to corneal epithelial abrasion: distinct adhesion molecule requirements. Invest. Ophthalmol. Vis. Sci. 47, 1947–1955 (2006)

Park, M. et al. Visualizing the Contribution of Keratin-14+ Limbal Epithelial Precursors in Corneal Wound Healing. Stem Cell Reports 12, 14–28 (2019)

Ljubimov, A. V. & Saghizadeh, M. Progress in corneal wound healing. Prog. Retin. Eye Res. 49, 17–45 (2015)

Chapter 4

Highly Scalable Biocartographic Surveying At The Cellular Level Using Dimensionally Reduced Fluorescence In Situ Hybridization

Hemminger, Zachary; Tam, Gabriella; Xie, Fangming; Underwood, Thomas; Dong, Hong Wei; Wollman, Roy

Abstract

Single Cell technologies have allowed the molecular profiling of hundreds of thousands to millions of cells. These datasets have been used to generate catalogs of the cell types present in tissues and organs often referred to as atlases. Predominantly these technologies lose the spatial information of where in the tissue each cell originated from. Advances in spatial transcriptomics has allowed researchers to generate anatomical maps of where these cataloged cell types are within tissues. These approaches rely on single molecule imaging which requires high optical resolution. This limits the use of these technologies to relatively small regions of interest. In order to profile whole organs as well as larger tissues like human samples, multiple orders of magnitude scale increase is needed. Here we present Dimensionally Reduced Fluorescence In Situ Hybridization or dredFISH. dredFISH works by using cell type catalogs to design a linear projection matrix which can be used to measure a highly informative low dimensional representation of the cells gene expression. This measurement can be performed at the cell level as opposed to the single molecule level meaning that low magnification imaging can be used. With larger fields of view, more tissue can be profiled in the same amount of time by orders of magnitude. Using the mouse brain as a model, we show that this low dimensional gene expression measurement is highly informative, containing the cell type identity as well as information that can be used to reconstruct individual genes. dredFISH provides the scale improvements needed to begin generating detailed anatomical maps for entire organs as well

as larger tissues. The generation of these maps will be transformative for a number of biological fields and their understanding of physiology.

Introduction

A primary goal of biology is to understand how biological systems function. The scale of these systems can range from molecular to organismal and even beyond. A common approach to elucidate the function of a system is to determine the structure and composition of that system. For the organ and organismal level, the connection between the structure (anatomy) and the function (physiology) has been a key guiding principle. Historically this has been carried out through histological stains which measure a single anatomical component. Given the complexity of multicellular organisms, a single or even a handful of anatomical measurements is vastly insufficient to generate comprehensive anatomical maps. Complete structures of proteins have been pivotal in the understanding of their function. In order to understand the biological physiology at the organ and organismal level, complete anatomical maps are needed.

The importance to physiology of anatomical maps that detail the cellular composition of tissues as well as the location of every single cell is comparable to the importance of a fully mapped and annotated genome to genomics. Every subfield of physiology including:, neuroscience, oncology, developmental biology, immunology and every tissue specific physiology field will gain immediate and transformative understanding of their systems if provided a comprehensive single cell map (Lien et al 2017, Close et al 2021, Smith et al 2019, Baron et al 2020, Moncada et al 2020, Berglund et al 2018, Lohoff et al 2020, Mantri et al 2021, Chen et al 2022, Nerurkar et al 2020, Allam et al 2020) . Even beyond basic physiology, the generation of these maps will transform the field of pathology and is likely to lead to greater understanding of disease which should lead to novel therapeutics that are unlikely to be

discovered without these maps. The need for and impact of cartographic mapping of tissues at the cellular level is clear, yet the generation of these maps has been slow.

Since the 1700's histology has been the primary technique for generating maps of tissues (alturkistani et al 2015). While not surprising, the approach of taking samples and viewing a single or at most a handful of molecular markers fails to map the majority of the complexity present in a tissue. This can be overcome by performing different stains on different samples although something is lost when molecular markers cannot be visualized on the same sample. Sequential staining has allowed the number of markers mapped to increase. For proteins, the limitation here is the cost to generate a large number of highly specific antibodies and the need to measure one at a time. The introduction of multiplexed measurements has allowed scientists to generate spatial maps of many primarily nucleic acid molecular markers. These approaches do generate comprehensive maps of tissues but fundamental limitations in scalability prevent the scale of cartographic mapping that is needed to map all of the tissues, organs and organisms that anatomy currently studies (Lubeck et al 2014, Eng et al 2019,Shah et al 2017, Moffit et al 2016, Moffit et al 2016, Chen et al 2015, Rodriques et al 2019, vickovic et al 2019, salmen et al 2018, stahl et al 2016, Liu et al 2020, Chen et al 2021, Qian et al 2020, Lee et al 2014, Gyllborg et al 2020, Wang et al 2018, Ke et al 2013, Alon et al 2020).

All of these approaches rely on single molecule imaging or barcoded sequencing. Sequencing approaches are limited by their nucleic acid capture efficiency, predominantly lower spatial resolution, and the high cost of sequencing. Image based approaches are limited by their need to work with high optical resolution. This significantly reduces the area that can be imaged. While exceptions exist for image based approaches that can handle large areas, they typically come at the expense of nucleic acid capture efficiency or the number of molecular markers that can be measured at a time. There is a clear need for techniques that can capture a large

amount of molecular information at high spatial resolution at scales that far exceed the current approaches.

Extensive single cell transcriptomics is actively being performed to identify cell type populations present in the organs of common model systems as well as humans and to characterize them transcriptionally (Svensson et al 2020). The outcome of these efforts is cell type labeled single cell gene expression data. In order to improve the scalability of spatial transcriptomics is to only target highly informative genes. For imaging based spatial transcriptomics, these datasets have been used to identify which genes should be targeted. While this decreases the number of target genes dramatically, a large number of genes are still needed in order to map closely related cell types. Sequencing based approaches are predominantly untargeted and so do not gain any benefit in scalability from these composition datasets. For both image and sequencing based spatial transcriptomics these datasets have been used to assign cell type labels to their measured data (Korsunsky et al 2019). With cell type assignment being a key goal of these technologies, this raises the question as to what is the minimal amount of information necessary to call cell types accurately and how few measurements are needed to capture that information?

Dimensionality reduction directly addresses this question. Linear dimensionality reduction in the form of principal component analysis is often a first step in the analysis of single cell transcriptomic data including spatial transcriptomics data. It is computationally expensive and a bit noisy to compare the entire gene expression profile of every single cell to every other single cell. Dimensionality reduction compresses the full gene expression vector into a few dozen highly informative numbers.This dimensionally reduced transcriptome is a weighted linear sum of gene expression. This allows cell to cell comparison in a computationally efficient manner to group cells of similar gene expression profiles as defined by convention. Not only is dimensionality reduction used to group cells from the same dataset, it is also used to integrate

multiple datasets and to transfer labels across datasets. Clearly these dimensionally reduced transcriptomes still contain the cell type information. This raises the question as to why it is necessary to measure individual genes at all.

Here we present dimensionally reduced fluorescence in situ hybridization or dredFISH which is a nucleic acid and image based technique for designing highly informative dimensionally reduced measurements of gene expression which contains sufficient information for cell type classification which can be measured without the need for high optical resolution. dredFISH also includes the computational framework necessary to recover cell type information from these measured stains. This allows labeled single cell cartographic maps of large tissue areas with multiple orders of magnitude speed increase due to large field of views. We performed dredFISH on the mouse brain to generate spatially resolved single cell cartographic maps at throughputs that far exceed alternatives. We also present a framework for taking these single cell maps and identifying anatomical features which are visually interpretable to classically trained physiologists and pathologists as well as statistically accessible for bioinformaticians.

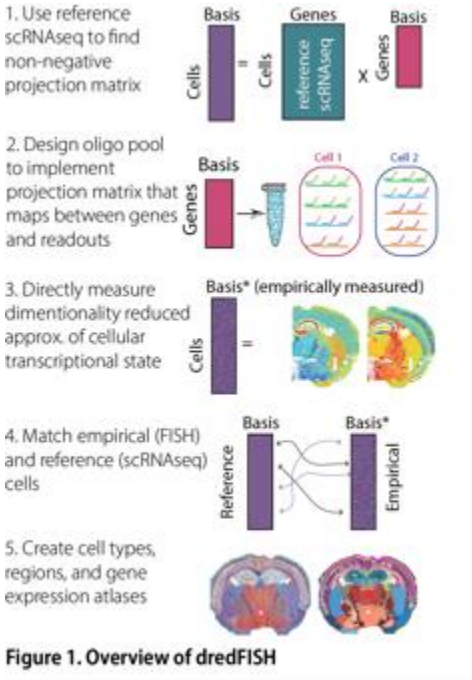

Figure 4.1: dredFISH Methodology

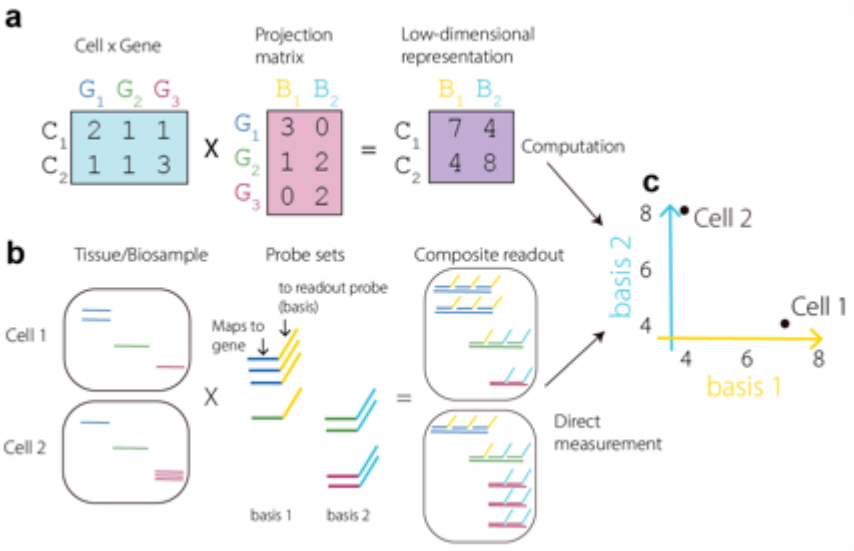

Figure 4.2: dredFISH Molecular Example

A. Block matrix diagram showing the expression of three genes for two cells, a projection matrix showing the weight that each gene has in the two basis measurements and a low dimensional representation of the gene expression of both cells for each basis. B. Molecular diagram of A showing the expression of three genes within two cells. The projection matrix is implemented by generating bivalent dna probes for each non negative value within the projection matrix. These probes contain 2 sets of binding sequences the first targets the gene and the second targets a fluorescent readout probe that will be used to measure the basis. The number of probes that connect a gene to a basis is equal to the value within the projection matrix (i.e. a value of 3

means 3 probes that connect G1 to B1). By binding these probes to the transcripts within the cell a composite readout is generated. The sum of the basis sequences for each cell is equal to the low dimensional basis for that cell. Each of these basis sequences will be read out with a unique fluorescent readout probe. C . Together this low dimensional representation or dredFISH space can be used to separate cells by their gene expression in a way that does not require measuring each gene individually. The positions of cells in this space are the same if measured directly with dredFISH or if measuring genes individually like scRNAseq and projecting with the projection matrix.

dredFISH Methodology

dredFISH can be broken into three key parts, encoding, measurement and decoding. Encoding consists of using reference single cell RNA seq data to design a projection matrix that will be used to compress the gene expression into basis factors without losing cell type information. The experimental measurement consists of implementing the designed projection matrix into molecular probes that will be used to stain samples and the acquiring of the in situ basis factor measurement for each cell spatially as scale. Decoding consists of grouping measured cells with similar basis factor measurements and then using the original reference single cell RNAseq data to transfer labels to the measured cells.

Encoding

The reference single cell RNA sequencing data that was used to generate the encoding was the Allen Brain Atlas which consisted of ~70 thousand cells measured with smart seq and ~1.3 million cells measured with 10X. These cells were labeled with multiple hierarchical labels, the coarsest of which are inhibitory neurons, excitatory neurons, and non-neurons while the finer labels delineated to canonical cell types and even finer clusters. Lowly expressed genes were excluded due to their noisy nature in the reference data. Highly expressed genes were also excluded because they would disproportionately affect the optimization function compared to other normally expressed genes. The projection matrix was calculated with discriminant projective non-negative matrix factorization or DPNMF. The optimization function for this algorithm was to maximize the gene reconstruction as well as minimize variability within a cell

type and maximize the variability between cell types (Song et al 2021). A mu parameter weighted the reconstructive versus the discriminant aspects of the cost function. A large mu parameter of 50 was chosen to provide greater discriminant performance at the price of reconstruction accuracy. DPNMF was chosen over PCA as the resulting weights were non-negative and sparse which is essential when converting the projection matrix to molecular probes.

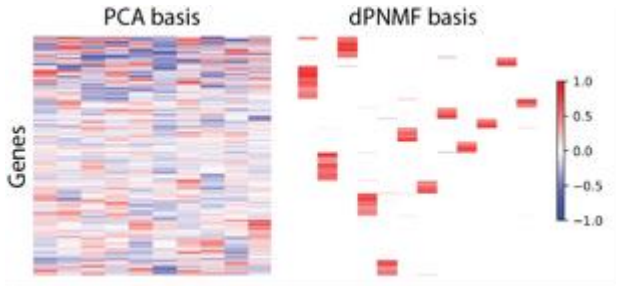

Figure 4.3: dredFISH Encoding Example
A. The first 12 basis of PCA and DPNMF calculated using the Allen institute for Brain Science scRNAseq reference data. DPNMF contains only non-negative values and the majority of gene to basis projections are zero.

Converting Encoding to Molecular Probes

The projection matrix was scaled and integerized so that the total sum of integers across all genes and basis would be ~90 thousand. This number was chosen based on the maximum number of DNA probes that could be purchased at the time. Integerized projection weights were directly converted into the number of FISH probes that needed to be designed for each gene. Genes that required more probes than could be designed had their projection matrix values clipped to the maximum designable probes. For each probe 3x20bp sequences were added depending on which basis the gene had weight in on the integerized projection matrix. Probes were ordered as a pool from custom arrays.

Measurement

Samples were prepared in a method sharing most steps with MERFISH. In short, Samples were sectioned to 10 um and placed on treated coverslips. Samples were fixed with PFA, permeabilized with Triton in PBS and stored at -20C in 70% Ethanol. Samples had their RNA modified with melphaX to allow anchoring to a later hydrogel. Ordered Probes were amplified using PCR to add a T7 promoter then IVT to increase the number of copies and lastly RT to convert ssRNA to ssDNA. Probes were hybridized in a 30% formamide hybridization solution at 37C for 36 hours and washed at 37 C with a 30% formamide wash buffer for 1 hour. Samples were embedded in a thin 4% polyacrylamide hydrogel and then cleared in a proteinase k buffer with CaCl overnight at 37C.

Fluorescent readout probes were hybridized in an automated fluidics system for 10 minutes and then imaged with a custom epifluorescent microscope. Fluorophores were stripped off of the readout probes with a reducing agent between rounds of imaging. Background images were acquired prior to the first readout probe as well as after stripping of each round's fluorophore. Images for each round and background were stitched together using a nuclear stain. Cells were segmented using cellpose on the nuclear stain as well as a polyT staining. Intensities for each cell for each measurement was pulled from the stitched images and collapsed to generate a vector of basis expression for each measured cell. This provides a cell by basis matrix as well the spatial coordinates for each cell's location within the sample.

Basis measurements show clear spatial patterns that change for each basis measurement which align visually with known anatomical features. Cells within this dimensionally measured space also show separable patterns of basis factor expression when projected into UMAP. This shows a visualization of the amount of cell type specific information that was preserved during the measurement. dredFISH measurements are shown to be highly informative for generating cartographic maps as shown when the cells umap localization is

projected onto their spatial coordinates. This visualization shows how highly informative the dredFISH measurements are for generating spatial maps of tissues.

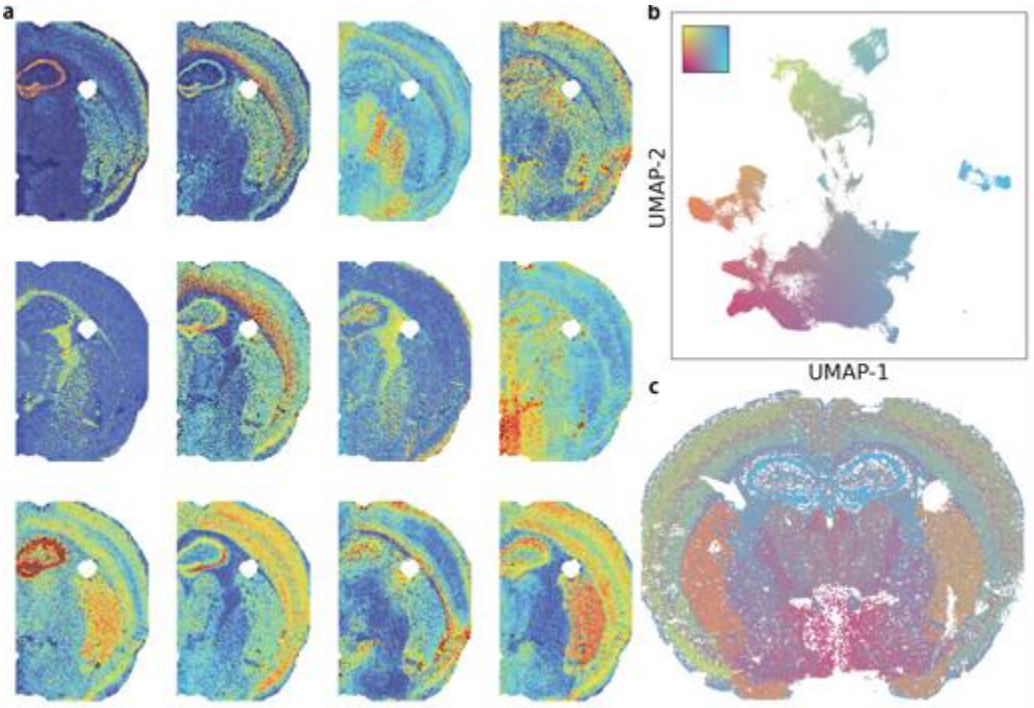

Figure 4.4: dredFISH Measurement.
A. 12 representative experimental dredFISH Measurements out of 24 in one hemisphere of the mouse brain coronal section containing ~50k cells. Cells expressing more of the genes which contained non zero values within a single basis of the projection matrix show higher dredFISH signal when measured. Distinct patterns are present as the result of different basis measuring different weighted sets of genes. B. UMAP visualization of all 24 dredFISH measurements. Colors show the position within this UMAP space. C. Spatial coordinates of a whole coronal section ~100k cells. Cells are colored according to B. Clear known anatomical features are visible within this visualization.
TMG

Visual inspection of cartographic maps is useful but lacks the quantitative nature necessary for rigorous statistical analysis. For that reason, we developed a computational architecture for analyzing these spatially measured dimensionally reduced transcriptomes. This can be done in approaches that mimic how other forms of single cell and spatial single cell transcriptomics data are analyzed. On the gene expression side: cells can be grouped together based on their shared gene expression measurement, labels can be transferred and gene expression can be imputed across datasets. On the spatial side: space can be used to define

the granularity at which cell types should be resolved, zones of homogeneous cell types can be found, and more complex regions making up homogeneous or heterogeneous populations can be defined quantitatively.

Supervised Decoding

Cells present in regions that were captured within the reference can use their highly informative dimensionally reduced gene expression vector to find cells with similar gene expression in the reference dataset. Reference single cell RNA seq datasets were projected using the same projection matrix that was used to generate the molecular probes. Normalization operations were performed on the reference as well as measured data. Basis vectors for each cell were normalized by their sum to account for uneven staining efficiencies as well as differences in segmentation and total RNA content. Basis measurements were normalized using Zscore to correct for differences in staining efficiency across rounds of imaging and to move the sequencing data and the dredFISH data to the same shared space. By aligning the measured and reference datasets into the same space, we can transfer information known about the reference data onto dredFISH cells that are nearby in the shared space as defined by a cosine similarity metric. The simplest but possibly most useful bit of information that can be transferred is the labeled cell type.

dredFISH is also shown to be useful for cells that were not present in the reference dataset. Areas outside of the Hippocampus and Cortex were not sampled in the current Allen Mouse Brain Atlas. Despite this for coarse level cell types like excitatory neuron, inhibitory neuron and non-neuron, the dredFISH cells that were outside of the hippocampus and cortex shown spatial organization that qualitatively aligns well with spatial transcriptomic data generated by other technologies.

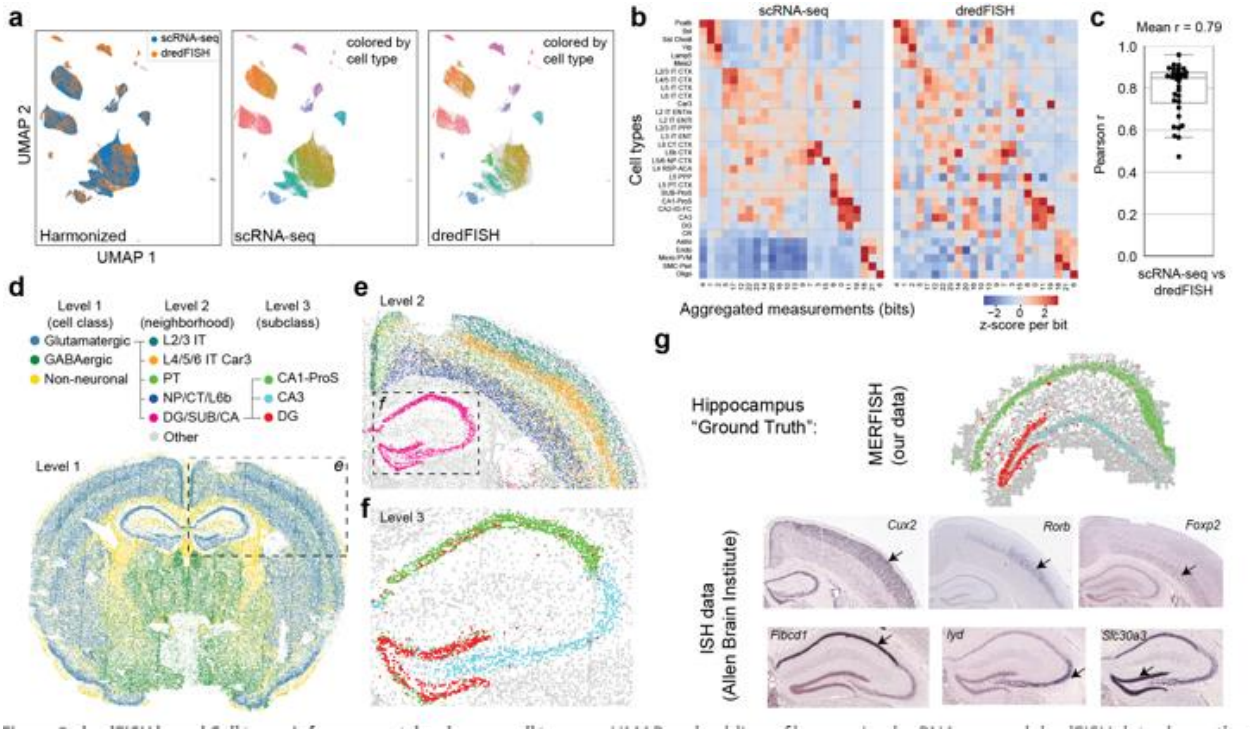

Figure 4.5: dredFISH Cell Type Labeling.
A. UMAP Embedding of measured dredFISH data as well as scRNAseq data that has been projected using the DPNMF projection matrix. Datasets were harmonized by iterative zscore normalization through 3 hierarchical cell type annotations. B. Visual of the cell type average low dimensional gene expression projection for measured and reference data. C. Average Pearson correlation of 0.79 between reference and measured data for each cell type. D. Iterative hierarchical label transfer results showing coarse cell type of excitatory, inhibitory and non-neuronal then finer cell types like the layers of the cortex as well as even finer cell type annotations like excitatory hippocampal neuronal cell types (CA1,CA3, and DG). G. Fine cell type localization of excitatory hippocampal neuronal cell types (CA1, CA3, and DG) detected using a gold standard MERFISH. H. Immunofluorescence standards for cell type markers showing agreement of dredFISH and MERFISH cell type localizations.

Unsupervised Clustering

Fine cell types are notoriously difficult to transfer labels between datasets .For this reason unsupervised clustering of cells at finer resolution than transferred labels is common. This can also be applied to measured dredFISH data. Clustering of cells can be highly sensitive to the resolution you allow. Too high of resolution and you may split cells of the same type arbitrarily. Too low of resolution and you may lump different cell types into the same type. One approach to do this is to define distinct cell types as those that differ transcriptionally as well as spatially. By using leiden clustering and optimizing the resolution parameter to maximize

difference in the entropy in the gene expression label space and the spatial zone space. This approach splits cells into finer and finer cell types until the types are no longer heterogeneously situated in space. Doing this generates ~100 cell types of finer resolution than label transfer.

Regions

Cell types and their locations are useful visually but can also be used to generate quantitative definitions for regions within tissues. Using these finer cell types we can cluster cells based on the identities of their neighbors. Like clustering of cell types this can also be sensitive to resolution. By using space as well as neighbor composition, we can set resolution limits that generate informative spatial regions. These regions align visually remarkably well with known anatomical features. Surprisingly this was true for anatomical features whose gene expression was outside of the cortex and hippocampus and as such was not designed for. This suggests a generalizable nature to the dredFISH stain for cells not sampled in the reference.

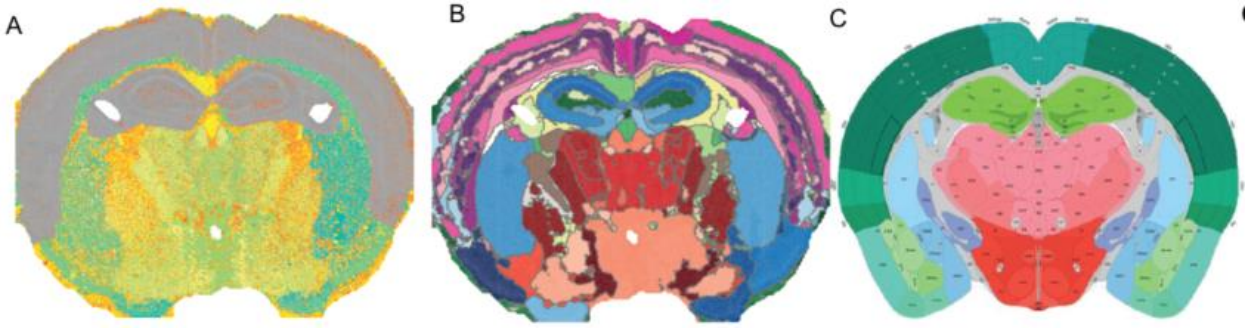

Figure 4.6: Unsupervised Clustering and Region Identification.
A. ~100 cell types identified with leiden graph based cluster analysis shown to be different transcriptionally as well as spatially. B. Regions identified by nearest neighbors identity topic modeling. C Established anatomical regions defined by the Allen Brain Atlas showing strong agreement with de novo generated regions.

Gene Reconstruction

Labels are not the only information that can be gained by aligning measured dredFISH data and reference scRNAseq data. Gene expression can be reconstructed for measured data by using the average of the 10 nearest neighbors within the reference dataset. The reconstruction accuracy can be predicted by comparing the average of 10 nearest neighbors in

dredFISH space for the reference data to the actual measured expression for that cell. Reconstruction accuracy varied but correlated strongly with the weight that a gene has in the projection matrix as well as how much variance can be explained for that gene by PCA. Genes with low variance explained by PCA are likely noisy either due to technical noise in sequencing or other factors that make some of the variance unexplainable by linear approaches. To show the accuracy of reconstructed gene expression the reconstructed gene expression patterns for three genes are compared to measured in situ hybridization provided by the Allen Brain Atlas. Expression patterns show strong agreement between measured and reconstructed gene expression patterns.

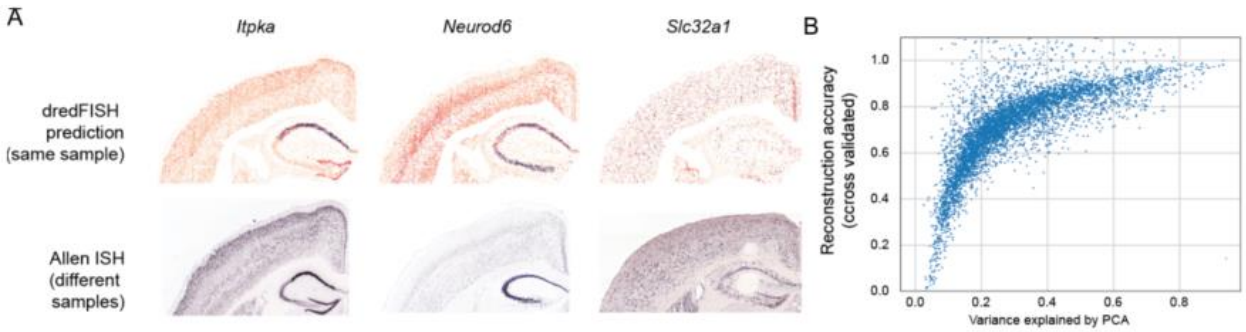

Figure 4.7: Gene Reconstruction.
A. Gene expression patterns reconstructed from dredFISH measurements as well as Allen ISH data for the same genes showing qualitative agreement. B. Expected gene reconstruction accuracy for each gene versus variance explained for that gene by PCA. Genes with >1 reconstruction occur when KNN out performs linear PCA.

Discussion

Detailed anatomical maps provide a clear path to increasing our understanding of the physiology of multicellular biology. Despite improvements to spatial single cell technologies, the scale needed to generate these anatomical maps is not present especially for larger complex tissues. As of 2022, the largest anatomical map generated by spatial transcriptomics was ~1 million cells with the average being closer to between 10,000 and 100,000 cells (Moses & Patcher 2022). Relatively small model systems such as the mouse brain contain ~100 million

cells. It is unlikely that most of the anatomical detail can be captured if less than 1 percent of the cells are measured spatially. When thinking about larger systems like the human brain, which contains ~100 billion cells, it is clear that we need technology that can measure cells at multiple orders of magnitude scale increase.

Here we present dredFISH, an imaging based spatial transcriptomic method that works by designing a weighted aggregate gene expression measurement that can be measured at the cell level using FISH. We show that DPNMF can be used to learn a projection matrix from an annotated single cell reference dataset. We show that this projection matrix can be implemented in a molecular probe and can be measured using standard epifluorescent microscopy. Lastly we show that the information contained within the low dimensional gene expression measurement is informative and can be used to identify cell types, identify spatial regions within a tissue and even reconstruct gene expression with similar performance to standard linear approaches. These show that dredFISH is as capable of generating highly detailed data as existing spatial transcriptomic methods. The key distinction between dredFISH and existing methods is the scale at which dredFISH can be performed.

Existing image based spatial transcriptomic techniques rely on detecting diffraction limited spots. Optical resolutions in the hundreds of nanometers are needed in order to detect a reasonable number of transcripts per round of imaging. dredFISH relies on detecting and quantifying cells rather than diffraction limited spots which are 10’s to 100’s of times smaller than cells. Due to this, dredFISH can operate with optical resolutions in the micron range and potentially higher. This means dredFISH can be measured with at least 10 fold lower magnification than single molecule techniques. This translates to a 10 fold larger field of view in each dimension. In two dimensions this means dredFISH can measure the same cells 100 fold faster while in three dimensions dredFISH should perform at 1000 fold faster.

The ability to generate detailed anatomical maps at single cell resolution hundreds to thousands of times faster than existing technologies is likely to cause a paradigm shift in how we approach investigating physiology. Having the ability to profile every cell in an organ with biological replicates and even across experimental conditions will generate enormous datasets that cannot be interpreted by eye. This is especially true if datasets are collected for entire human tissues and organs. Computational algorithms will need to be developed to quantify and summarize the detailed information present in these datasets. Just as single cell sequencing sparked a wave of bioinformaticians generating algorithms, we expect dredFISH to generate sufficient data to spark a new wave focused not only on the gene expression space but also on the anatomical space. Together dredFISH and these algorithms are likely to generate unique physiological insights that likely could not have been understood without the scale increase that dredFISH provides.

Materials And Methods

Encoding Probe Design and Synthesis

Encoding probes were designed using existing software to generate 30 base pair specific homology to RNA targets with a gc content of 45 to 65% and a melting temperature of 65 to 72C. Readout probe bonding sequences were concatenated to the encoding regions and amplification primers were designed and appended to both ends.

Coverslip functionalization

40mm round type 1.5 coverslips were cleaned in a 50:50 mixture of 37% concentrated HCl and Methanol for 30 minutes with sonication. Coverslips were rinsed with deionized water 3 times for 5 minutes each, once in Ethanol, and dried at 70C. Coverslips were modified with 0.2% allyltrichlorosiloxane in chloroform with 0.1% triethylamine for 30 minutes with sonication to facilitate hydrogel adhesion. Coverslips were rinsed once with chloroform, twice with ethanol

and dried for 1 hour at 70C. In cases where additional sample adhesion is necessary, Coverslips were modified with 2% aminopropyltriethoxysilane in acetone for 2 minutes. Coverslips were rinsed with deionized water twice, ethanol once, and dried at 70C.

Fixation

Samples were placed on functionalized 40mm round coverslips and fixed in cold 4% paraformaldehyde in 1xPBS for 5 minutes for cells and 15 minutes for tissues with agitation and washed three times in 1xPBS with 3 mg/mL poly vinylsulfonic acid (PVSA) and 0.1% triton x-100 for 5 minutes each with agitation. Samples were buffer exchanged into 70% ethanol and stored at -20C.

Permeabilization

Samples were rinsed with 1xPBS with PVSA and 0.1% triton x-100 three times for five minutes each with agitation. Samples were permeabilized with 1% triton in 1xPBS with PVSA for 30 minutes at 37 C with agitation. Samples were rinsed with 1xPBS with 0.1% triton x-100 and PVSA three times for five minutes each with agitation.

RNA Modification with MelphaX

Samples were rinsed with 30 mM MOPS ph 7.7 + 0.1% triton x-100 + 3mg/mL PVSA three times for five minutes each with agitation. To the same 50 uL of 1 mg/ml MelphaX in MOPS was added and a parafilm square was placed on top to prevent evaporation. Sample was reacted at 37 C overnight in a humidity chamber. Sample was washed with 1xPBS + 0.1% triton x-100 + PVSA three times for 5 minutes each with agitation.

Encoding Hybridization

Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation. Sample was rinsed at 37 C with 30% formamide in 1xTBS + 0.1% tween20 + 3mg/ml PVSA for ten minutes with agitation. To the sample 30uL of 2-5nM each encoding probe in 30% formamide + 1xTBS +0.1% tween20 + 3mg/ml PVSA + 10% dextran sulfate + 1mg/ml yeast tRNA + 1% murine RNase Inhibitor was added and a parafilm square was placed on top to prevent evaporation. 1 uM polyT acridite probe was added to hybe unless MelphaX was used. Sample was hybridized at 37 C for 36 hours in a humidity chamber. Sample was rinsed at 37 C with 30% formamide in 1xTBS + 0.1% tween20 + 3mg/ml PVSA four times for fifteen minutes with agitation. Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation.

Hydrogel Embedding

Sample was embedded in 50 uL of degassed 4% 19:1 acrylamide:bis-acrylamide in 1xTBS + 0.1% tween20 + 3mg/ml PVSA + 0.1% temed + 1% APS by inverting coverslip onto 50 uL of gel solution on a gel slick treated glass plate for 3 hours. Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation.

Clearing

Samples were digested in 1% proteinase k + 1xTBS + 0.1% triton x-100 + 3 mg/ml PVSA + 2mM CaCl2 800 mM Guanidine HCl ph 8 for 24-48 hours at 37 C with agitation. Sample was rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA three times for five minutes with agitation.

Readout hybridization

Samples were hybridized in a custom built fluidics system. Samples were rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA and stripped of previous fluorophores in 25mM TCEP

in 1xTBS + 0.1% tween20 + 3mg/ml PVSA for 10 minutes. Samples were rinsed with 1xTBS + 0.1% tween20 + 3mg/ml PVSA and 10% ethylene carbonate in 1xTBS + 0.1% tween20 + 3mg/ml PVSA. Readout probes were hybridized at 3nm in 10% ethylene carbonate in 1xTBS + 0.1% tween20 + 3mg/ml PVSA for 10 minutes. Samples were rinsed in 10% ethylene carbonate in 1xTBS + 0.1% tween20 + 3 mg/ml PVSA and 1xTBS + 0.1% tween20 + 3mg/ml PVSA. Samples were imaged in 0.1% rPCO + 2mM PCA + 2mM Trolox + 1xTBS + 0.1% Tween 20 + 3 mg/mL PVSA.

Imaging

Samples were imaged with a epifluorescent microscope at 10x with a ~~500 nm pixel size and a flir camera after readout hybridization and between the strip of the previous round.

Image Registration

Images for the first round of imaging were stitched together using a 10% overlap to correct for stage inaccuracy. In short a rigid transformation was calculated from phase cross correlation between neighboring images using nuclear stain images. For subsequent rounds of imaging, a rigid transformation was calculated from phase cross correlation between the nuclear stain images of the first round and the subsequent rounds.

Image Processing

Background images for each round were subtracted from readout images and a flatfield correction was applied. Flatfield was calculated for each pixel as the median after background subtraction and the entire flatfield was divided by the median across all pixels. A secondary background subtraction was performed using a minimum filter with a window of ~100 um.

Segmentation

Stitched nuclear images were binned into ~100 bins. A background subtraction was performed for each image using a minimum filter with a window of ~100 um. Images were segmented using cellulose with a diameter of ~10um. For Cytoplasm segmentation, the same operation was performed on a poly T readout.

Vector Pulling and Normalization

The median intensity of the pixels within a mask for each round of imaging was used as the basis measurement for that cell for that round. To account for uneven staining, the basis vector for each cell was divided by the sum of the vector. To account for round to round variation and to put the vector into a common space, the basis was zscored across all cells.

Supervised Classification

Reference scRNAseq data was first cell size normalized and then projected using the same projection matrix that was used to design the encoding probes. Each bit was normalized using zscore to put the cells in the same common space as the measured dredFISH cells. For each cell the 10 nearest neighbors in the reference data was calculated using cosine distance. The most comment label within these 10 neighboring cells was assigned to the measured cell. This was first performed for the coarsest level of annotation. Which contained 3 cell types. For each of the assigned cell types, the reference and measured data was subsetted to cells that contained that annotation. The whole process including z score normalization across cells within a bit was performed and the next level of annotation was transferred to the measured data using the 10 nearest neighbors within the subset. This process was repeated iteratively for the first 3 hierarchical annotations within the reference dataset.

Unsupervised Classification

Cells were clustered using leiden clustering on the normalized dredFISH vectors. The optimal resolution for this clustering was calculated to be the resolution that maximized the entropy difference between the number of cell types and the number of homogeneous spatial zones.

Regions

Regions were identified using Latent Dirichlet Allocation (LDA) on the local cell type composition for each cell using their unsupervised clustering labels. Cells were labeled to be part of a region by the topic that explained most of that cell's local composition.

Author Contributions

TU&ZH designed fluidics chamber. FX&ZH performed supervised label transfer. GT&ZH performed sample preparation and data acquisition. ZH performed data processing and instrumentation design. RW&ZH designed the project. ZH wrote the manuscript.

References

Alturkistani, H. A., Tashkandi, F. M. & Moha145mmedsaleh, Z. M. Histological Stains: A Literature Review and Case Study. Glob. J. Health Sci. 8, 72–79 (2015)

Song, D., Li, K., Hemminger, Z., Wollman, R. & Li, J. J. scPNMF: sparse gene encoding of single cells to facilitate gene selection for targeted gene profiling. Bioinformatics 37, i358–i366 (2021)

Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16, 1289–1296 (2019)

Lein E, Borm LE, Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. Science. 2017 Oct 6;358(6359):64–69. PMID: 28983044

Close JL, Long BR, Zeng H. Spatially resolved transcriptomics in neuroscience. Nat Methods. 2021 Jan;18(1):23–25. PMID: 33408398

Smith EA, Hodges HC. The Spatial and Genomic Hierarchy of Tumor Ecosystems Revealed by Single-Cell Technologies. Trends Cancer Res. Elsevier; 2019 Jul;5(7):411–425. PMCID: PMC6689240

Baron M, Tagore M, Hunter MV, Kim IS, Moncada R, Yan Y, Campbell NR, White RM, Yanai I. The Stress-Like Cancer Cell State Is a Consistent Component of Tumorigenesis. Cell Syst. Elsevier; 2020 Nov 18;11(5):536–546.e7. PMCID: PMC8027961

Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, Hajdu CH, Simeone DM, Yanai I. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol. nature.com; 2020 Mar;38(3):333–342. PMID: 31932730

Berglund E, Maaskola J, Schultz N, Friedrich S, Marklund M, Bergenstråhle J, Tarish F, Tanoglidi A, Vickovic S, Larsson L, Salmén F, Ogris C, Wallenborg K, Lagergren J, Ståhl P, Sonnhammer E, Helleday T, Lundeberg J. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat Commun. nature.com; 2018 Jun 20;9(1):2419. PMCID: PMC6010471

Lohoff T, Ghazanfar S, Missarova A, Koulena N, Pierson N, Griffiths JA, Bardot ES, Eng CHL, Tyser RCV, Argelaguet R, Guibentif C, Srinivas S, Briscoe J, Simons BD, Hadjantonakis AK, Göttgens B, Reik W, Nichols J, Cai L, Marioni JC. Highly multiplexed spatially resolved gene expression profiling of mouse organogenesis . bioRxiv. 2020 [cited 2022 May 12]. p. 2020.11.20.391896.

Mantri M, Scuderi GJ, Abedini-Nassab R, Wang MFZ, McKellar D, Shi H, Grodner B, Butcher JT, De Vlaminck I. Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. Nat Commun. 2021 Mar 19;12(1):1771. PMCID: PMC7979764

Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, Qiu X, Yang J, Xu J, Hao S, Wang X, Lu H, Chen X, Liu X, Huang X, Li Z, Hong Y, Jiang Y, Peng J, Liu S, Shen M, Liu C, Li Q, Yuan Y, Wei X, Zheng H, Feng W, Wang Z, Liu Y, Wang Z, Yang Y, Xiang H, Han L, Qin B, Guo P, Lai G, Muñoz-Cánoves P, Maxwell PH, Thiery JP, Wu QF, Zhao F, Chen B, Li M, Dai X, Wang S, Kuang H, Hui J, Wang L, Fei JF, Wang O, Wei X, Lu H, Wang B, Liu S, Gu Y, Ni M, Zhang W, Mu F, Yin Y, Yang H, Lisby M, Cornall RJ, Mulder J, Uhlén M, Esteban MA, Li Y, Liu L, Xu X, Wang J. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell . 2022 Apr 22; PMID: 35512705

Nerurkar SN, Goh D, Cheung CCL, Nga PQY, Lim JCT, Yeong JPS. Transcriptional Spatial Profiling of Cancer Tissues in the Era of Immunotherapy: The Potential and Promise. Cancers . 2020 Sep 9;12(9). PMCID: PMC7563386

Allam M, Cai S, Coskun AF. Multiplex bioimaging of single-cell spatial profiles for precision cancer diagnostics and therapeutics. NPJ Precis Oncol. 2020 May 1;4:11. PMCID: PMC7195402

Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nature methods. nature.com; 2014. p. 360–361. PMCID: PMC4085791

Eng CHL, Lawson M, Zhu Q, Dries R, Koulena N, Takei Y, Yun J, Cronin C, Karp C, Yuan GC, Cai L. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. Nature. Nature Publishing Group; 2019 Apr;568(7751):235–239. PMCID: PMC6544023

Shah S, Lubeck E, Zhou W, Cai L. seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus. Neuron. Elsevier; 2017 May 17;94(4):752–758.e1. PMID: 28521130

Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc Natl Acad Sci U S A . 2016 Sep 13; PMID: 27625426

Moffitt JR, Hao J, Bambah-Mukku D, Lu T, Dulac C, Zhuang X. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. Proc Natl Acad Sci U S A. 2016 Dec 13;113(50):14456–14461. PMCID: PMC5167177

Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, Rubinstein ND, Hao J, Regev A, Dulac C, Zhuang X. Molecular, spatial and functional single-cell profiling of the hypothalamic preoptic region. Science . 2018 Nov 1; PMID: 30385464

Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. American Association for the Advancement of Science; 2015 Apr 24;348(6233):aaa6090. PMID: 25858977

Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science. 2019 Mar 29;363(6434):1463–1467. PMCID: PMC6927209

Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, Äijö T, Bonneau R, Bergenstråhle L, Navarro JF, Gould J, Griffin GK, Borg Å, Ronaghi M, Frisén J, Lundeberg J, Regev A, Ståhl PL. High-definition spatial transcriptomics for in situ tissue profiling. Nat Methods. 2019 Oct;16(10):987–990. PMCID: PMC6765407

Salmén F, Ståhl PL, Mollbrink A, Navarro JF, Vickovic S, Frisén J, Lundeberg J. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. Nat Protoc. 2018 Nov;13(11):2501–2534. PMID: 30353172

Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, Mollbrink A, Linnarsson S, Codeluppi S, Borg Å, Pontén F, Costea PI, Sahlén P, Mulder J, Bergmann O, Lundeberg J, Frisén J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016 Jul 1;353(6294):78–82. PMID: 27365449

Liu Y, Yang M, Deng Y, Su G, Enninful A, Guo CC, Tebaldi T, Zhang D, Kim D, Bai Z, Norris E, Pan A, Li J, Xiao Y, Halene S, Fan R. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell. 2020 Dec 10;183(6):1665–1681.e18. PMCID: PMC7736559

Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, Yang J, Li W, Xu J, Hao S, Chen X, Liu X, Huang X, Lin F, Tang X, Li Z, Hong Y, Fu D, Jiang Y, Peng J, Liu S, Shen M, Liu C, Li Q, Wang Z, Wang Z, Yuan Y, Volpe G, Ward C, Muñoz-Cánoves P, Thiery JP, Zhao F, Li M, Kuang H, Wang O, Lu H, Wang B, Ni M, Zhang W, Mu F, Yin Y, Yang H, Lisby M, Cornall RJ, Uhlen M, Esteban MA, Li Y, Liu L, Wang J, Xu X. Large field of view-spatially resolved transcriptomics at nanoscale resolution . Cold Spring Harbor Laboratory. 2021 [cited 2021 Mar 4]. p. 2021.01.17.427004.

Qian X, Harris KD, Hauling T, Nicoloutsopoulos D, Muñoz-Manchado AB, Skene N, Hjerling-Leffler J, Nilsson M. Probabilistic cell typing enables fine mapping of closely related cell types in situ. Nat Methods. 2020 Jan;17(1):101–106. PMCID: PMC6949128

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly Multiplexed Subcellular RNA Sequencing in Situ. Science. 2014 Mar 21;343(6177):1360–1363.

Gyllborg D, Langseth CM, Qian X, Choi E, Salas SM, Hilscher MM, Lein ES, Nilsson M. Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. Nucleic Acids Res. 2020 Nov 4;48(19):e112. PMCID: PMC7641728

Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, Evans K, Liu C, Ramakrishnan C, Liu J, Nolan GP, Bava FA, Deisseroth K. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science . 2018 Jul 27;361(6400). PMCID: PMC6339868

Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, Nilsson M. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods. 2013 Sep;10(9):857–860. PMID: 23852452

Alon S, Goodwin DR, Sinha A, Wassie AT, Chen F, Daugharthy ER, Bando Y, Kajita A, Xue AG, Marrett K, Prior R, Cui Y, Payne AC, Yao CC, Suk HJ, Wang R, Yu CC (jay), Tillberg P, Reginato P, Pak N, Liu S, Punthambaker S, Iyer EPR, Kohman RE, Miller JA, Lein ES, Lako A, Cullen N, Rodig S, Helvie K, Abravanel DL, Wagle N, Johnson BE, Klughammer J, Slyper M, Waldman J, Jané-Valbuena J, Rozenblatt-Rosen O, Regev A, IMAXT Consortium, Church GM, Marblestone AH, Boyden ES. Expansion Sequencing: Spatially Precise In Situ Transcriptomics in Intact Biological Systems . Cold Spring Harbor Laboratory. 2020 [cited 2020 Dec 15]. p. 2020.05.13.094268.