UNIVERSITY OF CALIFORNIA SAN DIEGO

Spatial mapping of single cells in human cerebral cortex using DARTFISH: A highly multiplexed method for *in situ* quantification of targeted RNA transcripts

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

in

Bioengineering

by

Matthew Zhipong Cai

Committee in charge:

Professor Kun Zhang, Chair
Professor Jerold Chun
Professor Xiaohua Huang
Professor Bing Ren
Professor Sheng Zhong

2019

The Dissertation of Matthew Zhipong Cai is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California San Diego

2019

DEDICATION

For my wife Natalie, who stuck by me through all these years with unwavering support.
Thank you.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

DNA: Deoxyribonucleic acid

RNA: Ribonucleic acid

LNA: Locked nucleic acid

scRNA-seq: Single-cell RNA sequencing

NGS: Next generation sequencing

ISH: In situ hybridization

FISH: Fluorescence in situ hybridization

IHC: Immunohistochemistry

RCA: Rolling circle amplification

Rolony: Rolling circle amplified “colony”

smFISH: Single-molecule RNA fluorescence in situ hybridization

HCR: Hybridization chain reaction

bDNA: Branched DNA

BS(PEG)9: PEGylated bis(sulfosuccinimidyl)suberate

MDA: Multiple displacement amplification

PCR: Polymerase chain reaction

PE: Paired-end

UMI: Unique molecular index

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank everyone who supported me throughout my PhD. Without their guidance, encouragement, and support I would not be here today completing my dissertation.

In particular I want to thank my advisor, Dr. Kun Zhang, for taking me into his lab as an undergraduate student. He has always been patient and encouraging no matter which direction I was heading. He has fostered an environment that is easy to work in and where lab members can pursue bold ideas. His support throughout the years has been invaluable to me.

I would also like to thank everyone in the Zhang lab who helped me with my experiments: Ho Suk Lee for starting DARTFISH and teaching me the ropes; Dinh Diep for her contributions in probe design; Richard Que for partnering with me in the last year and sharing tech development responsibilities; and Kian Kalhor for his computational expertise and helpful inquisitiveness. I was also lucky to have some wonderful undergraduates to mentor: Justin Dang and Kefei Gao.

Finally, I would like to give a special thanks to Dr. Xiaohua Huang, who taught the class in my third year of undergrad that sparked my interest in next generation sequencing technologies that led me to this point.

Chapter 1 is coauthored with Lee, Ho Suk; Dang, Justin; Gao, Kefei; Yung, Yun; Kennedy, Grace; and Zhang, Kun. The dissertation author was the primary author of this chapter.

Chapter 2 is coauthored with Wu, Yan; Kalhor, Kian; and Zhang, Kun. The dissertation author was the primary author of this chapter.

Chapter 3 is coauthored with Diep, Dinh; Que, Richard; and Zhang, Kun. The dissertation author was the primary author of this chapter.

VITA

2013 Bachelor of Science, University of California San Diego

2019 Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Bioengineering

Professor Kun Zhang

ABSTRACT OF THE DISSERTATION

Spatial mapping of single cells in human cerebral cortex using DARTFISH:

A highly multiplexed method for *in situ* quantification of targeted RNA transcripts

by

Matthew Zhipong Cai

Doctor of Philosophy in Bioengineering

University of California San Diego, 2019

Professor Kun Zhang, Chair

The advent of high throughput single cell genomic technologies has revolutionized the study of cell biology. It has enabled scientists to discover rare cell types that were hidden in gene expression measurements of bulk cell populations. This led to many discoveries in complex tissues made up of heterogeneous cell populations, notably the mammalian brain. However, cells function in coordination with their environment and neighboring cells. Because these high throughput single cell technologies dissociate the cells from their native tissue, the spatial context is lost. *In situ* methods that examine cells in fixed tissue have existed for decades and are used routinely by doctors to diagnose diseases. But those traditional *in situ* methods do not have the capability to measure the expression of more than a handful of genes necessary to correlate with single cell. Presented here is one *in situ* approach for highly multiplexed RNA quantification that is also the first to be successfully used in human cortical sections, to the best of our knowledge.

The first chapter of my dissertation covers the development of DARTFISH, a method that enables highly multiplexed *in situ* digital quantification of targeted RNA transcripts in fresh frozen tissue.

The second chapter describes efforts to map cell types identified by single-cell or single-nuclei RNA sequencing to spatially defined cells from DARTFISH cortical sections.

The third and final chapter details ongoing improvements to DARTFISH to achieve better cell type classification of single cells in DARTFISH.

INTRODUCTION

The direct translation of biological discoveries to improving human health is a great incentive for studying living organisms. The more we understand healthy and diseased organisms, the better we can invent solutions to improve our own lives. In order to do so, we must have a complete description of cells, the basic unit of life. Cell populations are heterogeneous, containing multiple distinct cell types, and each cell can be defined by its molecular profiles, morphology, location, functional properties, etc. It is the goal of The Human Cell Atlas (HCA) to define the human cell types by these traits in a comprehensive reference. Much like the periodic table of elements is a fundamental tool in chemistry, the HCA will provide a framework for studying biology (Regev et al., 2017).

The monumental effort to catalog all cell types is only possible with the development of new technologies. History has shown that discoveries in cell biology have always been driven by technology; for example Robert Hooke observed cells for the first time in a cork slice when he viewed it under the microscope he invented. Recently, use of single-cell RNA sequencing (scRNA-seq) has exploded for two reasons: next generation sequencing (NGS) has greatly increased data collection throughput and significantly decreased the cost of sequencing DNA and RNA (Shendure & Ji, 2008); and advanced experimental techniques for isolating and lysing single cells as well as reverse transcribing and amplifying transcripts have made sequencing the picogram scale inputs possible (Trapnell, 2015). These technologies combined with the recognition that the transcriptome is a mediator of cellular phenotypes (Kim & Eberwine, 2010) has made scRNA-seq a powerful tool for defining cell types in a large population of cells based on transcriptomic state (Klein et al., 2015; Lake et al., 2016; Macosko et al., 2015; Zheng et al., 2017).

Cell type discovery and representation by scRNA-seq is important but a comprehensive description of cells must include more than just a transcriptome profile for each cell type. Other cell characteristics provide very useful information to differentiate cell types, are indicators of disease, and lead to a better understanding of cell biology. For example, chromatin accessibility profiles can be used to define cell types and can also shed light on transcriptional regulation when linked with transcription profiles (Cao et al., 2018; Chen, Lake, & Zhang, 2019). Other cell characteristics such as the connections between cells cannot be measured in isolated cells but require information about the cell in its native environment. MAPseq uses barcoded viruses and NGS to map where single neuron processes connect to, hopefully one day providing insight on how different cell types form the circuitry of the brain (Kebschull et al., 2016).

A more direct way to study cells in their tissue context, aka *in situ*, is fixed on a glass surface under a microscope. This is not novel today; histology has been a research discipline since scientists began examining dye-stained biological specimens under a microscope hundreds of years ago (Y. Wang, 2018). Early histologists used common dyes like indigo to study the structure of plant and animal tissue. Soon after, advancements in chemicals for tissue preservation, dyes for staining, and microscopes for observing led to many new histology techniques that began to have use for diagnostic pathology . There were stains for glycogen, (involved in many cancers), amyloid fibrils (indication of amyloidosis), myelin (demyelination is a sign of neuron degeneration), just to name a few (Titford, 2009). Covering all the various histology techniques is beyond the scope of this dissertation but the more recent developments in fluorescence *in situ* hybridization (FISH) are highly relevant and will be explored in depth.

In situ hybridization (ISH) and FISH, which uses fluorescence microscopy, are techniques for localization and quantification of specific DNA or RNA segments in a histologic

specimen using nucleic acid probes. Any ISH assay must have three components: probes, a label on the probes, and a way to detect the label. Developments in all three areas have led to ISH methods that are faster, more sensitive, higher spatial resolution, and have more flexibility of targets.

The probes are always single-stranded oligonucleotides that take advantage of the complementary base-pairing nature of nucleic acids for their target specificity. ISH probes can vary in length, number, and how they are produced. Initially, probes were cellular DNA and RNA that could be purified and selected from cultured cells. The first published experiments used rRNA and RNA transcribed from microsatellite DNA largely because there were no other probes available. Despite these limitations, the findings were impactful towards understanding chromosomal organization in nuclei (Gall, 2016). When gene cloning was established, ISH probes that targeted specific genes began to be used, however, these probes still had off-target activity due to repetitive sequences in the probes (Sealey, Whittaker, & Southern, 1985). With the reference genome created by efforts like the Human Genome Project, ISH probes could be strategically designed with bioinformatics tools to target specific sequences on chromosomes or RNA. In addition, with the increased ease and affordability of oligonucleotide synthesis on DNA microarrays, many probes can be used to tile a single chromosome or transcript, making detection of single molecules possible. The design of these probes is not trivial and there is continual development of computational tools for designing ISH probes (Beliveau et al., 2018). Finally, to address thermodynamic limitations of DNA and RNA, DNA analogues like locked nucleic acids (LNA) can be synthesized to make probes with higher specificity and melting temperature than DNA probes of the same length (Fontenete et al., 2016).

The first tags used to label probes were ^{3}H uridine that was incorporated by *in vivo* or *in vitro* transcription (Gall, 2016) and ^{32}P deoxynucleoside triphosphate that was incorporated by nick translation with DNA polymerase I (Rigby, Dieckmann, Rhodes, & Berg, 1977). These tags were able to be detected directly by autoradiography but the signal-to-noise ratio was low making it difficult to detect targets with low copy numbers (Jilbert, Burrell, Gowans, & Rowland, 1986). Haptens such as biotin and digoxigenin are used as tags that are detected indirectly by proteins conjugated to reporter molecules. One advantage of these indirect systems is that they could amplify the signal thus increasing sensitivity. It is worth mentioning other reporter molecules besides radioisotopes and fluorophores exist, namely enzymes such as horseradish peroxidase (HRP) and alkaline phosphatase (AP), which react with chromogenic substrates that precipitate into colored product. Chromogenic dyes are very convenient for pathologists because they can be viewed simply under a standard light microscope but are capable of less multiplexing and lower resolution (Gupta, Middleton, Whitaker, & Abrams, 2003).

As fluorescence microscopy became more accessible, FISH became the gold-standard for ISH assays (Gall, 2016). Using spectrally distinct fluorophores, multiple targets can be visualized in the same sample at higher resolution. Fluorophores can either be conjugated directly to nucleotides on the probe or to antibodies that bind to haptens on the nucleotide. Using a secondary protein to detect haptens amplifies the signal, which was necessary for improving detection sensitivity before incorporating fluorophore-labeled nucleotides into probes was efficient enough for direct detection (Huber, Voith von Voithenberg, & Kaigala, 2018). While fluorescence microscopy developed, fluorophore chemistry also improved with the invention of fluorophores that are brighter, more spectrally distinct, more stable, and that

come in a variety of spectral positions (Dempsey, Vaughan, Chen, Bates, & Zhuang, 2011). These technological advances in fluorescence microscopy enabled more powerful FISH techniques to be used.

Early RNA FISH was limited to high abundance targets in order for the fluorescence signal to be detected above background autofluorescence in the cell. But with the recent availability of synthetic oligo pools combined with the improved probe labeling and fluorophore chemistry, individual molecules of RNA can be detected and counted in fixed cells (Raj, van den Bogaard, Rifkin, van Oudenaarden, & Tyagi, 2008). This single-molecule RNA FISH (smFISH) method uses 48 singly labeled 20-mer probes that tile across the target to reliably generate uniform intensity diffraction-limited spots. Previous approaches for single molecule detection used a few multi-labeled 50-mer probes (Femino, Fay, Fogarty, & Singer, 1998), which meant one or two probes binding off-target could result in a false positive or a target missing one probe could become a false negative. The introduction of smFISH was a huge step for spatial transcriptomics. It has higher sensitivity than scRNA-seq (>95% detection rate) while also retaining the spatial information of the RNA and cell (Levesque & Raj, 2013). However, there were still two limitations preventing it from being used to create a human cell atlas: the number of targeted genes is limited to the number of spectrally distinct fluorescence channels available and the signal is not bright enough to overcome the higher autofluorescence typical of tissues.

This work aims to address the limitations of scRNA-seq by demonstrating a new technique that is capable of mapping cells onto a spatial map of a tissue section. In order to achieve this, it must be highly multiplexed for broad cell type classification, have signal amplification for use in tissue, and have a fast imaging protocol for higher throughput. The first

chapter will cover the development of this technique, dubbed DARTFISH for Decoding Amplified taRgeted Transcripts with Fluorescence in situ Hybridization. The second chapter will demonstrate proof-of-concept by mapping cell types from scRNA-seq data to the human and mouse cerebral cortex. Lastly, the final chapter will report ongoing improvements to DARTFISH that are in response to recent developments in the spatial transcriptomics field.

CHAPTER 1. DEVELOPMENT OF A HIGHLY MULTIPLXED IN SITU RNA METHOD

1.1. Abstract

Technology to analyze gene expression of a cell population with single-cell resolution and localization is critical for understanding the heterogeneity of structured tissues such as human brain and tumors. Advancements in single-cell sequencing allow the full transcriptome of isolated cells to be profiled, but it neglects the native spatial context of the cell. While several methods have been reported, in situ RNA mapping in post-mortem human specimens, remains to be demonstrated. We have developed a highly multiplexed method, called DARTFISH, for mapping of an arbitrary subset of RNA transcripts *in situ*. To achieve robust detection in complex tissue samples that has significant background autofluorescence, DARTFISH adopted an *in situ* cDNA transcription and rolling circle amplification strategy similar to FISSEQ. We leverage the multiplex capability and high specificity of padlock probes to capture thousands of targets *in situ* and include a hybridization-based combinatorial barcode scheme that allows amplicons to be decoded with quick reaction kinetics and under isothermal conditions.

As a proof of concept we performed DARTFISH on human culture fibroblast cells and cortical sections from mouse and human post-mortem brains. With a probe set targeting 240 genes we detected 800 amplicons per cell in fibroblast monolayer and 140 amplicons per cell in 10µm human cortical sections. In a 0.6mm2 cortical section, we decoded 27,812 amplicons representing 235 of the 240 genes. We also demonstrated that amplicon copy number can be used to quantify transcript abundance.

1.2. Introduction

Organs and tissues contain multiple cell types structurally organized in a way to carry out specific functions. High-throughput RNA sequencing of dissociated single cells has shown

that cells can be classified by their transcriptome profile (Lake et al., 2016) but to understand how cell types contribute to tissue function requires spatial information . Therefore there is a need for a method that can classify cells while retaining the spatial context of each cell.

One organ of particular interest is the human cerebral cortex due to its complexity, significance to our species, and how little we understand its pathologies. It also poses significant challenges such as high autofluorescence from lipofuscin, which accumulates with age, degraded RNA and tissue quality due to post-mortem tissue collection, and the complexity of cell types requiring many genes to differentiate between cell types.

There are a couple of methods that generate RNA sequencing libraries with spatial information encoded as sequences within the library. 'DNA microscopy' uses a mixture of unique DNA barcodes (UMI) that are added to fixed cells to tag RNA molecules. The UMI-tagged RNA is then amplified and concatenated to nearby UMI-tagged RNA as it diffuses such that the closer the proximity between two RNA molecules the more concatemers will be sequenced (Weinstein, Regev, & Zhang, 2019). While the idea is clever, the scale and orientation of the resulting spatial coordinates is only relative and cannot be integrated with *in situ* data like histological stains. 'Spatial transcriptomics' places a tissue section on an array containing reverse transcription (RT) primers with a positional barcode. RNA is then primed and reverse transcribed on the slide to create an NGS library such that every sequenced molecule can then be traced to a spot on the array (Ståhl et al., 2016). This method conveniently allows you to image the section using traditional microscopy techniques and map the transcripts to the tissue, but the array pattern and feature size of RT primers means the spatial resolution is not quite single-cell level.

The other approach in the field of spatial transcriptomics are *in situ* methods that chemically or enzymatically alter a tissue section and then use a microscope to localize and quantify the transcriptome. The last half century has seen these methods evolve from low resolution ISH to smFISH, which can detect single-molecules in cells using 40 fluorophores tiled along the target. However, in order for fluorescence signal to be detected above background in human cortical sections and other tissues, some method of signal amplification is required. As an aside, there is some work on reducing background rather than amplifying signal but retaining the RNA is not trivial (Chung et al., 2013; Sylwestrak, Rajasethupathy, Wright, Jaffe, & Deisseroth, 2016).

Signal amplification strategies for smFISH include DNA-based and enzyme-based methods. DNA-based methods use DNA oligos that assemble to each other on the target such that the site normally occupied by one fluorescently labeled probe can accommodate tens or hundreds. The two well-known methods are hybridization chain reaction (HCR) and branched DNA (bDNA) (Evanko, 2004; Sylwestrak et al., 2016). A newcomer called clampFISH combines click chemistry with multiple rounds of hybridization to achieve tunable signal gain and stably bound probes for applications like expansion microscopy (Rouhanifard et al., 2019). All of these methods use a more complex probe set limiting the ability to drastically multiplex the number of targeted genes.

Enzyme-based amplification involves rolling circle amplification (RCA) of circular single-stranded DNA target using Φ29 DNA polymerase and a primer. The resulting RCA colony (rolony) contains thousands of copies of the target in a submicron nanoball that FISH probes can hybridize to. ‘FISSEQ’ uses random reverse transcription (RT) primers to generate cDNA *in situ* and then circularizes the cDNA to serve as a template for RCA (Lee et al., 2014).

'*In situ* sequencing' (ISS) uses padlock probes that target gene-specific cDNA and the ligated padlock probes are the circular template for RCA (Ke et al., 2013). Because the only sequence dependence of RCA lies in the primer, this method can amplify all targets in one step as long as they contain a universal primer sequence. However, each enzymatic step such as RT, ligation, and RCA has less than 100% efficiency that overall leads to detection rates between 1 and 30% (X. Wang et al., 2018). FISSEQ and ISS also both use a sequencing-by-ligation approach instead of FISH to decipher the rolonies. This has the advantage of reading *de novo* sequences but the imaging process is much more time-consuming and costly.

The other important feature an *in situ* method must have to identify multiple cell types is being able to multiplex targets. Complex tissues like the cerebral cortex contain tens of different cell types, each defined by a combination of gene markers (Lake et al., 2016). Therefore determining the cell types in the same section requires a multiplexed assay that can detect tens of genes. The exact number of genes necessary depends on the number of cell types an experiment aims to resolve and how closely related the cells are. Pushing the limits of spectrally distinct fluorophores in a single FISH experiment reaches a maximum of seven before encountering problems of cross-talk between channels (Bhakdi & Thaicharoen, 2018). One possible solution to this is to label molecules with a combination of fluorophores in a spatial pattern that can be visualized through super-resolution microscopy (Lubeck & Cai, 2012). But the popular solution has been to do multiple rounds of fluorescence imaging in such a way that the color of each spot changes and after projecting across all rounds, the sequence of colors of each spot can be translated to its identity. The maximum number of genes able to be detected follows the equation $N = c^r$ where c is the number of colors and r is the number of rounds. In the seqFISH method, sequential rounds of smFISH are carried out on the sample

using 24 fluorescent probes hybridized to each transcript per round (Lubeck, Coskun, Zhiyentayev, Ahmad, & Cai, 2014). While simultaneously detecting hundreds of genes is possible this way, the probe set becomes very costly since each gene needs 24 times r fluorophore-labeled probes, and r increases with number of genes. MERFISH cleverly reduces the number of different expensive fluorophore-labeled probes that must be used by utilizing unlabeled 'encoding probes' as an intermediary. The encoding probes hybridize to the RNA targets and leave an unbound readout sequence that is hybridized by a fluorophore-labeled 'readout probe'. Many genes share the same readout sequence at each round, thus reducing the number of fluorescent probes needed (Moffitt et al., 2016). While both seqFISH and MERFISH can detect hundreds of genes with smFISH-like sensitivity, they also have limited smFISH-like signal-to-noise ratio.

To summarize the existing *in situ* methods, there are three vital characteristics that make an idea assay for spatial transcriptomics in tissue: number of genes detected, signal amplification, and detection rate. All the methods mentioned above only excel at two of the three (see Figure 1). In addition, to create a cell atlas of a whole tissue, data acquisition should be fast to feasibly process the hundreds or thousands of sections required. Given the existing methods, to spatially map cell types in tissue sections there is a need for an *in situ* RNA method that can be multiplexed to detect hundreds of genes simultaneously, has high signal amplification, detection rate, and throughput.

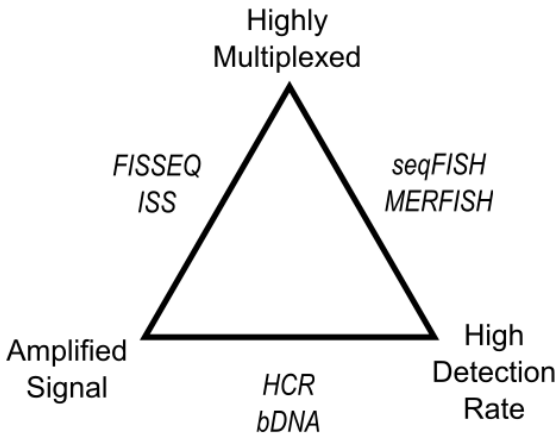


Figure 1: The iron triangle for *in situ* RNA method characteristics. The ideal method would have all three: high number of genes, signal amplification, and detection rate. Current methods require choosing two and sacrificing the third.

Our approach was to improve the detection rate and imaging throughput of enzyme-based amplification methods like FISSEQ and ISS. FISSEQ demonstrated RCA of circular templates to be robust and introduced cross-linking chemistry that fixed rolonies in cells and mouse brain sections. It was revolutionary because it did not target specific genes using probes making it like RNA-seq *in situ* (Lee et al., 2014). The corollary to that is many reads in FISSEQ were of ribosomal RNA (rRNA) as well as housekeeping genes, and not informative for classifying the cell. Since the sequencing is confined to the volume of the tissue and can't be dispersed across a flow cell, maximizing informative reads is vital. FISSEQ reports an average of ~2,800 reads per cell of which 82.7% are rRNA in cultured fibroblasts (Lee et al., 2014). We chose to use padlock probes, which can be highly multiplexed and specific *in situ* (Ke et al., 2013), that serve two purposes. The first is that by targeting informative genes there will be no volume in the tissue wasted. Second, it obviates the CircLigase step in FISSEQ that has <1% efficiency. Although ISS also uses padlock probes, we target hundreds of genes rather than tens

and we use a FISH-based barcode. FISH's compatibility with isothermal conditions and its fast kinetics makes the microscope 'decoding' simpler and quicker than ISS' sequencing. The method presented here is called DARTFISH, for Decoding Amplified taRgeted Transcripts with FISH.

1.3. Method Development

1.3.1 DARTFISH Overview

We initially used the FISSEQ protocol to fabricate rolonies in cultured fibroblasts and tried to multiply the number of targeted gene rolonies by using padlock probes that hybridize to the 'primary' rolonies. Theoretically every primary rolony contains thousands of targets so that even padlock probes with 1% efficiency will lead to a 10-fold increase in number of rolonies. However our experiments showed no improvement in rolony counts (see Figure S1). The finalized protocol uses padlock probes targeting cDNA and is outlined in Figure 2.

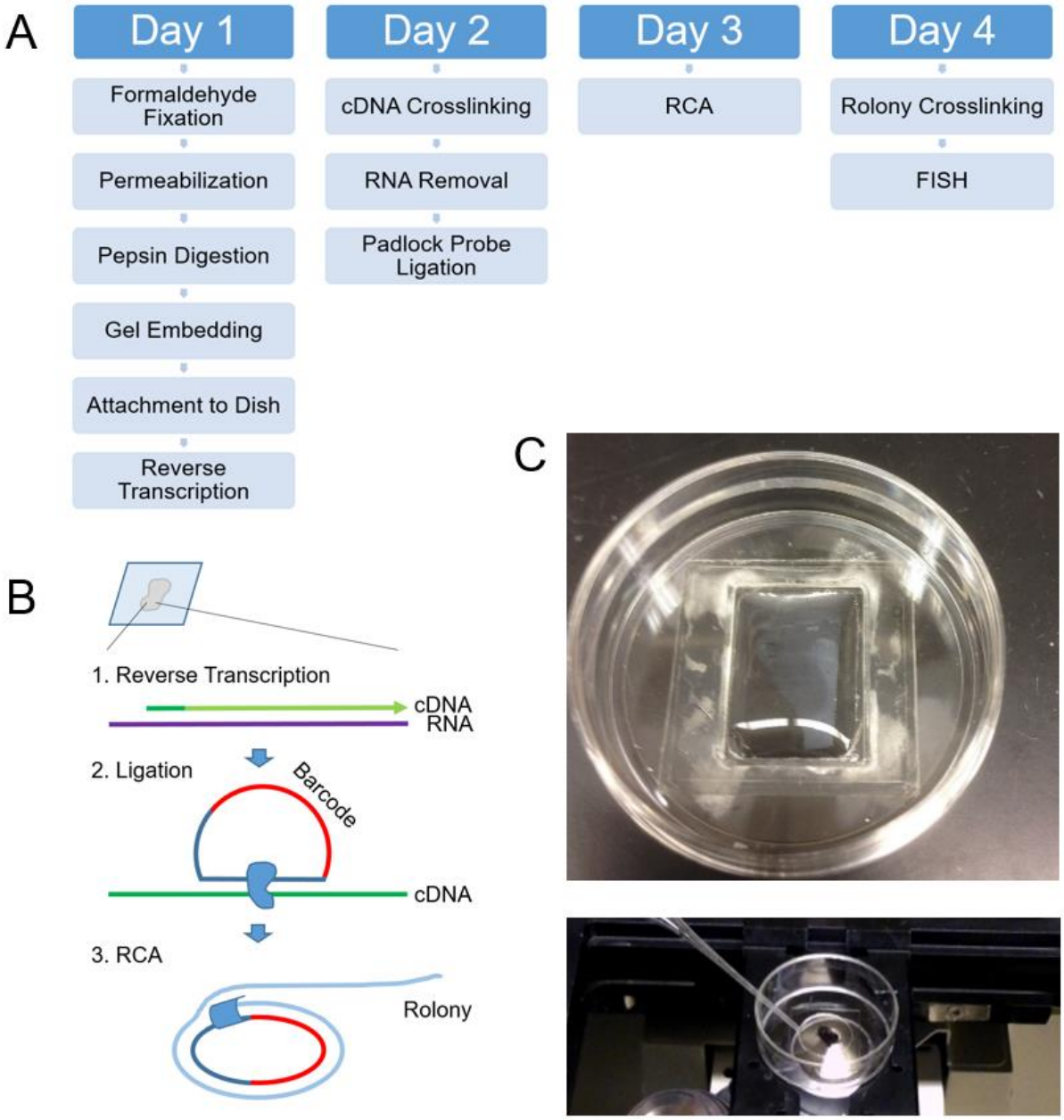


Figure 2: An overview of DARTFISH. (A) Outline of the protocol steps starting with a fresh frozen tissue section on a coverslip. If the sample is a cultured cell monolayer in a glass-bottom dish then gel embedding and attachment to dish is skipped. (B) Depiction of the three main enzymatic reactions that fabricate rolonies from RNA *in situ*. Maximizing the efficiency of each reaction is critical for increasing the detection rate. (C) The sample in a custom made glass-bottom dish with 12x17 mm hole. The dish is a format that can easily be stained with FISH probes and washed by manually pipetting on the microscope stage.

The sample, either a 10 μm fresh frozen tissue section or a monolayer of cultured cells, must be on a #1.5 glass coverslip to be compatible with the high resolution microscopy necessary to resolve rolonies. The tissue section is placed on a VECTABOND (Vector Labs) treated coverslip embedded in a hydrogel to prevent degradation during high temperature incubations and the coverslip is mounted to a petri dish to create a glass-bottom dish.

Fixation and permeabilization conditions must be optimized for each sample and some general guidelines can be found here (Lee et al., 2015). The general principle is to fix enough to retain the RNA without over-fixing the tissue, making it inaccessible to probes and enzymes. Permeabilization is to open the cell membrane but too much and the tissue will disintegrate. Finally, pepsin digestion is critical for removing RNA-binding proteins that disrupt RT. However precise control of pepsin concentration and reaction time is necessary to prevent the tissue from being completely digested.

SuperScript IV Reverse Transcriptase (ThermoFisher), an M-MLV mutant, reverse transcribes *in situ* RNA while incorporating free aminoallyl-dUTP (aa-dUTP) into cDNA, which is then covalently cross-linked by BS(PEG)9. The RNA is removed from the DNA:RNA hybrid so that padlock probes can hybridize to target sequences on the cDNA. RCA with Φ29 and the circularized padlock probes synthesizes rolonies that also have aa-dUTP to be cross-linked in the tissue.

The padlock probes are designed with two target-complementary sequences at the 3' and 5' ends meant to hybridize adjacently on the cDNA and become ligated by Ampligase (Lucigen). In between the hybridizing ends is a common linker sequence for the RCA primer and a hybridization-based barcode. The barcode is associated with the target gene such that decoding the barcodes on rolonies reveals the location of transcripts within the tissue.

Decoding the cross-linked rolonies is a process of sequential rounds of FISH using decoding probes. The hybridization protocol can be done at room temperature on the microscope stage simply by pipetting 0.5 **μ**M of each decoding probe in a 2X SSC and 30% formamide buffer onto the sample and incubating for 10 minutes. Stripping the decoding probes is done by adding 80% formamide in 2X SSC buffer. Imaging can be done on any fluorescence microscope with an objective capable of resolving 0.5 **μ**m-diameter features. We like to use a laser scanning confocal because it generates 3D image stacks and reduces out-of-focus background.

Images are analyzed using an in-house MATLAB package but can also be analyzed using Starfish, a python package developed by the SpaceTx consortium to analyze data from image-based transcriptomics methods like DARTFISH. We worked closely with Starfish developers so the two pipelines are very similar. The general principle is to align the images across all imaging rounds and then examine the intensities of each pixel across all channels and rounds. The intensities form a barcode and if adjacent pixels all share the same barcode forming a spot of the expected size, then we can confidently identify the spot as a rolony.

1.3.2 Padlock probe design

The first step of designing a DARTFISH experiment is to design padlock probes that hybridize and ligate with high specificity to genes of interest. At both ends of a padlock probe are hybridization ends, labeled H1 and H2 in Figure 3A, which have a sequence complementary to the target. The 5’ end must be phosphorylated in order for the 3’ and 5’ ends to be ligated, forming a circular ssDNA. The high specificity of padlock probes is attributed to the requirements of both ends hybridizing and a perfect sequence match at the ligation site in order for circularization to occur. The hybridization ends are designed to have a melting temperature

(Tm) of 55 °C to 65 °C so they are compatible with the thermostable DNA ligase, Ampligase, at 55 °C. The higher reaction temperature reduces non-specific hybridization that occurs between short complementary sequences at room temperature.

Selecting the 40-50bp sequence on the gene for the padlock probe to hybridize is not trivial. We chose to target all exons and no intronic or untranslated regions, but one could design padlock probes to target introns if so desired. We did a comparison of padlock probes targeting all exons versus only constitutive exons and found that targeting all exons had better results (see Figure S1). To design the actual hybridization sequences, we use ppDesigner, a program developed in our lab that (Diep et al., 2012). The genome and coordinates of the selected exons are fed into ppDesigner, along with desired parameters such as Tm and length of probes, and ppDesigner outputs the best hybridization sequences based on a trained neural network that predicts probe efficiency. These sequences are then filtered for specificity by mapping to the genome and transcriptome and then removing sequences that have multiple alignments.

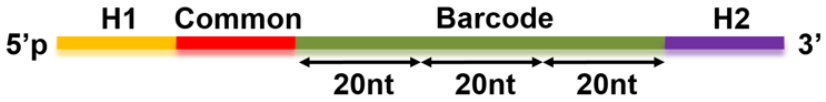
Connecting the two ends is a common linker sequence universal to all padlock probes and serves as the primer site for RCA. On rolonies it can also be hybridized by a FISH probe to detect all rolonies. Also on the backbone of the padlock probe is a hybridization-based barcode that can be read by sequential rounds of FISH. The barcode consists of three to five 20-nucleotide sequences depending on the number of rounds of sequencing. The length is limited by oligonucleotide synthesis technology but as the error rate improves, the potential barcode length of DARTFISH can as well.

The barcode and error-checking decoding scheme is an adaptation of Illumina's BeadArray. It was developed to decode the identity of randomly ordered DNA-linked beads on an array, which could achieve higher feature-density than an ink-jet printed microarray

(Gunderson et al., 2004). This problem is very similar to decoding the identity of rolonies in cells. The 20-nucleotide sequences used in the barcode and decoding FISH probes are shortened versions of the 22-nucleotide sequences designed to have minimal cross-complementarity, similar GC content and Tm, no runs of a single base longer than five, and low similarity to human genomic sequences.

The decoding scheme includes error-checking, shown in Figure 3B, that accounts for the different probabilities of each error type. In our version, every round has three possible ON values, corresponding to three fluorophores, and an OFF state. To incorporate an OFF state into a rolony, simply omit the 20-nucleotide sequences for that round. Since the most common error type is a misclassification of an ON as an OFF, we can use a checksum that equals the number of ON states detected in the barcode. The most likely error where one of the rounds transitions from an ON state to OFF state would result in a checksum of one less than expected. The less likely error where an OFF state is misread as an ON state would result in a checksum of one greater than expected. Only if both error types occur concurrently would the rolony be misclassified. To implement this, our decoding scheme uses barcodes that have two OFF states, represented by zeroes in the barcode, and at least two different ON state values, represented by one, two, or three.

A



B

Signal Intensity	Barcode	Checksum	Result
	130220	4	Used Valid
	100220	3	Invalid
	131220	5	Invalid
	131020	4	Unused Valid

C

$$\#\ \text{Unique Barcodes} = (3^{r-2} - 3) \times \binom{r}{2}$$

Rounds	Unique Barcodes
7	5,040
6	1,170
5	240
4	36

Figure 3: Padlock probe design. (A) Padlock probes for DARTFISH contain a common linker sequence for the RCA primer and a barcode for identifying the probe via FISH-based decoding. (B) The signal of each channel (color) is measured for every round and a threshold on the brightest channel for every round is used to generate a barcode. The barcode utilizes an error-checking strategy that requires the checksum, the number of rounds greater than the threshold, to be exactly two less than the number of rounds. Barcodes that meet the checksum requirement are considered valid, but may be unused for that experiment. The number of unused valid barcodes is indicative of the misclassification rate. (C) The number of unique barcodes scales exponentially with the number of rounds of imaging, making the simultaneous detection of hundreds of genes possible in just five rounds of FISH.

1.3.3 Padlock probe production

A differentiator between DARTFISH and ISS is that DARTFISH uses thousands of ssDNA probes >120bp making it too infeasible to order each probe individually. Instead, an oligo pool synthesized in parallel on a DNA microarray was used (Agilent Technologies, CustomArray, or Twist Bioscience). The low yield of an oligo pool required PCR amplification to produce enough DNA for a DARTFISH run and is described below.

The oligo pool was first amplified by PCR in a 100 **µ**l reaction with 0.1 nM template oligonucleotides, 400 nM each of AP1V4IU and AP2V4 primers, and 50 **µ**l KAPA SYBR fast Universal qPCR Master Mix at 95 °C for 30 s, 20 cycles of 95 °C for 30 s; 55 °C for 45 s; and 72 °C for 45 s, and 72 °C for 2 min. The amplicon products were purified with Qiaquick PCR purification columns and then re-PCR'd in 96 x 100 **µ**l reactions. The conditions for the second round of PCR were the same except starting with 0.02 nM template and consisting of only 12 cycles. The resulting product volume was reduced by ethanol precipitation and then purified with Qiaquick PCR purification columns.

In order to make the PCR amplicons single-stranded, <200 **µ**g of amplicon was digested with 240 U of λ Exonuclease (New England BioLabs) in 1X λ Exonuclease Reaction Buffer for 2 hr at 37 °C. The ssDNA product was then purified using Zymo ssDNA/RNA Clean & Concentrator columns. To remove one of the flanking sequences used for PCR, the ssDNA was digested with 25 U USER (New England BioLabs) for 2.5 hr at 37 °C in 1X DpnII Buffer. To remove the other flanking sequence, 1 **µ**M RE-DpnII_V4 guide oligo was added, the sample heated to 94 °C for 2 min, followed by 37 °C for 3 min to anneal the guide oligo. Then 250 U DpnII (New England BioLabs) was added and the DNA was digested overnight at 37 °C in 1X

DpnII Buffer. The resulting DNA was then purified with Zymo ssDNA/RNA Clean & Concentrator columns.

The digested probe needs to be size selected to remove partially digested probes and background from synthesis errors of the oligo pool (see Figure S2). The probes were purified with 6% denaturing TB-urea PAGE and then ethanol precipitated to the highest possible concentration, typically ~ 2.5 **μ**M.

1.3.4 Hydrogel embedding of tissue section and fabrication of a glass-bottom dish

One issue we constantly faced during early phases of development was the tissue section sloughing off after the pepsin digestion step. It was especially noticeable after overnight incubation steps at elevated temperatures, likely due to reversal of formaldehyde cross-linking (David, Fowler, Cunningham, Mason, & O'Leary, 2011). Inspired by CLARITY and Expansion Microscopy, we developed a hydrogel embedding method that uses Acryloyl-X, SE (ThermoFisher) to create covalent bonds between primary amines of the tissue and the surrounding polyacrylamide gel to reinforce the structural integrity of the tissue. The gel is 5% T acrylamide cross-linked with 0.5% C bis-acrylamide, which keeps the pore size large enough for diffusion of polymerases, ligases, and 150bp ssDNA probes (see Figure 3). The thickness of the gel is 70 **μ**m at time of polymerization by clamping the section to another coverslip with a layer of polyimide tape (Kapton) as a spacer in between as shown in Figure 4A.

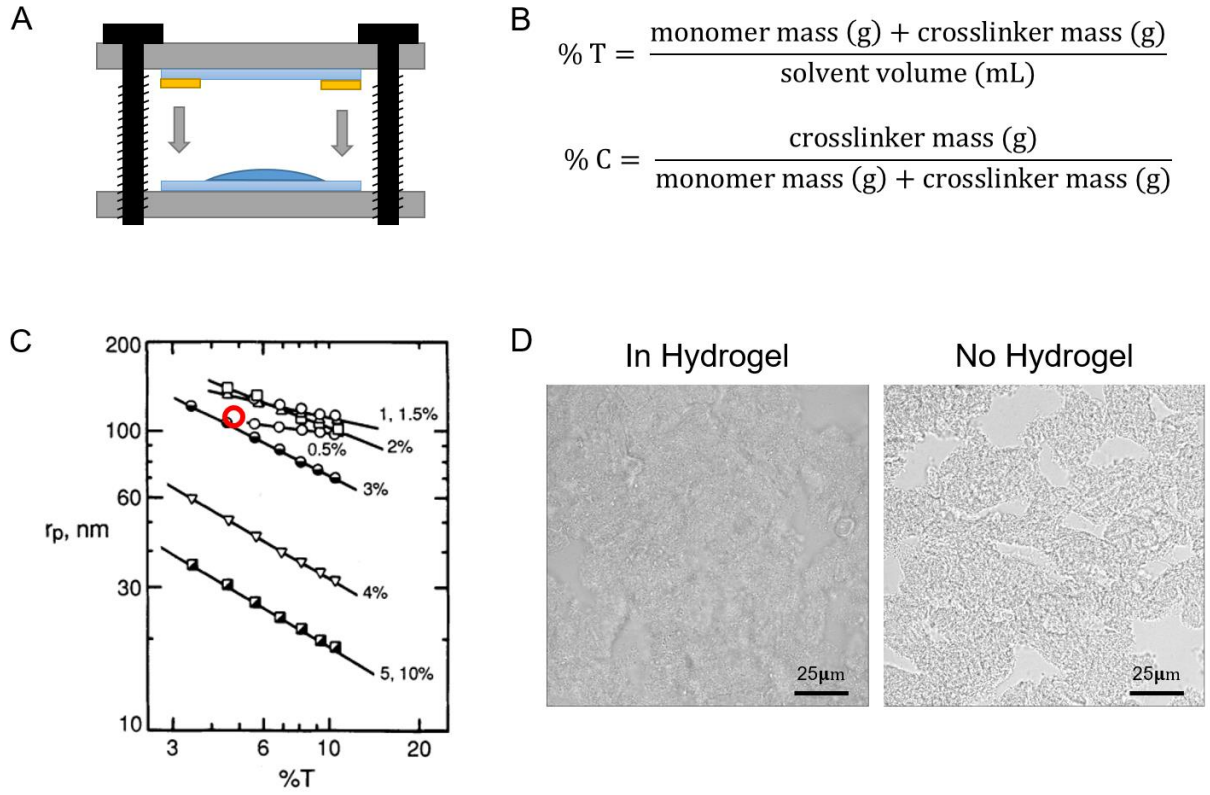


Figure 4: Design of hydrogel embedding technique to preserve tissue integrity. (A) Hydrogel polymerization clamp device to control gel thickness uses 70 **µ**m thick Kapton tape in yellow as a spacer around the tissue. (B) Formula for calculating standard polyacrylamide monomer and cross-linker concentrations used to prepare gel. (C) There is a linear relationship between pore size and %T of a polyacrylamide gel for a given %C (Holmes & Stellwagen, 1991). Extrapolating from the curve for 0.5% C, the hydrogel used for embedding has a pore radius of >100 nm, marked by the red circle. Padlock probes are approximately 50 nm in length and polymerases are approximately 20 nm in diameter. (D) Differential interference contrast microscopy of human cortical tissue shows significantly more degradation from DARTFISH processing without hydrogel embedding.

Another issue we faced when switching from fibroblasts cultured in a glass-bottom dish to tissue sections on a #1.5 coverslip is how to mount the coverslip to the bottom of a 35 mm cell culture dish with minimal risk of leakage. The solution we came up with was to use a laser cutter (LaserCAMM) to cut a 10mm circle in the center of a sterile 35 mm cell culture dish (Corning Falcon). Covering both sides of the dish with painter's tape prevents ablated polystyrene from leaving a cloudy residue on the surface as well as minimized the protrusion at the edge of the cut caused by melted polystyrene. Any protruding lip was shaved off to create a flat surface for the coverslip to adhere to. The ideal adhesive we found was ARcare 90106 (Adhesives Research), a clear medical grade double-sided pressure-sensitive tape used for *in vitro* diagnostic applications. We used a vinyl cutter (Roland) to cut appropriately shaped adhesives with a 10 mm hole. After embedding the tissue section in hydrogel, the edges of the coverslip were wiped clean and attached to the bottom of the dish using light pressure on the adhesive.

1.3.5 Selection of *in situ* reverse transcriptase and primer using NGS

The M-MuLV Reverse Transcriptase (Enzymatics) used by FISSEQ was not optimal for the purposes of DARTFISH. M-MuLV has ribonuclease H (RNase H) activity that cleaves the RNA as it polymerizes the cDNA strand. Paired with a random nonamer primer, it is expected that the same RNA transcript could be primed in multiple locations and produce multiple cDNA fragments. RNase H activity would lead to truncated cDNA, which may be functional for FISSEQ as CircLigase has lower efficiency for long oligonucleotides. However for DARTFISH, having less truncated cDNA product is better as it increases the number of potential targets for the padlock probes.

Comparing potential reverse transcriptases and primers is not a trivial experiment *in situ*. It cannot be done *in vitro* because the fixation of RNA within cells changes the conditions of the reaction. However quantifying RT efficiency by FISH is not practical. The natural difference in gene expression between different regions of heterogeneous tissue is a confounding variable that is difficult to control for when comparing rolony counts. Even in cultured cells there is variability, which requires sampling large areas to remove. Second, confocal microscopy with high numerical aperture (NA) objectives is not the ideal tool for imaging large areas. Finally, the numerous conditions to test requires handling multiple dishes, which is impractical during the rolony synthesis stage as well as imaging.

Here we devised a method for quantifying *in situ* RT efficiency that leverages the throughput of NGS to count. The protocol is exactly the same as DARTFISH through the RT step in order to keep all variables constant. But instead of cross-linking the cDNA to the tissue, the tissue is scraped off the coverslip with a sterile scalpel and placed in a PCR tube. The DNA is extracted using the Zymo Quick-DNA kit and then qPCR with technical triplicates is used to quantify. We used primers for 18S rRNA to quantify RT efficiency and primers for hLINE as a standard to normalize for number of cells in each sample (see Table 3 for sequences).

The reverse transcriptases we tested were M-MuLV (Enzymatics) from the FISSEQ protocol, RevertAid H Minus from BaristaSeq, and SuperScript IV. RevertAid H Minus and SuperScript IV have no RNase H activity and should yield more cDNA than M-MuLV. The primers we tried were a random nonamer (N9), a quasi-random primer from Sigma Aldrich's Transplex Whole Transcriptome Amplification kit (KN2), and a quasi-random LNA primer of our own design. The KN2 primer is 9 random G's or T's followed by 2 N's to create a primer that hybridizes randomly and is also less likely to form primer dimers, thus leading to more

priming for RT. However, having 9 K's limits the number of possible sites in the transcriptome so the LNA primer was designed. The LNA primer is a random hexamer of A's, T's, and +C's where +C is an LNA nucleotide. The LNA ribose is modified with a bridge connecting the 2' oxygen and 4' carbon that locks it into the 3'-endo conformation making hybridization more energetically favorable. Effectively, the Tm is higher so a shorter primer, which complements more potential sequences of the transcriptome, can hybridize at the same temperature.

Using this method, we quantified RT efficiency in human brain with different combinations of reverse transcriptases and primers and calculated each condition relative to M-MuLV Reverse Transcriptase paired with the random nonamer primer (see Table 1). We found SuperScript IV Reverse Transcriptase paired with a random nonamer primer had almost 15 fold more cDNA from 18S rRNA per cell.

Table 1: Relative Normalized RT Efficiency as measured by NGS

	M-MuLV	RevertAid H Minus	SuperScript IV
N9	1	N/A	14.7
KN2	1.9	2.33	4.05
LNA	0.76	N/A	N/A
N9 + LNA	1.39	N/A	4.13

1.3.6 Decoding by sequential hybridization

One of the strengths of DARTFISH is the quick and simple decoding process shown in Figure 5A that can be done by manually pipetting. The optimal hybridization buffer is 30% formamide in 2X SSC and the optimal stripping buffer is 80% formamide in 2X SSC (see Figure S3). Imaging is done on a Leica SP8 laser scanning confocal microscope with a 63X 1.4NA oil-immersion objective and lasers with wavelengths of 448, 552, and 647 nm. The higher

resolution of a high NA objective was found to increase the detection rate when compared to a 0.75NA objective (see Figure S4). The pixel size is set to 0.15 μm, which satisfies the Nyquist-Shannon sampling theorem for detecting a 0.5 μm rolony.

Each round, three decoding probes with fluorophores Alexa488, Cy3, and Cy5, are hybridized to rolonies with barcodes containing a complementary 20-nt sequence for that round. An important detail is that not every rolony will be hybridized by a probe in every round, the error-detection strategy purposely omits two rounds from every barcode as shown in Figure 5B. The sample is washed in 1 mL 2X SSC buffer twice before imaging in 2X SSC. Stripping is done by adding 80% formamide in 2X SSC, which disrupts the hydrogen bonding between probes and rolonies, and washing it away by repeating three times. 1X PBS is used to wash the sample twice before the next round of hybridization and as a storage buffer if the sample is to be kept at 4 °C.

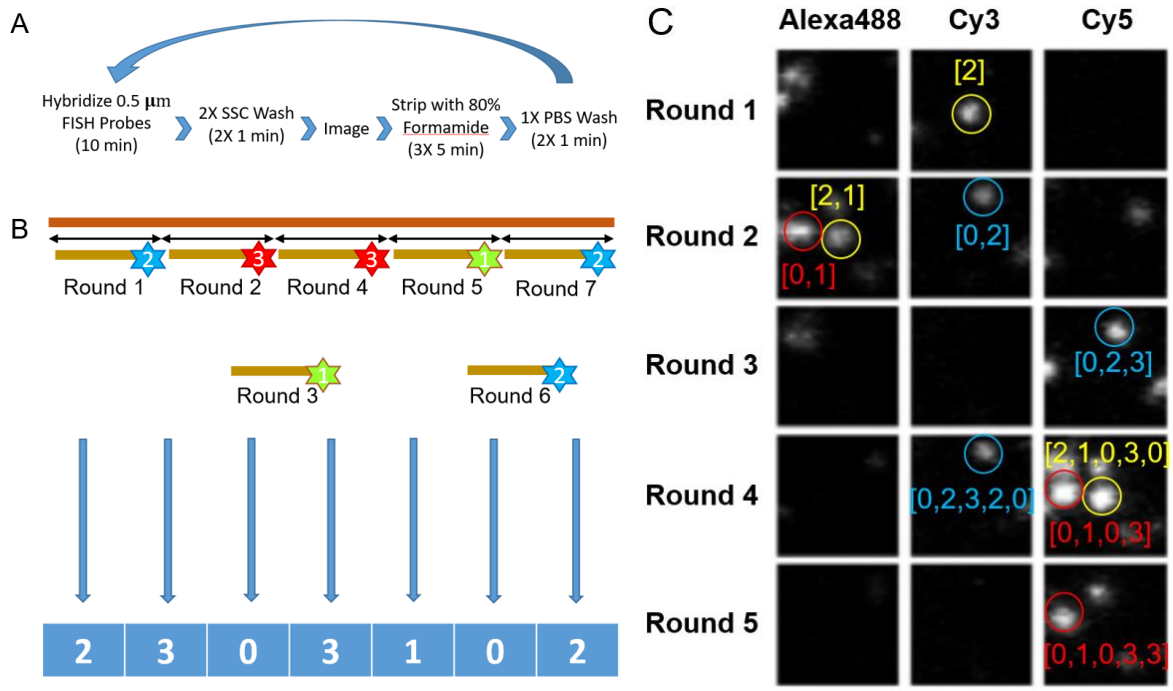


Figure 5: (A) Protocol for decoding rolonies with sequential rounds of FISH on a microscope stage. (B) Example of how a barcode on a rolony is decoded through 7 rounds of hybridization, imaging, and stripping. Only one of the three probes is shown for each round for illustration purposes. Rounds 3 and 6 were designed to be OFF states so there is no site for those probes to bind on the rolony. (C) Raw maximum intensity projected images for every channel of every round showing what rolonies look like under the microscope and how the decoding works.

1.3.7 Image analysis software

There are two algorithm paradigms for decoding a barcode or reading a sequence when the location of the features are not known *a priori*. The first is spot-based, where there is an image that is known to contain all features and image segmentation can identify the location and boundaries of all features. Then every feature is assigned a barcode based on the signal at that location for every round. The other approach is pixel-based, which assigns a barcode to every pixel. Adjacent pixels with the same barcode that also form a spot of the correct size is called a feature.

For DARTFISH and most other *in situ* RNA methods, pixel-based decoding can detect more features than spot-based decoding. Spot-based decoding relies on image segmentation that is inaccurate in areas of high rolony density or high background (see Figure S5). However, pixel-based decoding can have its flaws as well. Stochastic noise is more likely to be misinterpreted as a barcode and lead to false positives. Having an error-checking barcode scheme like DARTFISH can reduce that. The clustering step of neighboring pixels can also lead to inaccurate quantification if rolonies with the same barcode overlap. This issue can be addressed by restricting the size of a feature and applying a watershed transformation. Lastly, the pixel-based algorithm is more memory intensive since it has to find the barcode of every pixel instead of narrowing it down to just rolonies beforehand.

In DARTFISH, a FISH probe targeting the common linker sequence can hybridize to all rolonies to make it compatible with spot-based decoding but we use pixel-based decoding for the higher detection rate. An outline of the image processing and decoding algorithm for each field-of-view is shown in Figure 6A. After localizing rolonies in every field-of-view, the

field-of-view can be stitched together using their coordinates from the microscope or tile-stitching function in ImageJ.

The first step of decoding is to pre-process the images to improve data quality by removing noise and correcting for microscope stage drift. Since the confocal microscope produces image stacks in the format of multiple single-plane tiffs, a maximum intensity projection along the z-axis creates a single tiff image for every channel of every round. Then a Gaussian blur is applied with sigma of 100 nm to every image. A top-hat filter can also be used to remove spots larger than the expected rolony size (see Figure S6). Images from each round are then registered to a reference cycle using the SimpleITK package in python to find the affine transformation using the DIC images. Any round can be used as reference. If a round of imaging was done with no FISH probes to measure background autofluorescence, the background images are subtracted from every image in that channel.

After the images are pre-processed, the FISH images are decoded. Due to differences between channels including fluorophore quantum efficiency, laser, and PMT properties, the brightness of pixels varies among channels as shown in Figure 6B. Therefore the intensities are normalized by the maximum value in each image. Then for every round, OFF pixels are called for any pixel where none of the three channels has a value greater than 0.5 (see Figure S7). These pixels have ‘0’ at that digit of their barcode. For remaining pixels, the digit of the barcode is determined by the distance between a vector of the normalized intensities in 3-dimensional space to the nearest axis shown in Figure 6C.

With a digit called for every round, the barcode for each pixel is simply constructed by concatenating the rounds together. Invalid barcodes, those that don’t have two OFF states and at least two different ON state values, are discarded. Neighboring pixels of the same barcode

are then clustered and filtered by area and shape to remove noise. Finally a function like MATLAB's regionprops is used to get the location and area of every decoded rolony to create a table.

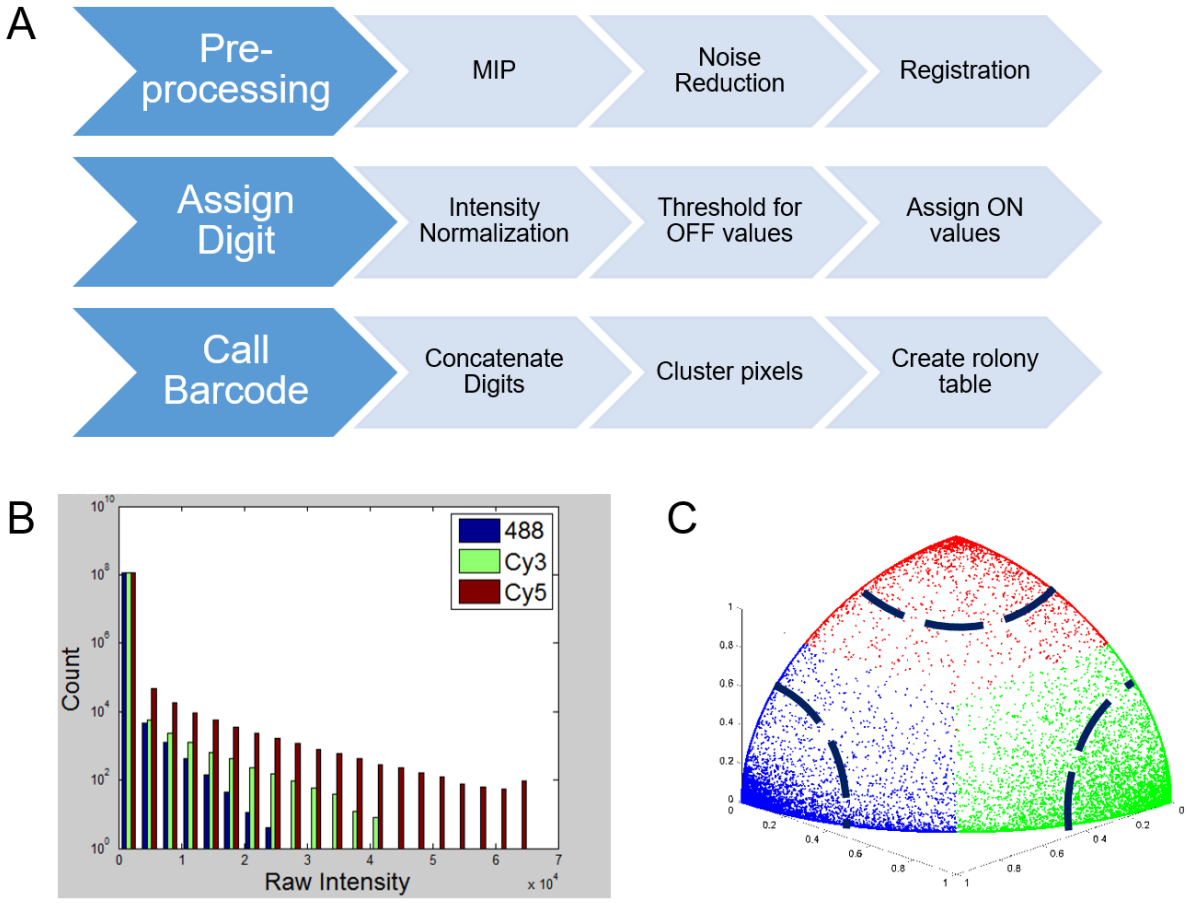


Figure 6: Hybridization-based decoding of rolony barcodes. (A) Protocol for decoding rolonies with sequential rounds of FISH on a microscope stage. (B) Histogram of pixel intensities of one round of images show how the different channels have different distributions and need to be normalized for an unbiased comparison. (C) Vector of channel intensities for each pixel plotted with threshold cut-offs 30° from the axes. (D) Example decoding of pre-processed images.

We also worked with the SpaceTx team at Chan Zuckerberg Initiative, who developed a Python library called Starfish that can be used to nearly replicate the MATLAB pipeline we created. The major difference is that the MATLAB pipeline processes the images of each round independently to call a digit before concatenating them to create a barcode. Starfish stores the intensity of every channel and round in a single N-dimensional vector and finds the nearest used barcode in that N-dimensional space. Despite this fundamental difference, they perform fairly similarly, suggesting they are both accurate. The advantage of using the Starfish package is the flexibility in trying different pipelines with relatively little effort (see Figure S8).

1.4 Results and Discussion

1.4.1 Validation in fibroblasts

The initial experiments used PGP1 fibroblasts cultured in glass-bottom petri dishes (MatTek). They are large, quick dividing, and have been previously used for FISSEQ studies. Cultured cells are also inherently easier than tissue to obtain FISH data because they have less background, are in a monolayer, and suffer no RNA degradation that occurs in tissue between the harvesting and fixation timeframe.

We started with one padlock probe targeting just MALAT1 to establish the optimal padlock probe hybridization and ligation conditions (see Figure S9). MALAT1, also known as Nuclear Enriched Abundant Transcript 2 or NEAT2, is a highly expressed non-coding RNA found in the nucleus. The high abundance, ubiquity, and nuclear specificity of MALAT1 transcripts makes it useful for evaluating different parameters like concentration and temperature of padlock probe capture. The drawback is that most other gene targets are expressed many orders of magnitude less than MALAT1 and the number of MALAT1 rolonies cannot be expected for other genes.

The first probe set was ordered as part of a 12,000 oligo pool from CustomArray that targets 240 genes determined to be the most differentially expressed in 1,000 human cortical nuclei. Between 5 and 30 probes target each gene with 3,500 probes total (see Figure S10). Padlock probes were designed and produced using methods outlined in Section 1.3. Another set was ordered in the same pool targeting 165 of the same genes with 2,500 probes but targeted only constitutive exons instead of contigs like the first set.

We established that the padlock probes could be used quantitatively by doing an experiment where we captured FirstChoice Human Brain Reference RNA (ThermoFisher) *in vitro*. The RNA was first reverse transcribed using ProtoScript First Strand cDNA Synthesis Kit (New England BioLabs) following the standard protocol with the randomized primer mix. Then we used the padlock probes to capture cDNA. In another tube we set up the same reaction for capturing human genomic DNA. The padlock probe capture protocol is the standard from our lab (Diep et al., 2012). Since padlock probes have very high variance in efficiency, we characterize efficiencies by capturing genomic DNA that should have a uniform number of every target so that the count of captured padlock probes reflects their efficiencies (see Figure 7B). Then to quantify the number of targets in cDNA we can normalize for the differences in efficiencies by dividing. When comparing the normalized padlock probe counts of each gene to the FPKM values from RNA-seq of the same reference RNA, we found a good Pearson's correlation of 0.87 shown in Figure 7A.

By comparing the two probe sets that share 165 of the same gene targets, we investigated possible biases our decoding protocol may have for certain barcodes (see Figure 7C). The two probe sets had barcodes randomly assigned to each gene and were used to fabricate rolonies in parallel PGP1 fibroblast samples. As expected, if there were no biases among barcodes, we

found no correlation between barcode counts of the two samples. We also verified that the result was not due to non-specificity in either sample by showing that the gene counts between the two still had high correlation. The Pearson's r of less than one can be explained by the small sample size, stochastic gene expression, and that the probe sets target different sequences of each gene.

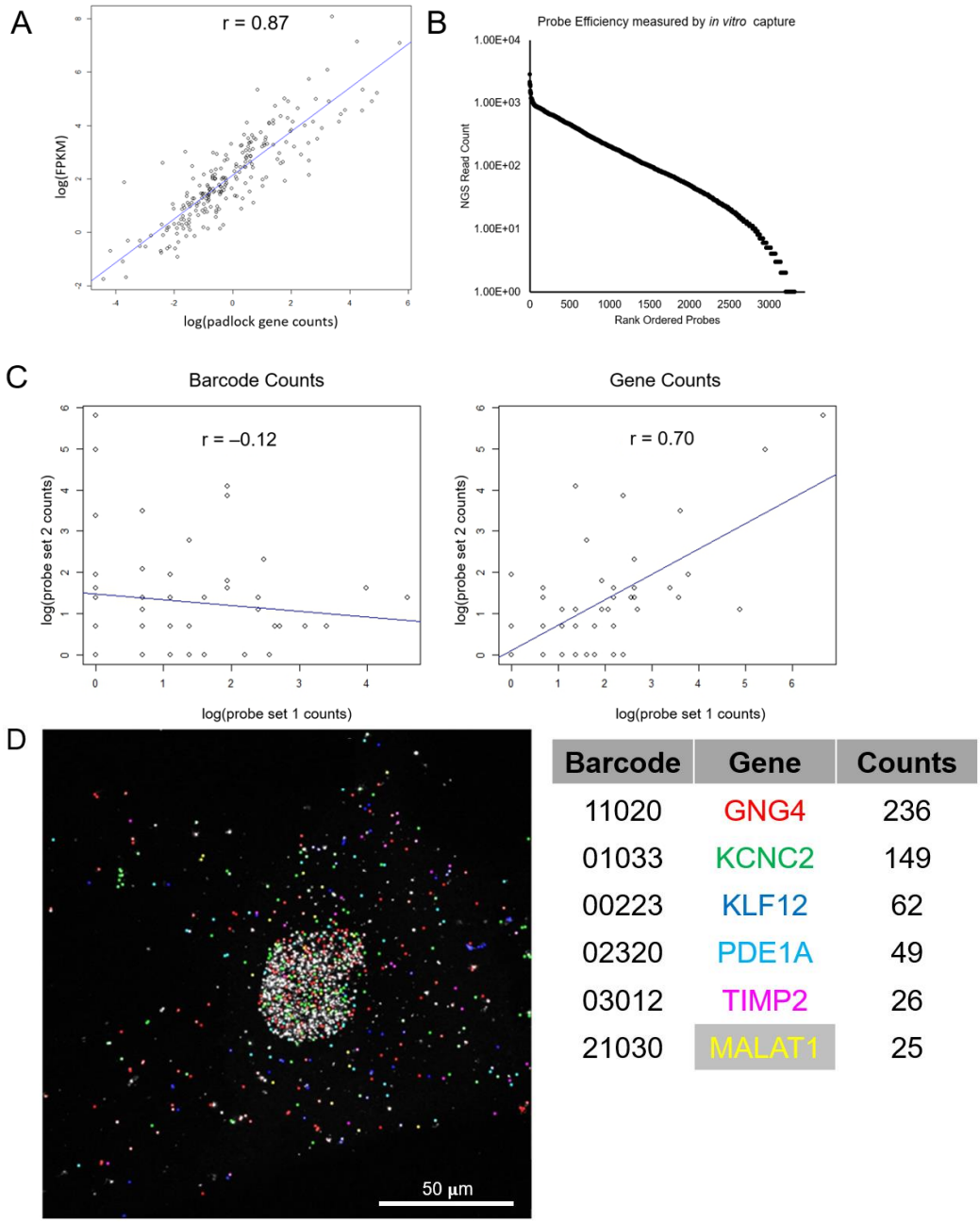


Figure 7: Padlock probes can be used to quantify target sequences *in vitro* and *in situ*. (A) By normalizing DARTFISH gene counts for padlock probe efficiency, we show at the bulk level that DARTFISH gene expression measurements in fibroblasts have high correlation with RNA sequencing. (B) Padlock probe efficiencies can vary over four orders of magnitude and is measured by capturing genomic DNA *in vitro* and sequencing. (C) Using two different probe sets targeting the same genes but with the barcodes shuffled, it is evident that there is no bias in terms of barcodes. There is no correlation between barcodes of each probe set but there is good correlation between genes of each probe set. (D) Example of rolonies decoded in a fibroblast. Padlock probes targeted exon contigs and no suppressor oligos were used.

In fibroblasts we detected an average of 70 genes per cell and 802 rolonies per cell without the use of suppressor oligos. For comparison, FISSEQ reports an average of 200 mRNA reads per cell in fibroblasts. Similar to findings in the FISSEQ publication, we noticed an enrichment of rolonies in the nuclei as can be seen in Figure 7D. This is despite the fact that the padlock probes target only exonic regions. A possible explanation is that the genes we chose to target were picked from an analysis of RNA sequenced from single-neurons. Alternatively, it could possibly be due to the *in situ* RT chemistry. In fibroblasts cultured at low cell density, the rolonies within the cell form a cloud that can be used for cell segmentation (see Figure S4).

Ordering thousands of probes as an oligo pool rather than thousands of individual oligonucleotides is cost effective and obviates the need to handle thousands of tubes. However, it makes modifications to the probe set tricky. Here we have also developed a way to suppress specific padlock probes in a probe set by using suppressor oligos. The suppressor oligos have complementary sequences to both hybridization arms of the padlock probe that are separated by two extra bases. Therefore padlock probes will hybridize to the suppressors but be unable to ligate. Through *in vitro* capture of genomic DNA and sequencing, we identified the highest efficiency probes were ones that targeted repetitive regions. By adding suppressor oligos we could drastically reduce the number of captured probes *in vitro* (see Figure S11). Five suppressor oligos for the genes KCNC2, GNG4, PDE1A, and PTPRK were used for all experiments shown in this section unless explicitly stated.

Utilizing a longer barcode with six rounds of decoding creates 1,170 valid barcodes that satisfy our error-detection scheme. By leaving a fraction of the valid barcodes unused, we can calculate the misclassification rate with the following formula where B is the number of decoded barcodes and C is the number of designed barcodes (Gunderson et al., 2004).

$$Misclassification\ Rate = \left[\frac{B_{unused}}{B_{used} + B_{unused}}\right] \times \left[\frac{(C_{used} - 1)}{C_{unused}}\right]$$

To measure the misclassification rate of our decoding algorithm, we designed a probe set that used only 391 out of 1,170 valid barcodes. We fabricated rolonies in fibroblasts and human brain tissue from Brodmann area 8 and achieved a misclassification rate of 1% and 4%, respectively. The higher misclassification rate in tissue is likely due to autofluorescence background that is sometimes punctate. This highlights the importance of an error-detection scheme. We also used spot-based decoding to inspect rolonies decoded to invalid barcodes and found that the majority of invalid barcodes assigned to rolonies were due to a single round transition from an ON state to OFF state or OFF state to ON state (see Figure S12). This confirms the assumptions of the error-detection scheme.

1.4.2 Validation in mouse brain and human brain

When moving onto experiments on mouse and human cortical sections, we had to establish the optimal fixation and permeabilization conditions. Since the experimental procedure for DARTFISH is a modification of FISSEQ, we used the density of FISSEQ rolonies as a metric to evaluate the fixation and permeabilization protocol (see Figure S13). The finalized fixation protocol was to immediately dry fresh frozen sections for 3 minutes on a 50 °C hot plate when removed from -80 °C followed by fixing in 4% formaldehyde buffered in 1X PBS at 37 °C for 15 minutes. The permeabilization was with 0.25% Triton X-100 in 2X SSPE at room temperature for 10 minutes. Pepsin digestion was with 0.01% pepsin in 0.1 N HCl at 37 °C for 90 seconds. Another key change we made was to use DEPC-treated buffers if we could not guarantee they were RNase-free.

For mouse brain we designed a probe set using genes curated by a consortium of scientists working on the SpaceTx project at CZI. These genes were picked for their high expression, many are canonical cell type markers in the mouse brain, and many more were computationally determined to have some cell type specificity in single-cell RNA-sequencing data. Ultimately, we targeted 478 genes using 7,133 probes designed similarly to the probes used in previous fibroblast experiments.

DARTFISH with this probe set was done in a coronal section of a C57BL/6 mouse harvested postnatal 3.5 months of age. Imaging covered the visual cortex over an area of 0.8 mm^2 with image stacks 10 **μ**m thick and optical sections every 0.3 **μ**m (see Figure 8A). Staining DNA with DRAQ5 (ThermoFisher) to segment nuclei, we counted 668 cells within the imaged volume. Starfish decoding localized 15.1 rolonies per cell. Within the total volume imaged, 31 genes had greater than 50 rolonies and 70 genes had greater than 25 rolonies. The rolony counts of top genes is shown in Figure 8B. To validate the specificity of DARTFISH, we compared the spatial expression of genes with high rolony counts to ISH images from the visual cortex of similarly aged mice available on the Allen Brain Atlas. As seen in Figure 8C there is very high concordance in gene expression. For example, there is a notable absence of Snap25, which encodes a t-SNARE protein, in Layer 1 of both images. In addition, Ptgds, which encodes Prostaglandin D synthase and is known to be expressed in glial cells and meninges, also shows very high expression in the pia mater of both images. Dysregulation in prostaglandin D synthase has been implicated in many neurological diseases including multiple sclerosis, Parkinson's disease, and schizophrenia (Harrington, Fonteh, Biringer, Hühmer, & Cowan, 2006). Given the high efficiency of padlock probes for Ptgds, they may be a good candidate for a study looking into the spatial gene expression of Ptgds in diseased and control mice.

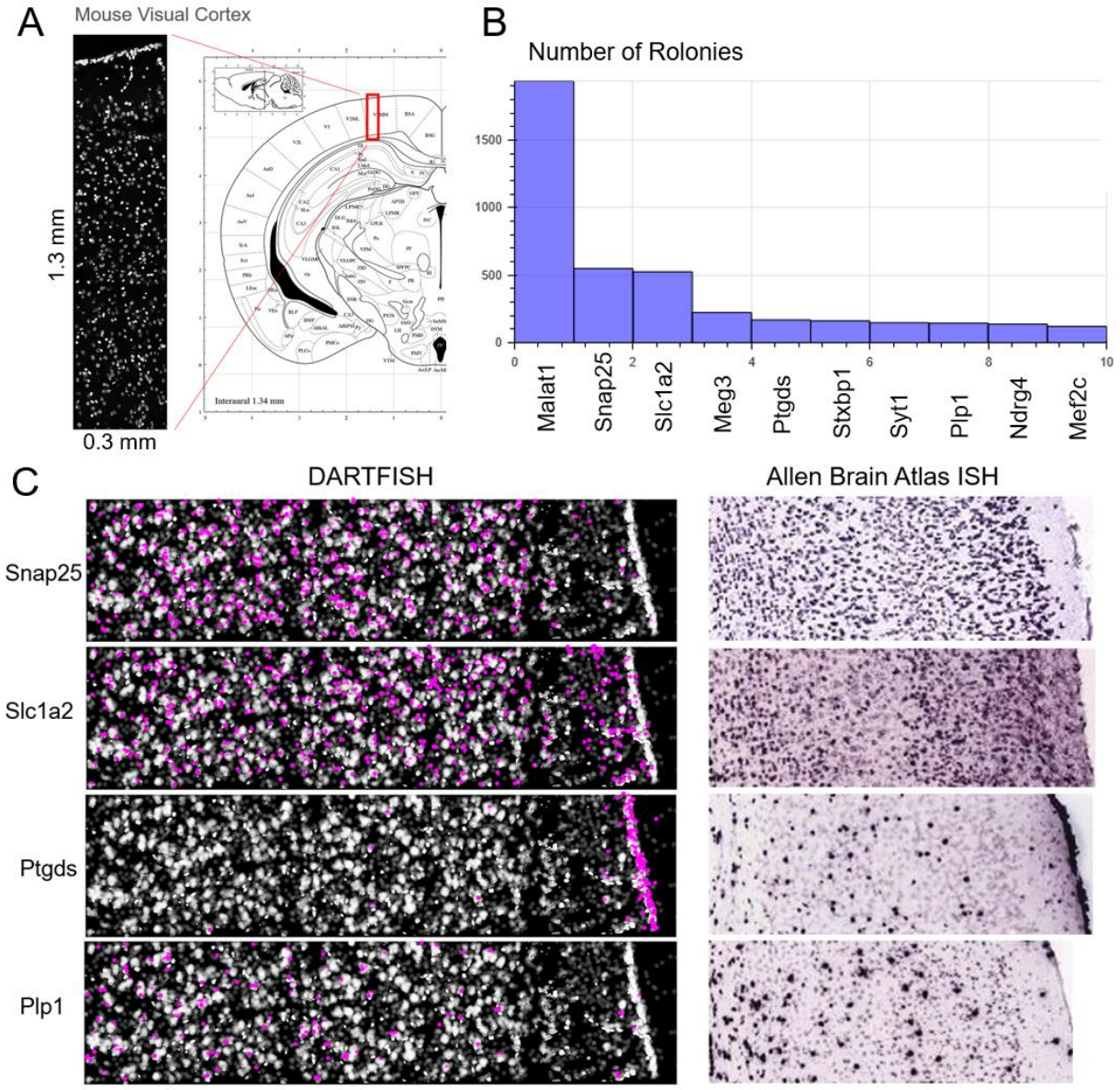


Figure 8: Validation of DARTFISH in a mouse visual cortex using a 478 gene probe set. (A) To measure the gene expression profile through the cortical layers a 1.3 mm tile approximately 1.5 mm from the longitudinal fissure was imaged. Shown here is the nuclei stained with DRAQ5. (B) The number of rolonies for each gene drops off sharply after the top three genes. (C) Comparing the images from DARTFISH and ISH data in the Allen Brain Atlas shows that DARTFISH is able to accurately capture the spatial gene expression in a mouse cortical section. The rolonies of each gene in DARTFISH are highlighted in magenta on the tile image from (A).

1.4.3 Validation in human brain

DARTFISH in human cortical sections is even more challenging than mouse cortical sections. Since human brain tissue isn't readily available and can only be collected post-mortem, the RNA quality as measured by RNA integrity number is low and the tissue is fragile, which necessitates the hydrogel embedding we developed. The old age of the donors means there is more autofluorescence background because lipofuscin accumulates with age, and the large size of the human brain makes imaging through all cortical layers time-consuming. Moreover, the area of cells in human brain tissue are approximately $1/7^{th}$ the area of fibroblasts we used previously, which makes obtaining a high density of rolonies in cells particularly important. The difficulty in applying highly multiplexed *in situ* RNA methods to human brain tissue is evidenced by the lack of any reported studies.

For initial trials in human brain sections, we decoded rolonies in white matter where the tissue was more intact, cell density was higher, and rolonies were more abundant. On average we saw 18 genes per cell and 137 rolonies per cell, which matches our results in fibroblasts after accounting for the difference in cell size.

For human brain we designed a probe set using genes identified in Lake et al., 2016 to be marker genes for neurons, glial cells, neuronal subtypes, as well as genes used in the Allen Human Brain Atlas for ISH. Ultimately we targeted 368 genes using 4,978 probes designed similarly to the probes used in previous fibroblast experiments. This is the same 6 round barcode probe set used to measure the misclassification rate in fibroblasts.

DARTFISH with this probe set was done in 10 **μ**m sections of fresh frozen occipital cortex (OCtx) and frontal cortex (FCtx) from patient #5342. The post-mortem interval was 14 hours. In order to image through the cortex from pia mater to white matter a thin tile image of

0.77 mm^2 was taken for each sample. The image stacks had optical sections every 0.3 **μ**m. Like the mouse cortex experiment, after decoding, the DNA was stained with DRAQ5 (ThermoFisher) to perform image segmentation of nuclei. The number of cells and rolonies covered in each sample are summarized in Table 2 below. To validate the specificity of DARTFISH, we compared the spatial expression of genes that had distinct layer specificity with ISH images from the Allen Human Brain Atlas. We also selected a few genes for RNAscope validation in adjacent sections of the same sample. As seen in Figure 9D there is very high similarity in gene expression distribution when comparing DARTFISH OCtx and RNAscope OCtx. As a negative control, we compared DARTFISH OCtx to RNAscope FCtx and found the distributions to be significantly different using the 2-sample Kolmogorov–Smirnov test.

Table 2: Experimental Summary of DARTFISH in Mouse and Human Cortex

Sample	Human Frontal Cortex	Human Occipital Cortex	Mouse Visual Cortex
Probe Set	4,978 probes targeting 368 genes	4,978 probes targeting 368 genes	7,133 probes targeting 478 genes
Imaged Volume Dimensions	150um x 5,000um x 10um (38 FOVs = 1.7mm^2)	150um x 5,000um x 10um (38 FOVs = 1.7mm^2)	300um x 1,300um x 10um (18 FOVs = 0.8mm^2)
Number of cells	1,035	786	668
Rolonies per cell	21.5	22.8	15.1
Genes detected	64 genes > 50 rolonies	62 genes > 50 rolonies	31 genes > 50 rolonies 70 genes > 25 rolonies

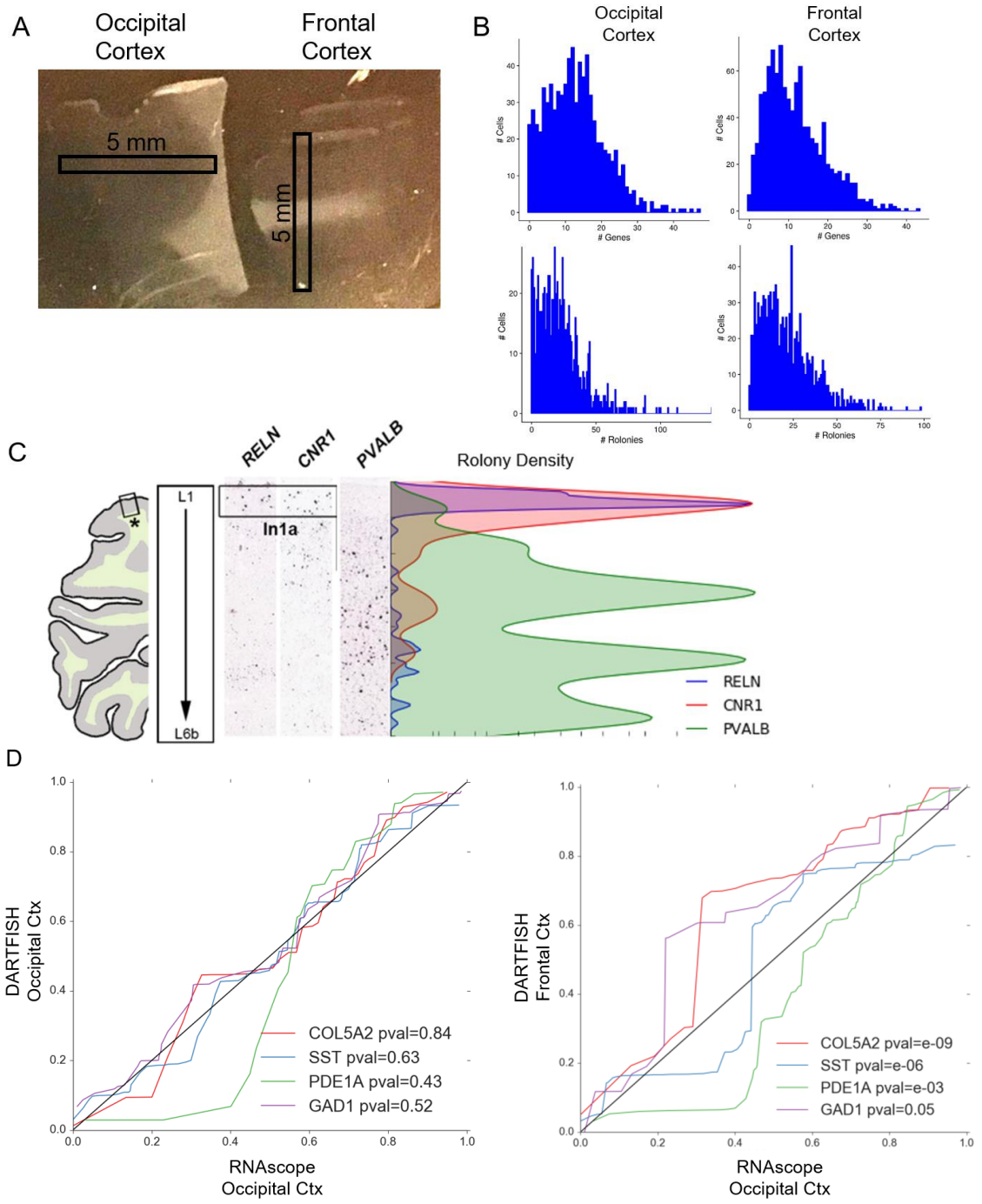


Figure 9: DARTFISH in 10 **µ**m human cortical sections containing all six cortical layers. (A) 5 mm x 150 **µ**m tiles were imaged in each section. (B) The distribution of number of genes per cells (top) and number of rolonies per cell (bottom). (C) Spatial distribution of select genes with layer specificity represented in a KDE plot of using a Gaussian kernel visualization. The expression matches ISH images on the left. (D) Two-sample KS test comparing DARTFISH gene distributions in OCtx and FCtx to RNAscope distributions in OCtx shows similarity between OCtx samples and dissimilarity between OCtx and FCtx as expected.

1.4.4 Conclusion

DARTFISH uses thousands of padlock probes with a hybridization-based barcode to capture hundreds of reverse transcribed RNA transcripts *in situ*. With rolling circle amplification, the signal-to-noise ratio is high enough to be used in tissues with high autofluorescence. DARTFISH is unique by being the only *in situ* RNA method with high signal amplification from RCA and is decoded by sequential rounds of FISH. Using FISH instead of sequencing is important in keeping the imaging process fast, simple, and affordable. In Chapter 1 we showed there is no bias in the barcodes by using two probe sets that shared targets but had different barcodes assigned to each target. We also showed padlock probes can be used to quantify relative target abundance if the counts are normalized by padlock probe efficiency. We first captured cDNA from human brain in a tube and found high correlation with RNA-seq. Then we carried out DARTFISH on cultured fibroblasts and again found high correlation of total captured padlock probe counts with RNA-seq ($r = 0.87$). DARTFISH is the only method we know of that can use thousands of padlock probes simultaneously and at concentrations in the picomolar range.

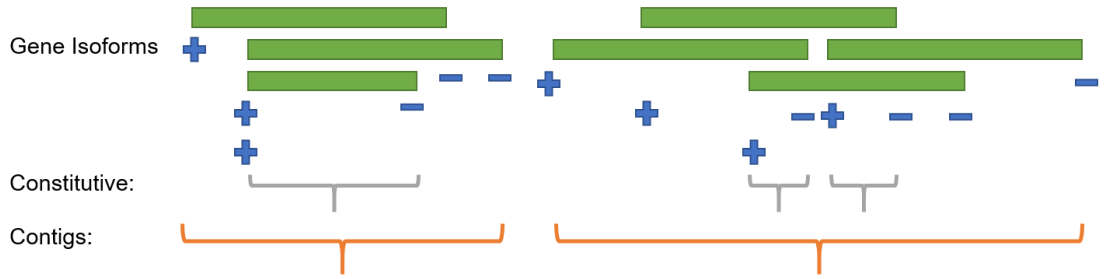
The main motivation for developing DARTFISH is for mapping cells in tissue sections, a much more challenging task than cultured cells. We demonstrated the use of DARTFISH in both mouse and human fresh frozen cortical sections as well as showed layer specific gene expression from DARTFISH matched reference ISH images from the Allen Brain Atlas and RNAscope data from serial sections. As expected, the human cortex was most difficult to process and we had to solve many issues such as tissue degradation and low RNA integrity. As far as we know, this is the only highly multiplexed *in situ* RNA data from human brain tissue.

1.5 Appendix to Chapter 1

Table 3: DNA Oligonucleotides for DARTFISH

Oligo Name	Sequence
AP1V4IU	G*T*AGACTGGAAGAGCACTGTU
AP2V4	/5Phos/TAGCCTCATGCGTATCCGAT
RE-DpnII_V4	TGCGTATCCGATC
RCA_Primer	GATATCGGGAAGCTGA*A*G
dcProbe0-488	/5Alex488N/TGTATCGCGCTCGATTGGCA
dcProbe0-Cy3	/5Cy3/CGTATCGGTAGTCGCAACGC
dcProbe0-Cy5	/5Cy5/ACGCTACGGAGTACGCCACT
dcProbe1-488	/5Alex488N/TCTTGCGTGCGATACGGAGT
dcProbe1-Cy3	/5Cy3/AACGGTATTCGGTCGTCATC
dcProbe1-Cy5	/5Cy5/CTGGTTCGGGCGTACCTAAC
dcProbe2-488	/5Alex488N/AGAACTTGCGCGGATACACG
dcProbe2-Cy3	/5Cy3/CTACTTCGTCGCGTCAGACC
dcProbe2-Cy5	/5Cy5/GACGAACGGTCGAGATTTAC
dcProbe3-488	/5Alex488N/GAATTGTCCGCGCTCTACGA
dcProbe3-Cy3	/5Cy3/CGTTTGATCGTTCGACCGAG
dcProbe3-Cy5	/5Cy5/AACTGCGACCGTCGGCTTAC
dcProbe4-488	/5Alex488N/CGGAATACGTCGTTGACTGC
dcProbe4-Cy3	/5Cy3/TACCATTCGCGTGCGATTCC
dcProbe4-Cy5	/5Cy5/CAGGGATCGGTCGAGTACGC
dcProbe5-488	/5Alex488N/GAGTGTCGCGCAACTTAGCG
dcProbe5-Cy3	/5Cy3/ACGTCTGCGTACCGGCTTAG
dcProbe5-Cy5	/5Cy5/CATGCGATTAACCGCGACTG
dcProbe6-488	/5Alex488N/CACGCTTACGATCCCGCTAT
dcProbe6-Cy3	/5Cy3/TCGTAACCCGTGCGAAGTGC
dcProbe6-Cy5	/5Cy5/CTCTCGTAGCGTGCGATGAG
dcProbe-ALL	/5Cy3/CTTCAGCTTCCCGATATCCG
Common linker	CTTCAGCTTCCCGATATCCG
KN2	KKKKKKKKKNN
hLINE1_F	TCACTCAAAGCCGCTCAACTAC
hLINE1_R	TCTGCCTTCATTTCGTTATGTACC
18SRNA_F	CTCAACACGGGAAACCTCAC
18SRNA_R	CGCTCCACCAACTAAGAACG

A



B

Probe Set Name:	V4	V6	V7	V8
Exons Targeted	Contig	Constitutive	Contig	Constitutive
Target Molecule	cDNA	cDNA	Rolony	Rolony
Avg Rolonies per FOV	771	71	306.25	42.3
Avg Genes per FOV	61.75	17	37.75	12.3

Figure S1: Padlock probes targeting all exon sequences (contigs) versus padlock probes targeting only constitutive exons. (A) Contigs were defined as sequence fragments present in *any* gene isoform and constitutive exons were defined as only the sequence fragments that were present in *every* gene isoform. (B) Rolony and gene counts for an experiment comparing padlock probes that target constitutive exons versus contigs and targeting cDNA versus FISSEQ rolonies showed that targeting contigs has almost an order of magnitude higher detection rate and that targeting cDNA almost doubles the detection rate.

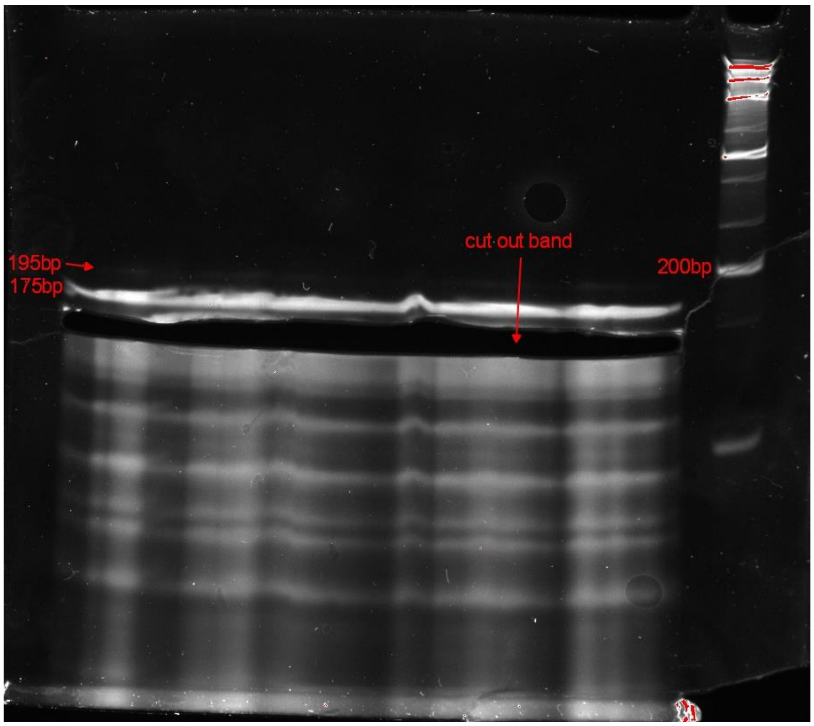


Figure S2: Image of a typical gel after size selection of correct band (155bp) during probe production. The two bands above the cut are the full length oligonucleotide after PCR (195bp) and the oligonucleotide with only one amplification arm removed (175bp).

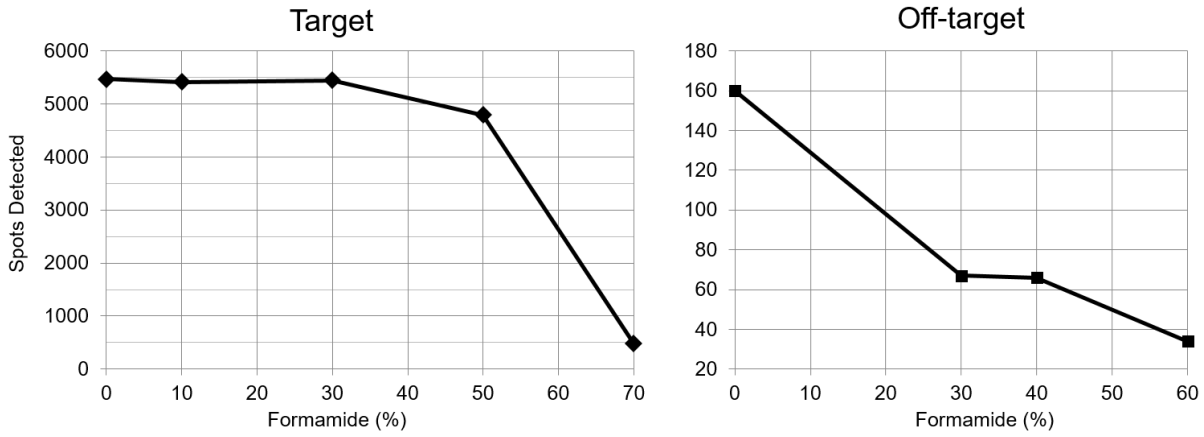


Figure S3: Experiment using two FISH probes, one with the correct sequence for hybridizing rolonies in the sample and the other with a sequence that isn't complementary anywhere to the rolonies. The results of testing FISH with varying formamide concentrations show that the best specificity without sacrificing sensitivity is at 30%.

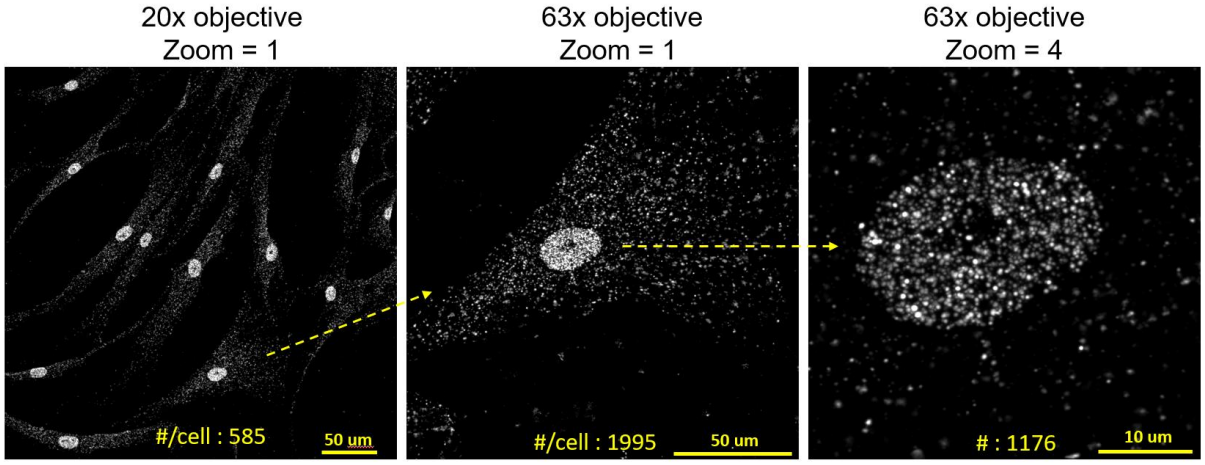


Figure S4: Comparison showing difference in detection rate by using two different objectives. The 20X objective is 0.75NA and the 63X objective is 1.4NA. Counting was done on a single maximum intensity projected 16-bit image using PISA. The rolonies here are FISSEQ rolonies and much higher density than typical DARTFISH experiments. The decoding algorithm is also fundamentally different than PISA, nonetheless, this still shows a high NA objective is better for resolving rolonies.

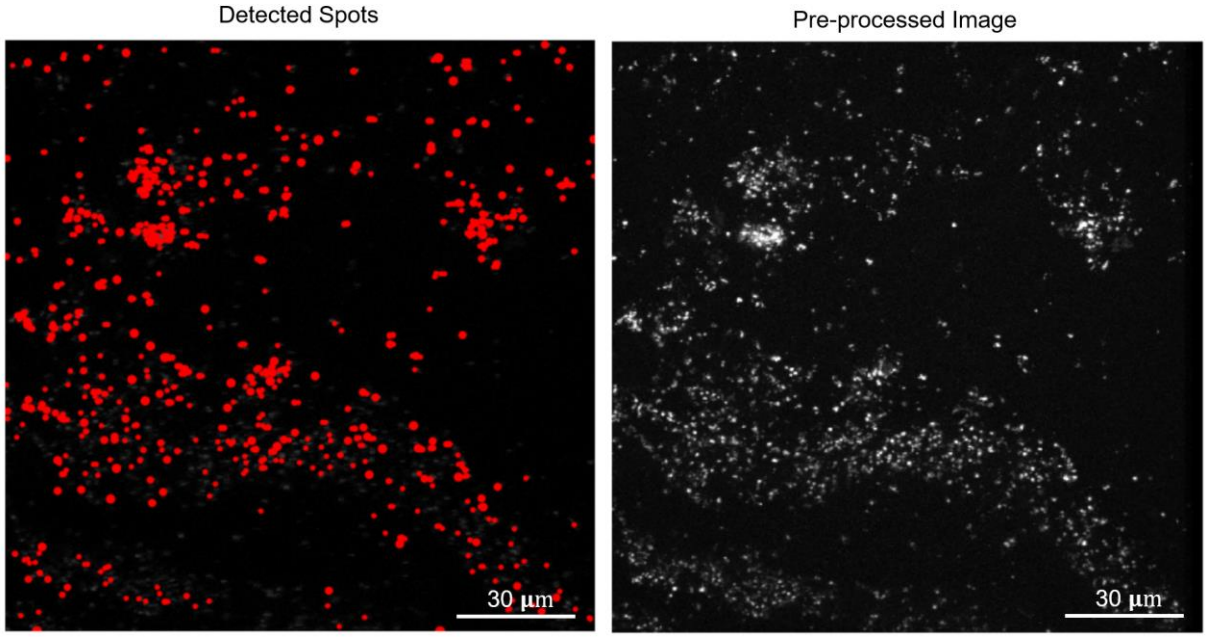


Figure S5: Example of spot-based detection performance on a DARTFISH image from human cortical section. Accuracy diminishes in high rolony density areas.

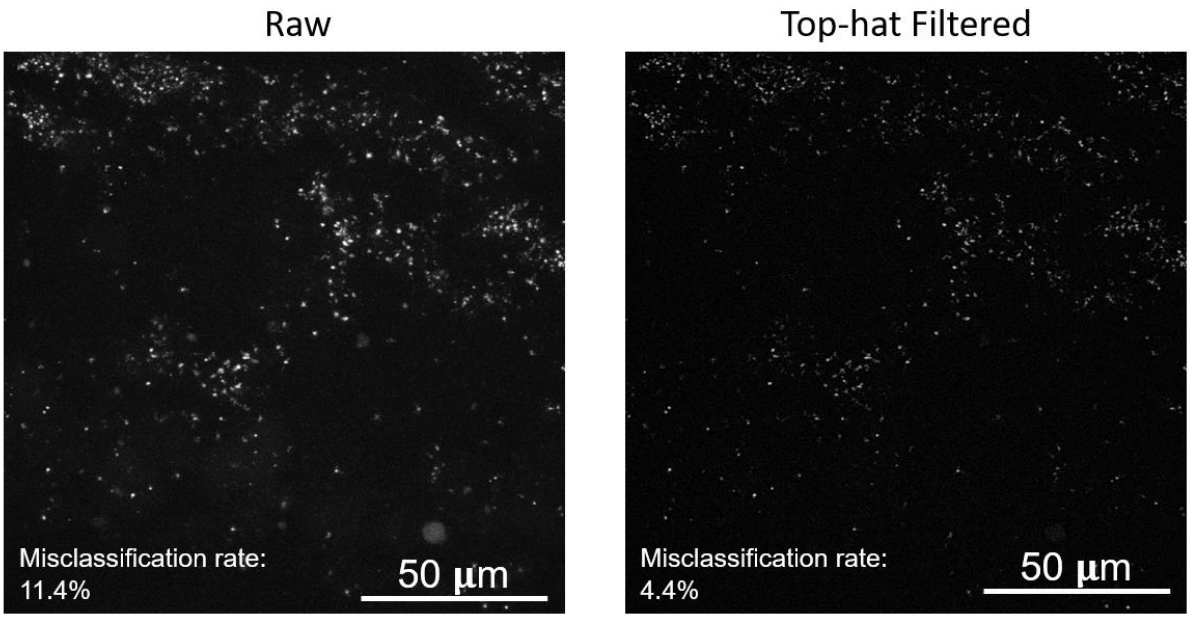


Figure S6: Top-hat filtering removes spots larger than typical rolonies and significantly improves the misclassification rate.

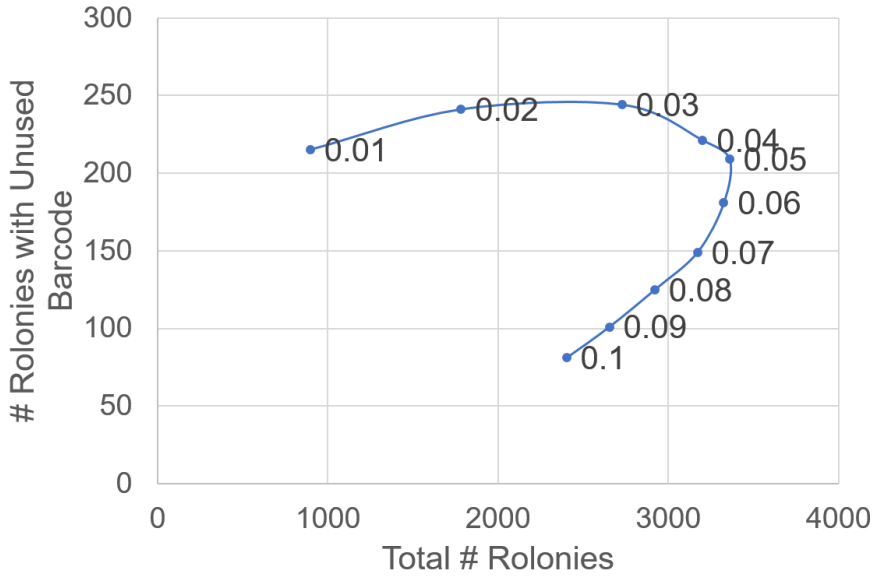


Figure S7: Finding the best threshold cutoff for calling an OFF state in the image analysis pipeline. Values between 0.01 to 0.1 were tested to see which threshold resulted in the most rolonies and lowest misclassification rate. A threshold of 0.05 consistently showed the best results across multiple experiments.

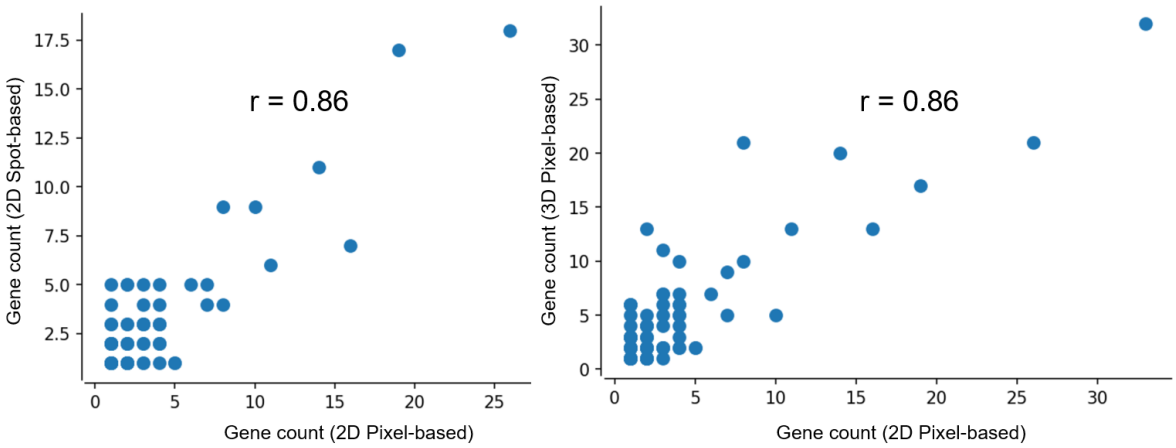


Figure S8: For field-of-view in Figure S6, images were decoded using three Starfish pipelines: 2D pixel-based, 2D spot-based, and 3D pixel-based. The correlation between 2D pixel-based and the other two pipelines is quite good with a Pearson's r of 0.86.

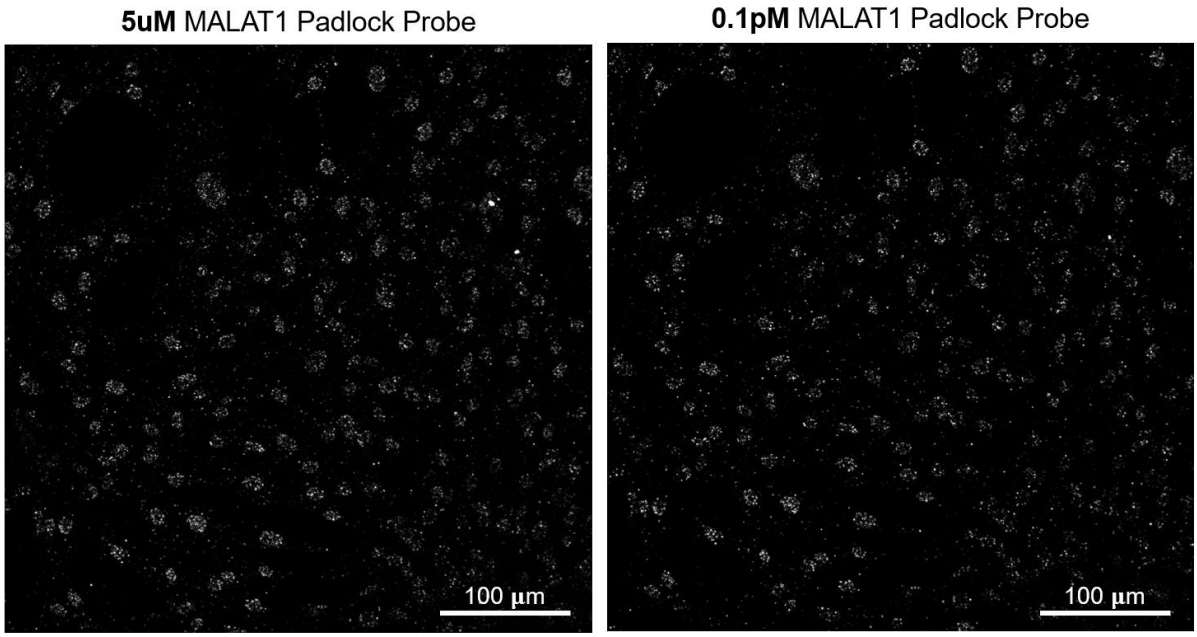


Figure S9: Tests using only one padlock probe targeting MALAT1, a highly expressed nuclear-enriched non-coding RNA transcript, were used to determine the best conditions for padlock probe hybridization and ligation. An experiment comparing 5 **µ**M and 0.1 pM concentration of padlock probe shows no significant difference in number of rolonies.

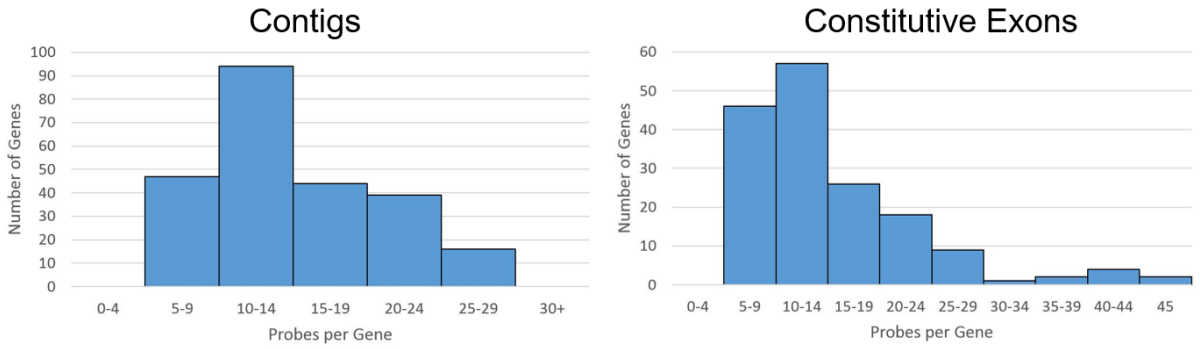


Figure S10: The distribution of number of probes designed per gene in the first probe sets used in PGP1 fibroblasts. The probe set targeting contigs targets 240 genes with 3,500 probes. The probe set targeting constitutive exon regions targets 165 genes with 2,500 probes.

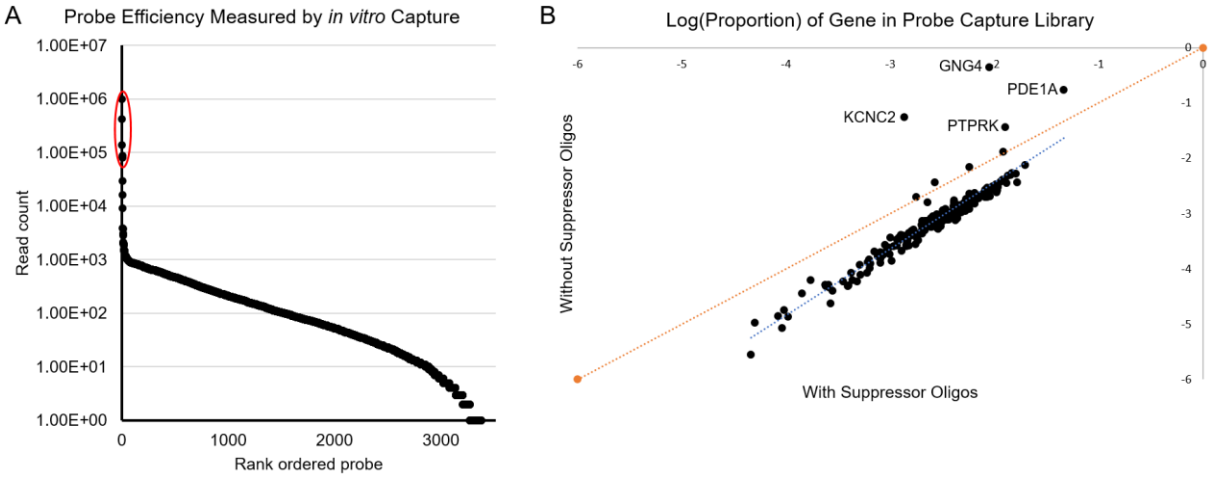


Figure S11: Suppressor oligos can be used to block the ligation of undesirably high efficiency padlock probes in a probe set obviating the need to reorder an oligo pool. (A) The first probe set used in fibroblasts had some probes targeting repetitive sequences in the genes: GNG4, KCNC2, PDE1A, and PTPRK. They also had extremely high efficiencies when measured by *in vitro* capture and sequencing. (B) Using suppressor oligos in an *in vitro* capture reaction successfully suppressed the number of ligated padlock probes for those genes as evidenced by the labeled markers left of the identity line. At the same time all other genes are shifted to the right because they make up more of the sequenced library.

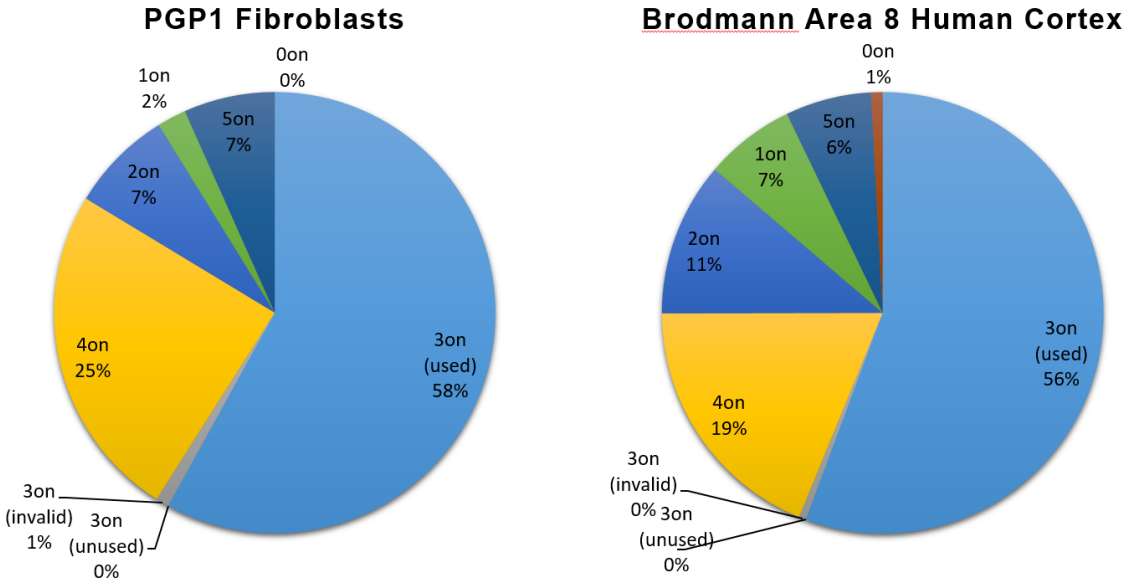


Figure S12: The distribution of barcodes decoded using spot-based decoding where rolonies are first identified using PISA on universal FISH probe. This probe set used a 5 round barcode with 3 ON states and used all 240 barcodes so there are no unused barcodes. It shows the most common error types are a single transition from an ON to an OFF state and from an OFF to an ON state.

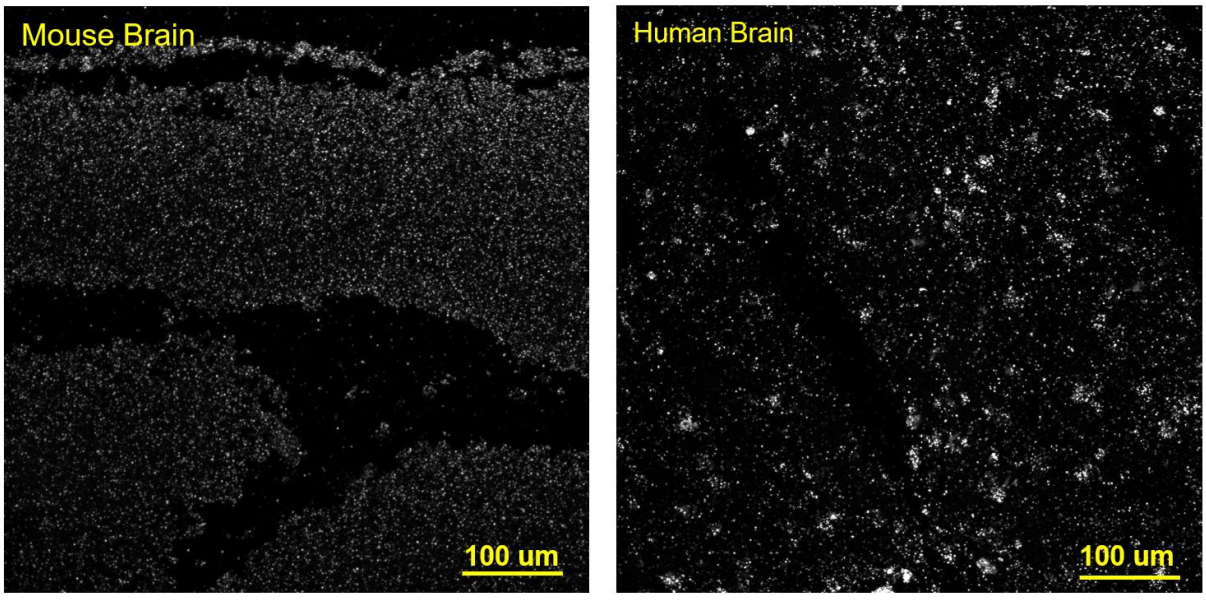


Figure S13: FISSEQ generated rolonies in mouse brain and human brain tissue sections show characteristic rolony density in the tissue. Cells in the human brain and larger with enrichment of rolonies in nuclei.

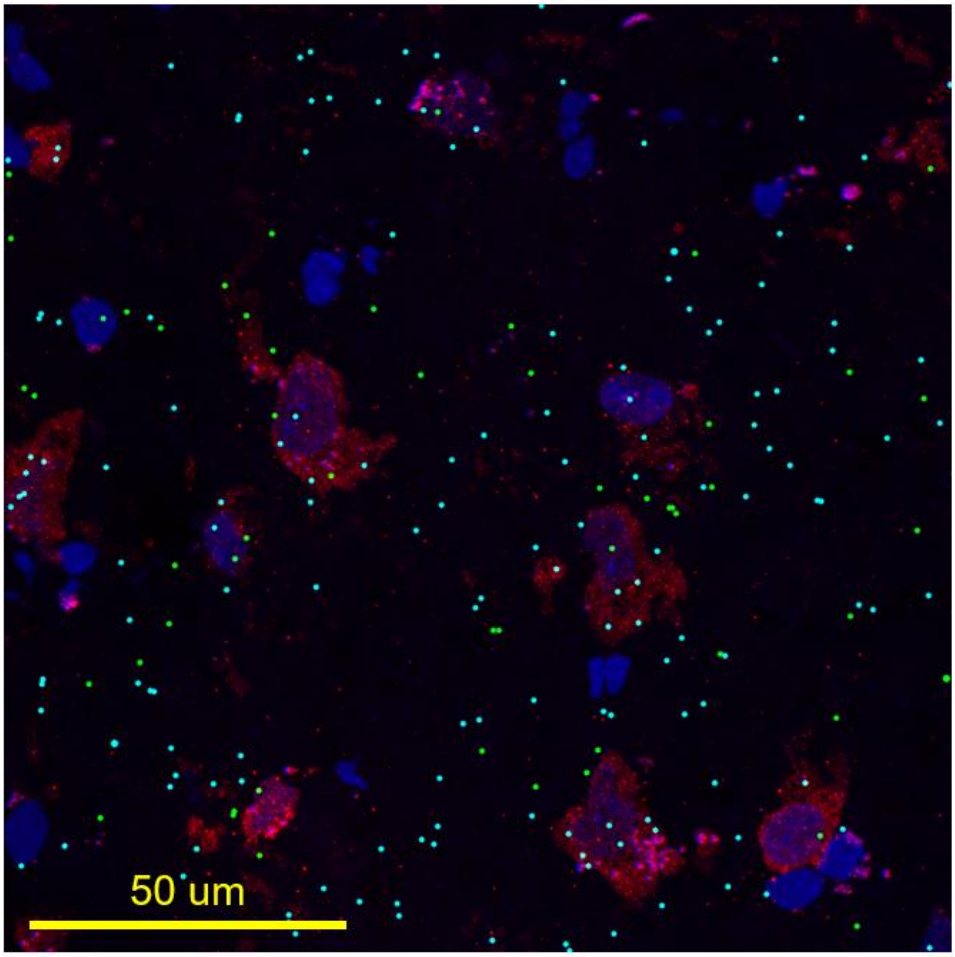


Figure S14: With some modifications to the protocol, DARTFISH is compatible with antibody staining after the rolonies are crosslinked. The pepsin digestion step and gel embedding were omitted from a DARTFISH experiment of the human cortex followed by staining with anti-NeuN and a fluorescent secondary antibody. The image here shows one field of view with nuclei in blue, anti-NeuN in red, neuronal-marker rolonies marked with cyan dots, and glial-marker rolonies marked with green dots.

1.6 Acknowledgement for Chapter 1

Much credit goes to Ho Suk Lee, who contributed greatly to the initial development of DARTFISH. Justin Dang helped develop the gel embedding protocol as an undergraduate volunteer. Kefei Gao performed experiments to optimize *in situ* reverse transcription. Justin and Kefei also assisted with a lot of the padlock probe production. Yun Yung and Grace Kennedy from the Chun lab provided the mouse and human brain tissue sections.

Chapter 1 is coauthored with Lee, Ho Suk; Dang, Justin; Gao, Kefei; Yung, Yun; Kennedy, Grace; and Zhang, Kun. The dissertation author was the primary author of this chapter.

CHAPTER 2. MAPPING SINGLE-CELL RNA-SEQ CELLS TO HUMAN AND MOUSE CORTICAL SECTIONS

2.1 Abstract

To generate a comprehensive reference like the Human Cell Atlas, cells not only need to be mapped spatially, they must also be linked with other data types such as molecular profiles and morphology. Data from *in situ* RNA methods like DARTFISH need to be integrated with single-cell or single-nuclei RNA-sequencing data. Dissociative single-cell methods like scRNA-seq have the advantage of high-throughput and deep *de novo* sequencing that can lead to discoveries of new rare cell populations and types. DARTFISH would best serve as a way to add spatial information to these cells by creating a map that scRNA-seq cells can be projected onto. In this chapter we attempted to do this with DARTFISH data from Chapter 1.

2.2 Introduction

Given the limited volume inside cells, the amount of transcriptome information obtainable through microscopy cannot compare to the throughput of NGS. Moreover, all *in situ* RNA methods besides FISSEQ use targeted probes, limiting *de novo* discovery of new transcripts. However, we have shown DARTFISH can provide the spatial information of hundreds of genes simultaneously in tissues with high autofluorescence like the human brain. In addition, high-throughput single-cell or single-nuclei sequencing data is rapidly being published for all human organs (Cui et al., 2019; Reyfman et al., 2019). In order to create a cell atlas with the spatial information and the full transcriptome, one popular approach is to project cells or cell types from single-cell RNA sequencing onto maps generated by *in situ* methods using genes that intersect both sets of data (Stuart et al., 2019).

There are a number of computational steps for mapping to *in situ* images. The first of which is how to cluster RNA spots or rolonies belonging to the same cell. Defining the boundary of a cell can be done by using the cloud of RNA spots (Tsanov et al., 2016) or using counterstains. However, both these methods work significantly better on cultured cell monolayers with separation between cells. Cell segmentation in tissue is not trivial. Without complete cell segmentation, assigning spots or rolonies to cells also becomes a hurdle. DARTFISH rolonies make the problem even more difficult than traditional RNA FISH because the rolonies are sparse.

The other aspect of mapping the cell types from scRNA-seq onto DARTFISH cells is the actual integration of the two types of data. The first published study to map single-cells from scRNA-seq to a spatial map used binarized WMISH data from the brain of a marine annelid (Achim et al., 2015). By matching gene expression profiles, it was able to map 81% of cells with high confidence using only 72 genes, but in the discussion also noted the information content of the genes was very important in determining the success of mapping. Genes that are more spatially restricted and have less overlap are most useful. This means the complexity and heterogeneity is also a determinant in the success of mapping. With the advent of reference atlases of much higher resolution and more gene information, more sophisticated methods (Stuart et al., 2019) are available. However, they are only compatible with spatial data nearing scRNA-seq level of sequencing depth.

2.3 Approach

Before cells from single-cell or single-nuclei RNA sequencing can be mapped, the DARTFISH data must be partitioned into cells. DARTFISH data after decoding is a list of RNA targets detected along with their spatial coordinates, rolony size, and a quality metric if

decoding was done using Starfish. The overlapping field-of-views are stitched together so rolonies are all on the same coordinate system and DIC and DRAQ5 stained tile images become one large image.

The first step is to segment cells. There are two paradigms for doing this on *in situ* data. One is using the density cloud and sometimes identity of RNA spots, or rolonies in the case of DARTFISH, to predict the boundaries of cells. This method requires a high density of rolonies and separation between cells. Of course cells in tissue often interact and make contact with other cells via cell junctions. In the cases where rolonies from two cells cannot be distinguished by rolony density, the identity of RNA may help separate them. For example if the RNA are gene markers for two distinct cell types and there's a clear division between the two types. However, if two of the same cell type are in close proximity, this method will count them as one cell. We tried spatially clustering DARTFISH data from human cortex where the cells were sparse and the rolonies were mostly in and around the nuclei using density-based spatial clustering of applications with noise (DBSCAN).

The other paradigm is to threshold the stained nuclei image and use the nuclei as an approximate center for the cell. The challenge here is accurately thresholding nuclei since their brightness from DRAQ5 staining can vary between and within cells. Also, some nuclei are packed tightly together and watershed algorithms are required (see Figure S15). We were unable to automatically threshold and segment our DRAQ5 stained images to our satisfaction so we had to rely on a supervised program to get 100% accuracy. There are also some approaches to combine both paradigms, using the nuclei to seed the center of rolony density clouds. Currently, no gold standard method exists and cell segmentation algorithms are still actively being developed by the community.

The second step is to assign rolonies to cells. This is trivial if the rolonies were used to segment the cells. If however the nuclei were segmented, a straightforward approach is to assign rolonies to the nearest nuclei. This works better with more separation between cells and for round cells. The way we implemented this was to use the centroids of segmented nuclei as seeds to generate a Voronoi diagram that partitions the image into Voronoi cells. The Voronoi cells are regions where every point within the cell is closer to the seed point than any other seed point. Therefore we can assign rolonies to whichever Voronoi cell they are in. The weakness of this approach, as previously mentioned, is cells like neurons with many projections where parts of the cell may be closer to other cell's nuclei than its own.

The most accurate solution would be to use a membrane stain to determine the boundaries of a cell but unfortunately there is no stain compatible with cortical tissue that has been permeabilized, that we are aware of. There is also an effort to use machine learning to train a convolutional neural network to segment cells by training it on tens or hundreds of thousands of human-segmented images. Creating the training data is an enormous task that will likely be crowdsourced.

After cell segmentation and rolony assignment, the DARTFISH data can be formatted as a single-cell gene expression matrix with spatial information linked to each cell. The count matrix is of the same format as that of scRNA-seq, so mapping one set of data to the other, or integrating the two, seems logical. The two major differences are that DARTFISH counts have padlock probe efficiency bias that can be normalized, and the DARTFISH matrix is much sparser. Instead of over a thousand genes per cell and over ten thousand reads per cell it is just a handful of each. We explored if with the proper normalization we could use dimension reduction techniques and machine learning classifiers to integrate the two types of data.

2.4 Results

Using DBSCAN from Python's scikit-learn with 'eps' set to 0.2 and 'min_samples' set to 10 on human cortex from Brodmann area 8 results in clusters that look accurate as shown in Figure 10. One feature of this method is that not every rolony has to be included in a cluster. However the optimal parameters need to be adjusted depending on rolony density and cell-to-cell proximity, which varies within a sample. This means a user has to manually run DBSCAN for every tile image, possibly many times.

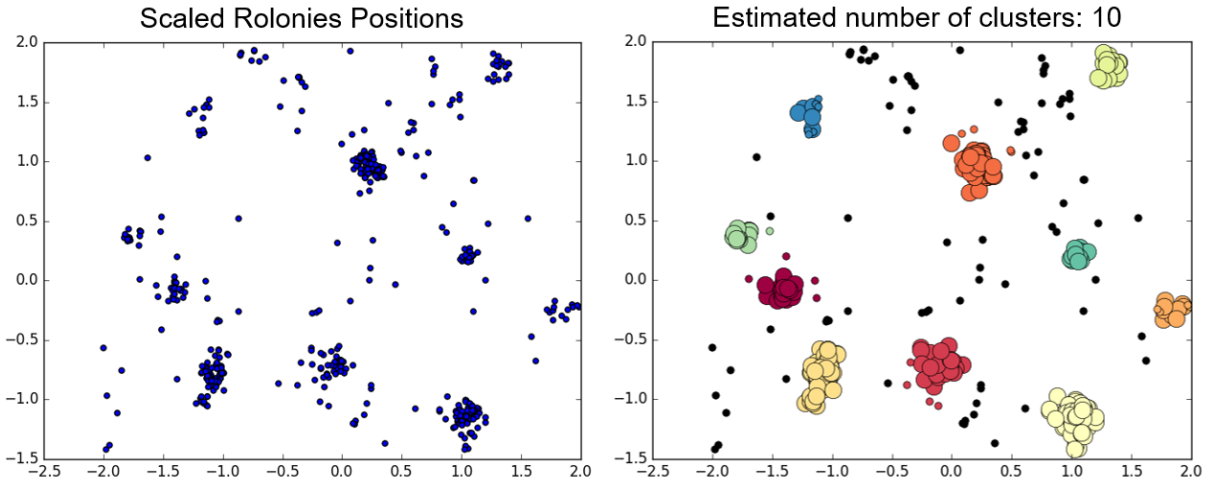


Figure 10: Density-based spatial clustering of applications with noise applied to rolonies in upper layers of human cortex from Brodmann area 8. Two dimensional spatial coordinates for decoded rolonies were standardized to have zero mean and unit variance on the left. The color coded clusters were found by scikit-learn's DBSCAN with 'eps' = 0.2 and 'min_samples' = 10.

We found thresholding nuclei-stained images and partitioning into Voronoi cells to be a more robust method, especially for deeper layers and white matter on brain tissue where rolonies were more distributed outside of the nucleus (see Figure S15). Images are preprocessed by subtracting a low offset value to remove obvious background and then contrast is enhanced with contrast-limited adaptive histogram equalization. Otsu's method is used to threshold and

binarize the preprocessed image. Small foreground objects are removed with an opening operation and holes are filled with a closing operation. Finally, connected components that represent nuclei are labeled and displayed to the user for inspection. The user sees the labeled nuclei and original image and can decide whether to approve it, remove it, or use the watershed algorithm to split the nuclei into multiple.

Voronoi tessellation only needs seed points, so regionprops in MATLAB is used to find the centroids of every connected component. The VoronoiLimit package in MATLAB is used to efficiently calculate the Voronoi cells. Then we assign every rolony to a labeled nuclei based on which Voronoi cell it is in. This can be organized as a gene expression matrix with genes and cells as rows and columns. It can also be formatted as a rolony table with every row being a rolony and columns being properties such as gene name, rolony size, spatial coordinates, and cell assignment.

Our first attempt to classify single cells using DARTFISH that had been spatially partitioned was to use the gene markers included in the probe set. Some gene markers are for broad cell types: neurons, microglia, astrocytes, oligodendrocytes, and endothelial. Other gene markers are for neuron types like interneuron or excitatory neuron. At the finest level, there are neuron subtypes identified in Lake et al., 2016. We aimed to classify cells into broad cell types and used many genes as markers for each cell type (see Table #). Cells were classified based on the plurality of gene markers expressed as long as the plurality exceeded 1/3rd and there were more than three rolonies assigned to that cell.

We tried this approach in a human tissue section from Brodmann area 8 using the same 4,978 oligo probe set from Chapter 1. We characterized the proportion of neurons in the cortex and white matter and found approximately half of all cells in the cortex were classified as

neurons using this plurality of DARTFISH rolony gene markers approach. Also no neurons were detected in the white matter. We validated these finding by staining an adjacent tissue section with anti-NeuN, which only stains the nuclei of neurons. As seen in Figure 11, anti-NeuN staining confirmed the absence of neurons in white matter and 53% neurons in 0.11 mm^2 cortex.

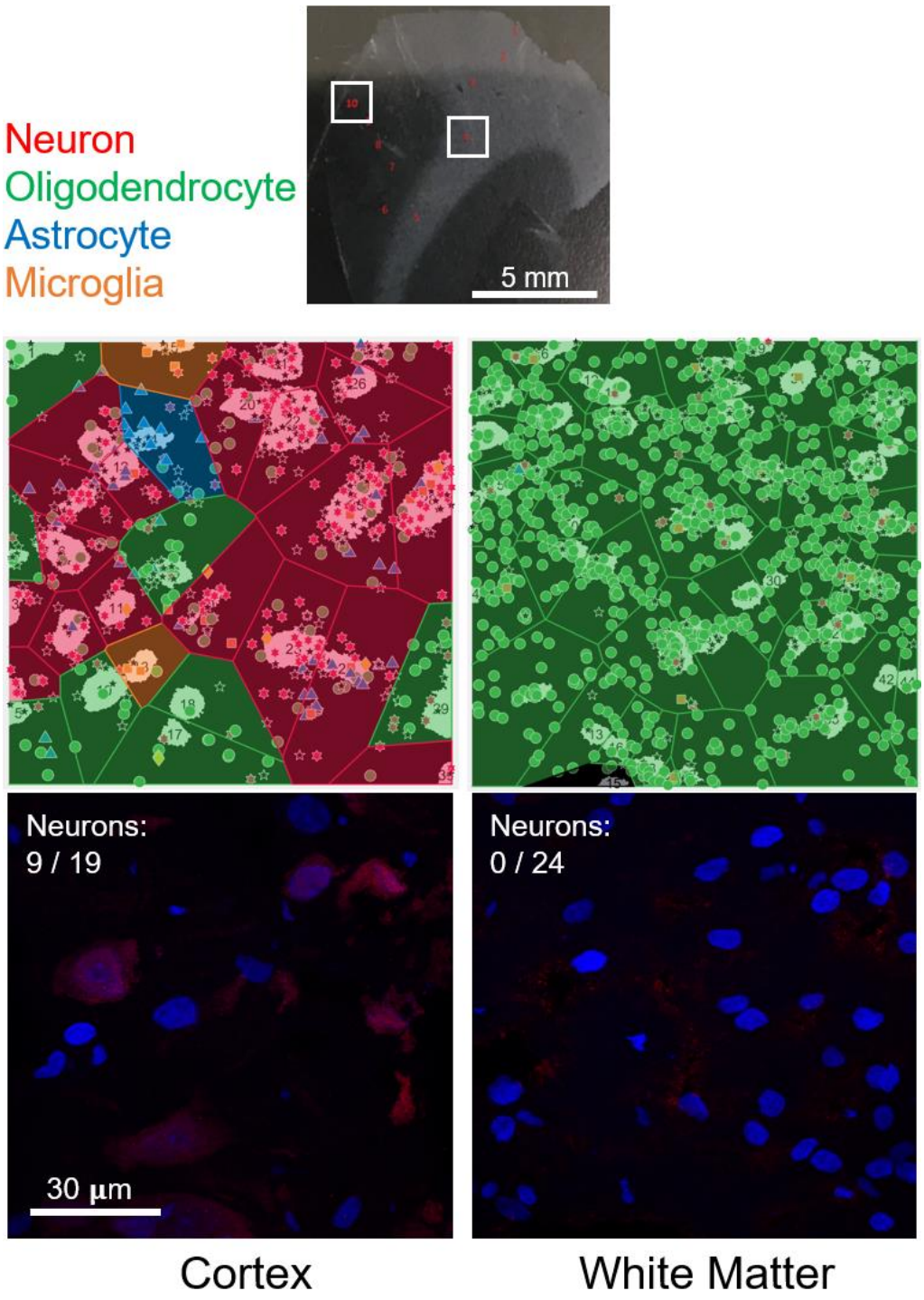


Figure 11: Classification of cells in cortex and white matter of Brodmann area 8 using marker genes of major cell types. Voronoi partitioned cells are color shaded by cell type and show approximately 50% neurons in the cortex and no neurons in white matter. This was confirmed by anti-NeuN staining (bottom images) in an adjacent section where anti-NeuN is red and DRAQ5 is blue.

We proceeded to apply the same method to human occipital cortex that we imaged across all six cortical layers including the pial surface. Like in Brodmann area 8, we detected approximately half the cortical cells to be neurons and the majority of cells in white matter to be oligodendrocytes. Most interestingly, at the pial surface there were a handful of neighboring endothelial cells that seemed to be part of a ring of cells (see Figure 12). Sectioning and attachment to the coverslip causes some damage to the tissue, especially at the edges, so it is expected to have worse RNA quality and therefore fewer rolonies for cell classification. We suspect if we had more rolonies in the cells that form a ring we would see that they are endothelial cells that are part of a pial vessel.

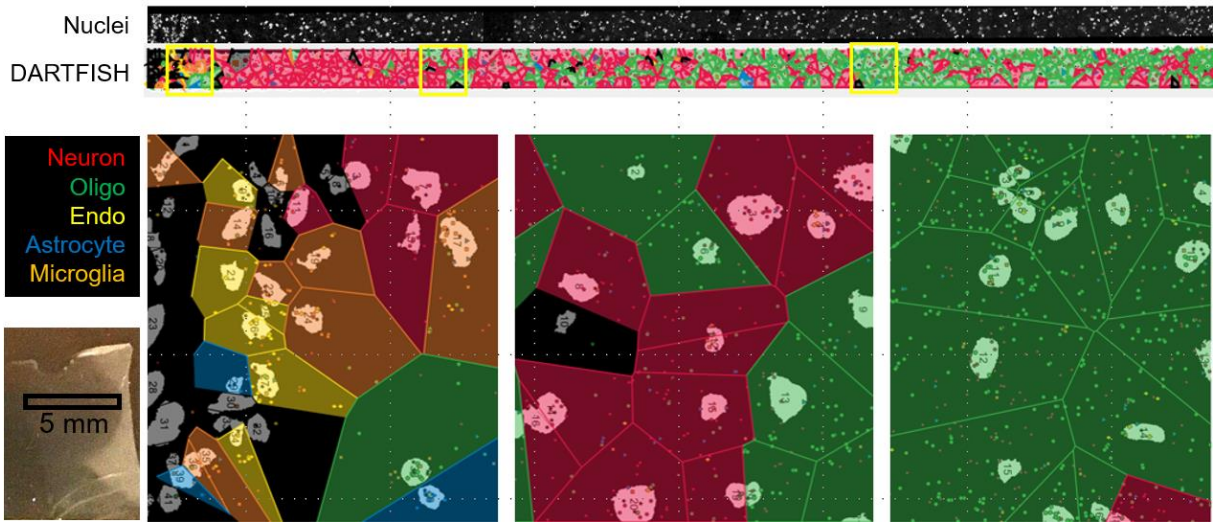


Figure 12: DARTFISH reveals single cell resolution spatial heterogeneity of human occipital cortex. 10 **μ**m fresh frozen sections were processed using DARTFISH targeting 368 genes. Bottom left: Rolonies were decoded in a 5mm long tilescan that spanned pial surface and the six cortical layers. Top: DRAQ5 stained nuclei were used to seed a Voronoi diagram that partitioned rolonies into single cells. Top left: Cells with a gene marker majority of a distinct cell type were labeled by color. Left inset: Subpial tissue shown containing cells for blood vessels and macrophages. Middle inset: Layer II/III has a mixture of neurons and oligodendrocytes.

In the human visual cortex we tile imaged across the tissue from end to end and could clearly see white matter running down the center. Looking at the abundance of RELN, a layer I gene marker, we were able to confirm both ends contained the full six cortical layers. Oddly, on the right end, the peak in RELN overlapped with PVALB and CNR1, which should not be in layer I. Looking at the tissue by eye, it was a bit denser and more opaque, suggesting the tissue might have folded on itself. We confirmed this along with the position of the white matter by staining the myelin with a Kluver-Barrera stain in a few adjacent sections from the same batch (see Figure 13). The cell type classification looked similar to what we expected based on the occipital cortex and the Brodmann area 8 sections we analyzed previously. It is important to remember that the cell segmentation and rolony assignment is not accurate for cells with projections that extend far away from their nuclei, such as neurons. So the relatively few astrocytes and microglia in the cortex could be due to the fact that the probe set has many more neuron markers and the transcripts could be in axons and dendrites that are then incorrectly assigned to a neighboring nuclei.

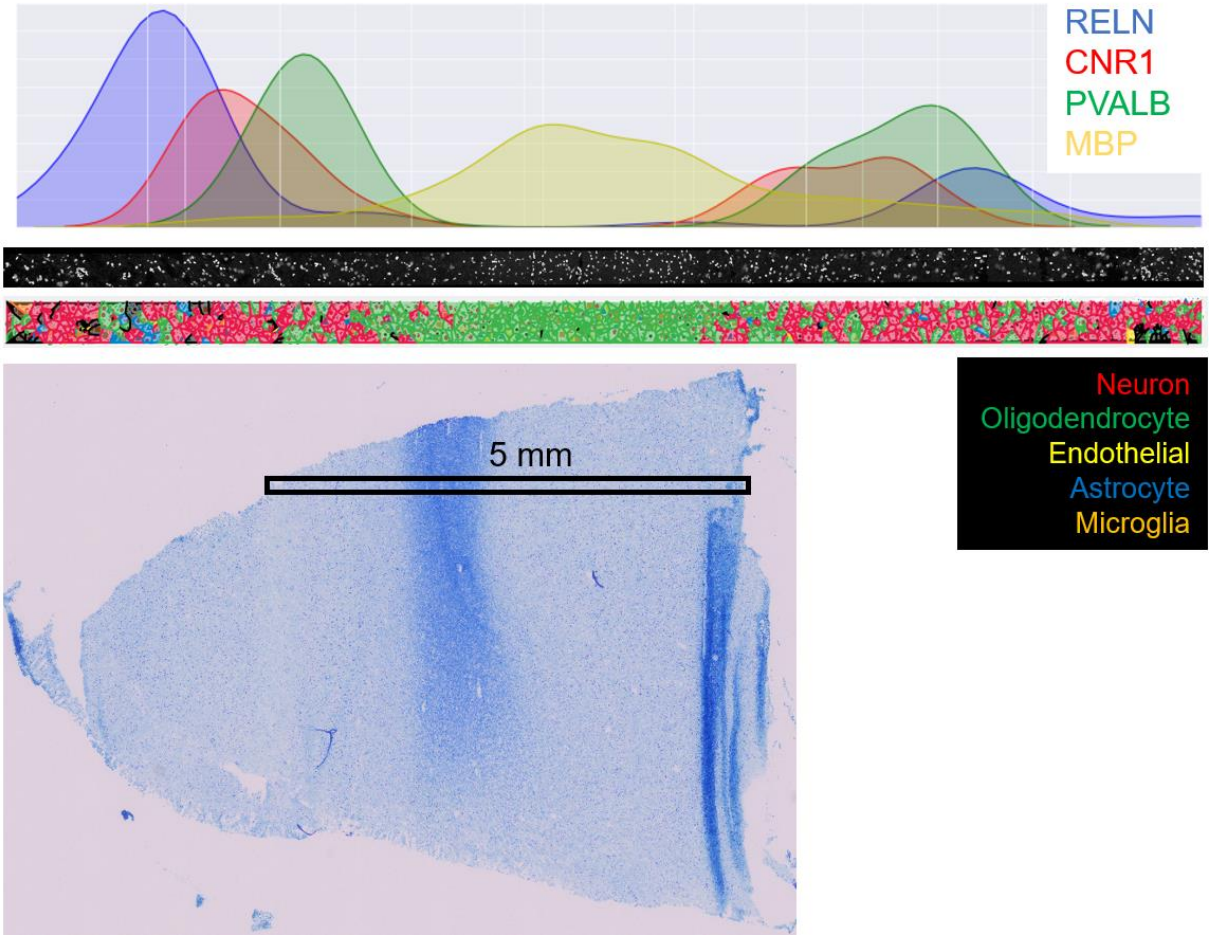


Figure 13: DARTFISH cell classification in human frontal cortex. Images were tiled across cortical layers through white matter and then another six cortical layers. Top tile: KDE plot of rolonies for four genes detected along 5 mm tile image. RELN is a marker for layer I of the cortex and can be clearly seen on the left end. The right end shows RELN overlapping with PVALB, which is due to tissue folding over during sectioning as seen in bottom Kluver-Barrera stain. Middle tile: Nuclei stained image. Bottom tile: Voronoi partitioned cells colored by cell type classification using gene markers.

More granular subtypes do not have single gene markers. Classifying the spatially segmented cells to subtypes requires considering the expression levels of sets of genes. Since the subtypes were defined from clusters found from scRNA-seq gene expression matrices, we attempted to apply a similar approach on our DARTFISH gene expression matrix. We tried two methods that follow the same principle: dimension reduction of scRNA-seq expression matrix,

train a classifier given knowledge of the cell types, and then apply the same transformation on the mouse visual cortex DARTFISH expression matrix. We chose to map to mouse because the rolonies were more enriched in nuclei, theoretically making the cell segmentation and rolony assignment more accurate (see Figure S16). For scRNA-seq we used publicly available data from mousebrain.org and selected the closest reference, middle cortex cells from a 60 day postnatal mouse. The reference included 2,361 cells with 1,564 genes per cell and 74,214 reads per cell.

The first method took the reference gene expression count matrix and removed any genes not present in the DARTFISH data. For the DARTFISH gene expression count matrix, any cells with less than four rolonies were removed and then counts were normalized by padlock probe efficiency. Both sets of data were then normalized by 'reads per cell' and log(x+1)-scaled. Then PCA was run on reference data with 10 principal components and a k-Nearest Neighbor (kNN) classifier was trained with reference cell types. Finally, we mapped the DARTFISH data with the kNN (see Figure S17). For validation we plotted the positions of rolonies and labeled them based on the cell type they belong to. The reference cell types have expected layer expression that we compared to. In particular, given the high expression of Ptgds in the pia mater and with Ptgds being a gene marker for VLMC2 (vascular leptomeningeal cells), we expected to see VLMC2 or VLMC1 labeled at the surface. Unfortunately, using this method classified those cells as mature oligodendrocytes, which also express Ptgds. However, given the location of those cells, it is much more likely that VLMC2 would be the correct classification.

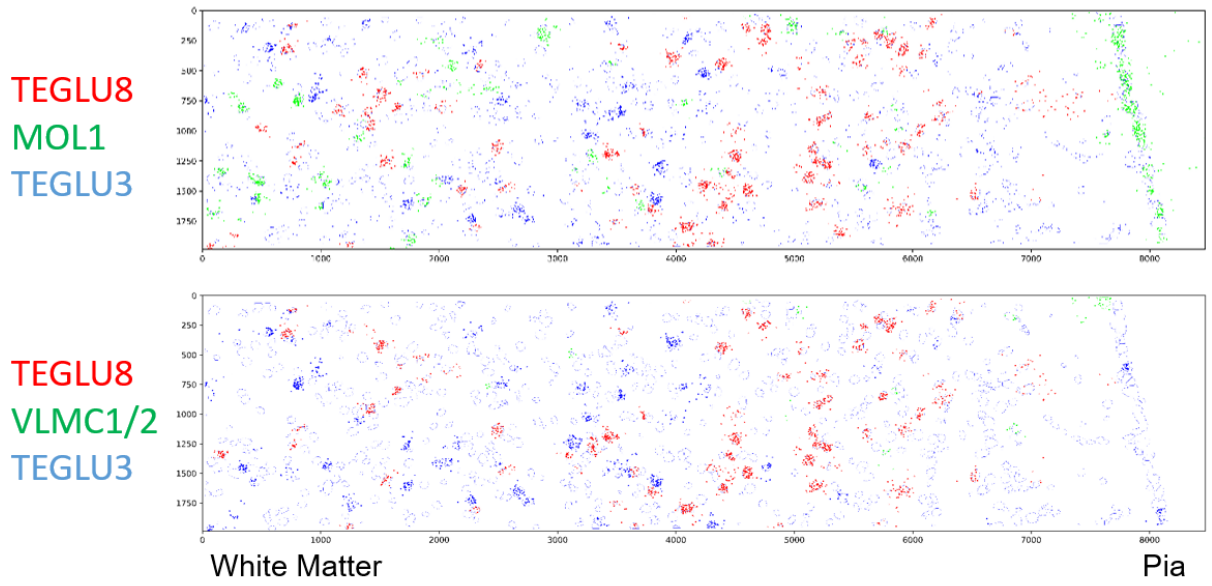


Figure 14: DARTFISH rolonies in mouse visual cortex labeled by the cell type they mapped to with kNN. TEGLU8 is an excitatory neuron cluster likely found in layer IV and TEGLU3 is an excitatory neuron cluster likely found in layer VI. MOL1 is a mature oligodendrocyte and VLMC1 and 2 are vascular leptomeningeal cells. Based on high expression of Ptgds in the pia mater (see Figure 8C), the cells at the surface should be classified as VLMC but instead they are labeled MOL1.

For the second method, we tried changing the algorithms. The first change was to use a more sophisticated dimension reduction technique called Uniform Manifold Approximation and Projection (UMAP). Secondly, we tried to break the classification down into hierarchical levels by first classifying broad cell types and then taking those cells and trying to map to a more specific subtype. The data scaling and normalization stayed the same.

As a proof of concept, we randomly split the scRNA-seq reference into training and test data sets. We then used the training set to fit a UMAP embedding and train SVC and kNN classifiers on six broad cell types. To validate, we transformed the test set with the UMAP embedding and classified with SVC and kNN with an accuracy of 98.8%. To demonstrate a few hundred genes is enough for accurate classification, we trimmed the reference scRNA-seq data to only the 375 genes present in the DARTFISH probe set and then repeated the same process.

With only 375 genes, the SVC and kNN classifiers both had an accuracy of 99.1% (see Figure S18).

We then followed the same procedure with scaled DARTFISH data and bootstrapped DARTFISH data where the gene counts were permuted for each cell. The DARTFISH cells are embedded into clusters with a similar topography to the reference cell clusters and the bootstrapped data only forms a cluster in the location of some neurons. This suggests the DARTFISH cells near non-neuron clusters are not random.

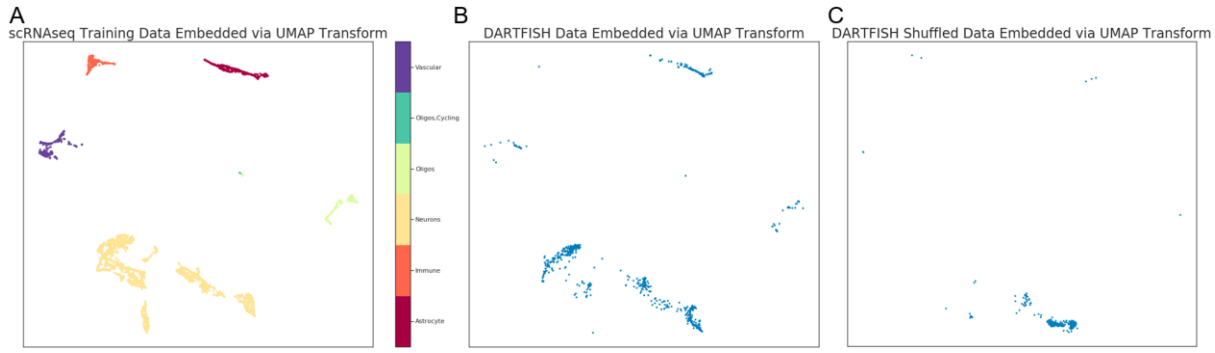


Figure 15: Supervised UMAP dimension reduction of scRNA-seq cells and DARTFISH cells. (A) scRNA-seq cells were embedded to cluster six broad cell types: vascular, cycling oligos, oligos, neurons, immune, and astrocyte. Neurons in yellow form a disperse cluster reflecting the diversity in neuronal subtypes. (B) DARTFISH cells embedded with the same transformation leads to clusters that match the reference. (C) Bootstrapped DARTFISH cells embedded with the same UMAP transformation only cluster in a location of part of the neuron reference cluster.

As in the first PCA and kNN method, we validated by plotting the UMAP embedded SVC classified cells onto the mouse visual cortex (see Figure 16). We expect the Ptgds expressing cells in the pia mater to be labeled as vascular cells. However, only four cells out of twenty five are classified how we expected. The rest are a mixture of neurons and astrocytes, which is unlikely to be true. Also, there should be a majority of oligodendrocytes in the white matter but this classification method resulted in very few cells labeled as oligodendrocytes.

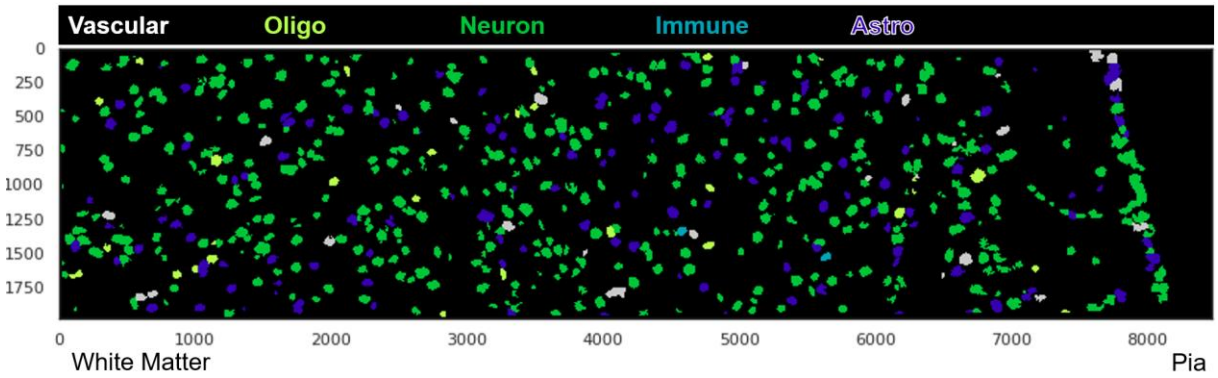


Figure 16: Mouse visual cortex cells labeled by a support vector classifier trained on UMAP embedded scRNA-seq data. Nuclei of vascular cells are colored white, and there are four on the pial surface. Based on Ptgds rolonies we would expect almost all cells in the pia mater to be vascular cells.

2.5 Conclusion

To generate a comprehensive reference like the Human Cell Atlas, cells not only need to be mapped spatially, they must also be linked with other data types such as molecular profiles and morphology. Within the scope of this dissertation, that means data from *in situ* RNA methods like DARTFISH need to be integrated with single-cell or single-nuclei RNA-sequencing data. As mentioned previously, dissociative single-cell methods like scRNA-seq have the advantage of high-throughput and deep *de novo* sequencing that can lead to discoveries of new rare cell populations and types. DARTFISH would best serve as a way to add spatial information to these cells by creating a map that scRNA-seq cells can be projected onto. In this chapter we attempted to do this with DARTFISH data from Chapter 1.

We first showed that cells can be classified into broad cell types (e.g. neurons, astrocytes, oligodendrocytes, etc.) by simply using gene markers included in the probe set of DARTFISH. In human cortical sections we can distinguish between white matter and cortex based on abundance of neurons. We are also able to detect distinct cerebral structures like blood vessels in the pia mater.

However, when we try to map cell clusters from single-cell RNA-seq the cells do not map to where we expect. After discussing with experts in the field who have experience integrating scRNA-seq data with other *in situ* methods, the conclusion was that the current data sets do not have enough rolonies per cell to integrate with scRNA-seq. We look into methods of improving DARTFISH in Chapter 3.

2.6 Appendix to Chapter 2

Table 4: List of gene markers used for cell type classification by plurality

Astrocytes	Endothelial	Microglia	Neurons	Oligodendrocytes
HPSE2	PAPSS2	SFMBT2	ITGA8	CTNNA3
GFRA1	RBM20	C10orf11	KIAA1217	OPALIN
DKK3	TESC	SWAP70	GAD2	NKX6-2
FIBIN	ATP10A	SLC15A3	CRTAC1	KLHL1
ABCC9	ITGA11	SLC25A45	ADRA2A	CLMN
FREM2	HCN4	NPAS4	NELL1	CAPN3
RYR3	MYH11	UNC93B1	SLC17A6	FA2H
ACOT11	RPH3AL	SLCO2B1	GPR83	ASPA
CDC14A	GRAP	IL10RA	HTR3A	EVI2A
VAV3	PECAM1	KCNJ5	SYT10	CNP
ANKRD35	ATP8B1	RASSF3	TAC3	GREB1L
EMID1	RNF152	TMEM119	TRHDE	MBP
LTBP1	GDF15	SELPLG	PAH	MAG
SERPINE2	COL24A1	GPR183	CUX2	PPP1R14A
SLC6A11	RCSD1	RAB20	NOS1	TMEM125
LRRC3B	NOSTRIN	PLD4	SRRM4	NHLH2
FEZF2	CCDC141	TNFAIP8L3	POSTN	TMEM63A
GABRG1	HPGD	HS3ST4	PCDH8	MAL
PDLIM5	SLC12A7	ITGAM	PCDH20	ERMN
ARSJ	ECSCR	ZFHX3	SYNDIG1L	PDE1A
SLC7A11	DSP	MAF	BCL11B	AOX1
DCHS2	LAMA2	CCL3	MEG3	MOBP
MYO10	GNG11	CCL4	MAGEL2	CLDN11
MCC	TEK	HMHA1	CELF6	PLD1
RNF182	ANXA1	GNA15	CRABP1	APOD
MDGA1		VAV1	CALB2	PDGFRA
HGF		RASAL3	NECAB2	SPOCK3
DLC1		LRRC25	ATP2C2	ENPP6
TOX		TYROBP	P2RX5	SEMA5A
SNTB1		TGFB1	RGS9	MOG
ADCY8		C5AR1	SDK2	THEMIS
CNTNAP3		CD37	GNAL	CREB5
RORB		CD33	CELF4	ANKRD18A
PCSK5		SLC2A5	SLC17A7	GJB1
NTNG2		TNFRSF1B	GPR153	PLP1
		C1QA	GRIK3	
		C1QC	NTNG1	
		C1QB	TNNT2	

		LAPTM5	SYT2	
		CSF3R	GNG4	
		ZC3H12A	GREM2	
		OLFML3	KMO	
		SYT6	PLD5	
		CTSS	PDYN	
		FCGR2B	LAMP5	
		OLFML2B	VSTM2L	
		RGS1	SLC32A1	
		NLRP3	TSHZ2	
		RIN2	PCP4	
		CRYBB1	PVALB	
		OSM	HPCAL1	
		NCF4	GALNT14	
		NFAM1	BCL11A	
		LIMS1	TACR1	
		IL1A	SMYD1	
		NR4A2	LYPD6B	
		SLC11A1	TBR1	
		INPP5D	KCNH7	
		CX3CR1	GAD1	
		CCR1	DLX1	
		CCR5	DLX2	
		CCRL2	PPP1R1C	
		TLR9	COL5A2	
		P2RY13	MARCH4	
		TLR2	IGFBP5	
		CSF1R	SLC6A1	
		RASGEF1C	ZNF385D	
		CD83	CCK	
		TNF	CACNA2D2	
		TREM2	SYNPR	
		NCF1	ROBO2	
		FOXP2	PLCXD2	
		IRF5	CLSTN2	
		MAMDC2	GPR149	
		GSN	SST	
		TLR7	RGS12	
			BEND4	
			SLC10A4	

			KIT	
			NMU	
			FRAS1	
			CXXC4	
			COL25A1	
			NDNF	
			TRPC3	
			CDH9	
			PLCXD3	
			PCSK1	
			TRPC7	
			ADTRP	
			LHFPL5	
			HCRTR2	
			MRAP2	
			CNR1	
			HS3ST5	
			MYB	
			VIP	
			SP8	
			NPY	
			NEUROD6	
			ZNF804B	
			DYNC1I1	
			DLX6	
			DLX5	
			TAC1	
			TMEM130	
			RELN	
			PNOC	
			BHLHE22	
			SULF1	
			CALB1	
			FREM1	
			ADAMTSL1	
			GRIN3A	
			LHX6	
			VAV2	
			OLFM1	
			GLRA2	

			IL1RAPL2	
			TRPC5	
			L1CAM	

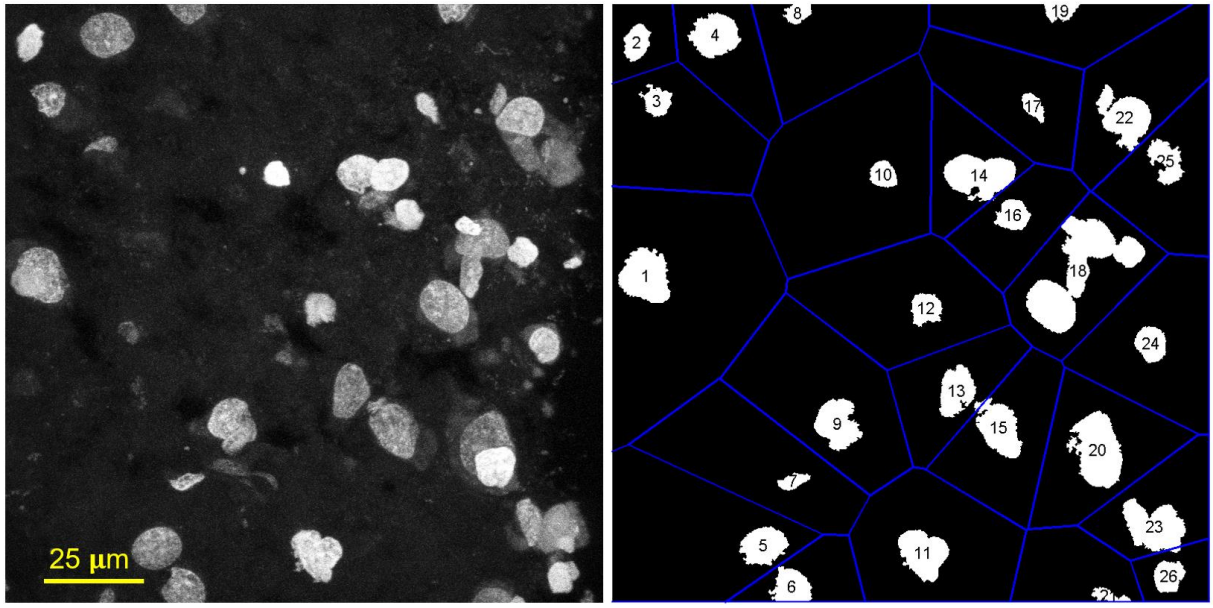


Figure S15: Cell segmentation using DRAQ5-stained nuclei image. Left: Maximum intensity projected image stack of DRAQ5 image. There are many nuclei with no separation or overlap, making accurate segmentation challenging. Right: Labeled nuclei after thresholding and Voronoi partitioning without user intervention. Cells like #14 need the user to apply the watershed algorithm to separate the nuclei in two.

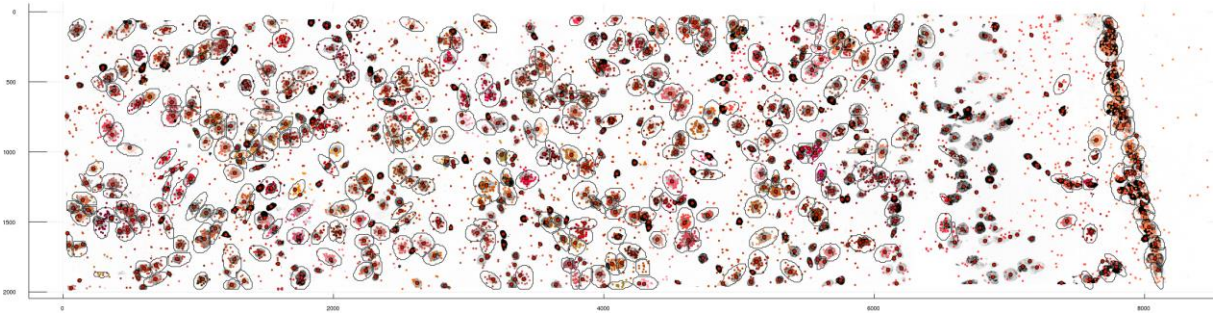


Figure S16: Cell segmentation of mouse visual cortex.

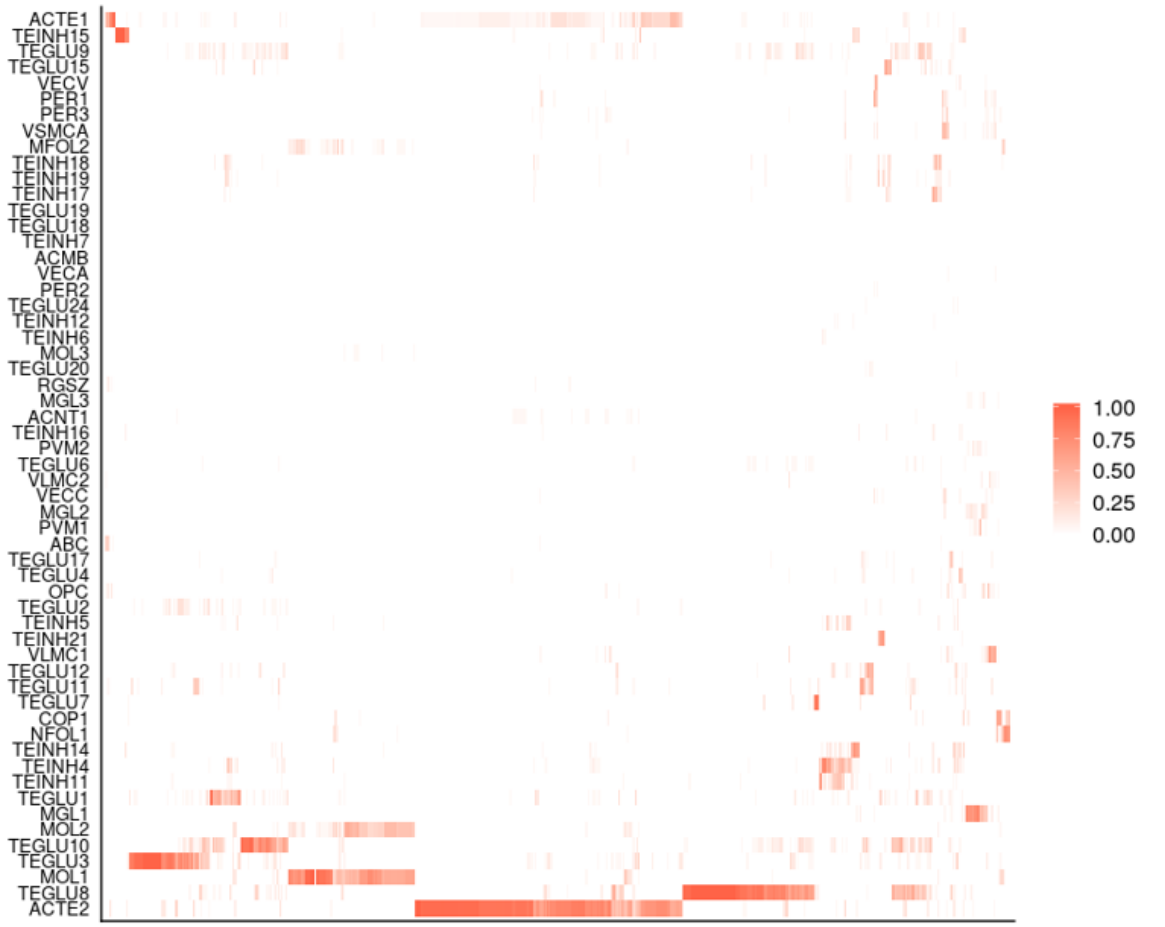


Figure S17: Similarity of each DARTFISH cell to reference scRNA-seq cell types using kNN classifier.

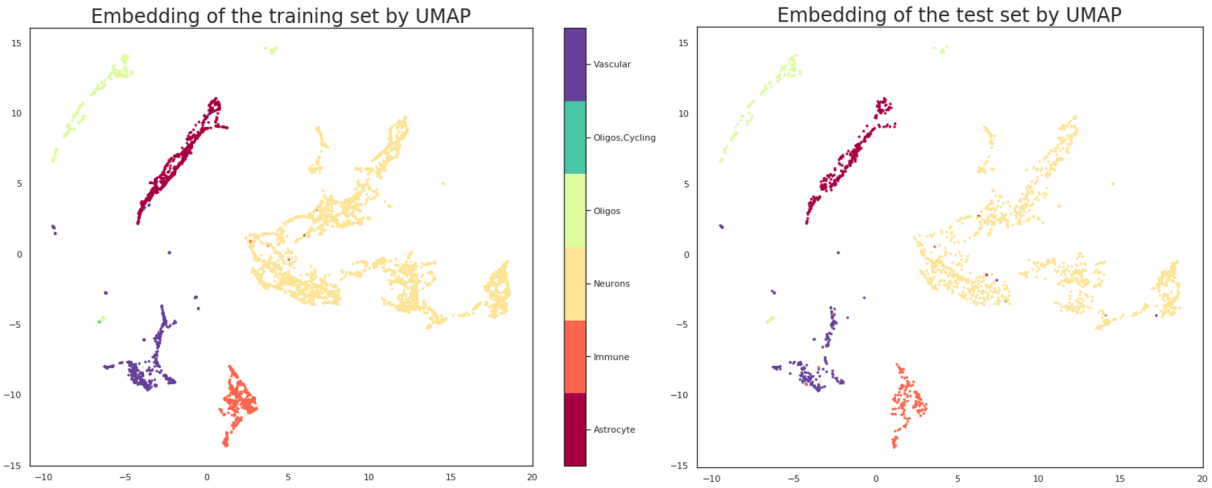


Figure S18: UMAP embedding of scaled scRNA-seq cells with only 375 genes randomly split into training and test sets. Trained SVC and kNN classifiers on embedded test set had 99.1% accuracy.

2.7 Acknowledgement for Chapter 2

Justin Dang for initial development of nuclei segmentation in MATLAB. Kian Kalhor for improving image registration with SimpleElastix and cell segmentation. Yan Wu for his cellMapper R package used for mapping scRNA-seq clusters to DARTFISH cells.

Chapter 2 is coauthored with Wu, Yan; Kalhor, Kian; and Zhang, Kun. The dissertation author was the primary author of this chapter.

CHAPTER 3. IMPROVING THE DETECTION RATE OF DARTFISH

3.1 Abstract

During the development of DARTFISH, the value in spatial transcriptome data has led to many *in situ* RNA methods being invented and improved. Many of them have been discussed already. At the same time we recognized in Chapter 2 that DARTFISH needed to have higher sensitivity for it to be integrated with single-cell RNA-sequencing data and useful as a tool for creating a reference Human Cell Atlas. DARTFISH has advantages in being highly multiplexed, having high amplification, and having a hybridization-based barcode. To improve the sensitivity, we explored methods that would obviate *in situ* reverse transcription, which is a bottleneck for detection rate. First we tried using PBCV-1 DNA ligase but ultimately found it lacked the specificity of our original DNA ligase. We then tested using SNAIL probes in place of padlock probes. The results have been an improvement of at least five-fold and are reported here.

3.2 Introduction

Fundamentally, DARTFISH needs to use a hybridization-based barcode for decoding because it has the advantage of being fast and simple for adoption. Biology labs that want to use an *in situ* RNA method to localize RNA or spatially map single cells would find value in a method that does not require expertise to set up an automated temperature-regulated fluidics system and budget to purchase a high performance fluorescence microscope. The imaging for DARTFISH in its current form can be done by manually pipetting on a microscope in a core facility. We also determined RCA to be the best signal amplification for use in tissues with high background so that the method is not limited by sample type. But given the complexity and

heterogeneity of human brain tissue, DARTFISH using padlock probes does not have enough rolonies per cell to map scRNA-seq cells.

One bottleneck in the process of converting RNA to rolonies has been *in situ* reverse transcription. The RNA is crosslinked by formaldehyde to proteins like RNA-binding proteins, which makes the transcript less accessible to primers and disrupts the reverse transcriptase as it polymerizes cDNA. Reverse transcription was necessary though because the DNA ligases used for padlock probe capture require a DNA splint oligo. In Chapter 1 we did our best to optimize this step by trying various primers and reverse transcriptases.

In 2016, we learned of a commercially available DNA ligase that could ligate ssDNA on an RNA splint. SplintR (New England BioLabs) is PBCV-1 DNA Ligase from *Chlorella* virus and reportedly has a hundred times higher activity than T4 DNA ligase (Jin, Vaud, Zhelkovsky, Posfai, & McReynolds, 2016). It opened the possibility of using padlock probes to capture RNA directly without having to reverse transcribe it to cDNA. One concern we had was that SplintR is not thermostable at high temperatures like Ampligase, which we had been using. That meant we could not raise the temperature to control probe hybridization specificity. We rationalized that some background non-specific noise would be acceptable if the number of rolonies was high enough, and that we could still use higher temperatures during probe hybridization if we made the ligation a separate step after washing away non-hybridized probes. The wash step becomes critical because any probes that did not hybridize but also remained in the tissue would become non-specific rolonies. This chapter will cover the testing of SplintR for use in DARTFISH and why it ultimately failed.

In 2018, a new type of probe called a SNAIL probe was published as part of the STARmap method (X. Wang et al., 2018). SNAIL probes solved the *in situ* reverse transcription

issue by cleverly adding another ssDNA oligo splint into the reaction for the terminal ends of the padlock probe to hybridize to for ligation. This way the traditional DNA ligases could be used. The target specific regions are no longer at the terminal ends of the padlock probe, but one is on the splint and the other is on the padlock such that both need to hybridize to the target for ligation to occur (see Figure S19). Since captured SNAIL probes are still circular ssDNA they can still be rolling circle amplified. Additionally, because they have a padlock probe component, we can insert our hybridization-based barcode into the backbone of the padlock and still carry out decoding with the exact same procedure. The second half of this chapter will cover our progress in testing DARTFISH with SNAIL probes in mouse and human cortical sections.

3.3 Results and Discussion

3.3.1 SplintR *in vitro* tests

Our initial SplintR testing was done using thousands of padlock probes in a probe set that was designed for capturing FISSEQ rolonies *in situ*. Since FISSEQ rolonies and RNA have the same strandedness, the same padlock probes have sequence complementary to both. The padlock probes were used to capture Universal Human Reference RNA (UHRR 740000-41) in a tube with 50 U SplintR circularizing the padlocks. The hybridization was done first by heating up the probe and template to 95 °C and slowly decreasing the temperature, as described in Chapter 1 for *in* vitro capture. Then SplintR was added along with the appropriate reaction buffer and incubated at room temperature overnight. Then RNase H and Riboshredder were added to digest the RNA, and Exonuclease I and Exonuclease III were added to digest non-ligated padlock probes. The remaining circular ssDNA was then quantified by qPCR using a pair of primers that hybridize to the common linker sequence on the backbone.

A range of RNA and padlock probe concentrations were used but the ratio between the two were kept constant. 8 ng of probes were used for every 30 ng of RNA. Quantitative PCR revealed that even without RNA template to serve as splints in the ligation reaction, padlock probes could still be ligated. The difference between C_t of the padlock probes capturing RNA and the C_t of the no template control (NTC) increased with lower padlock probe concentration. The qPCR curves are seen in Figure 17.

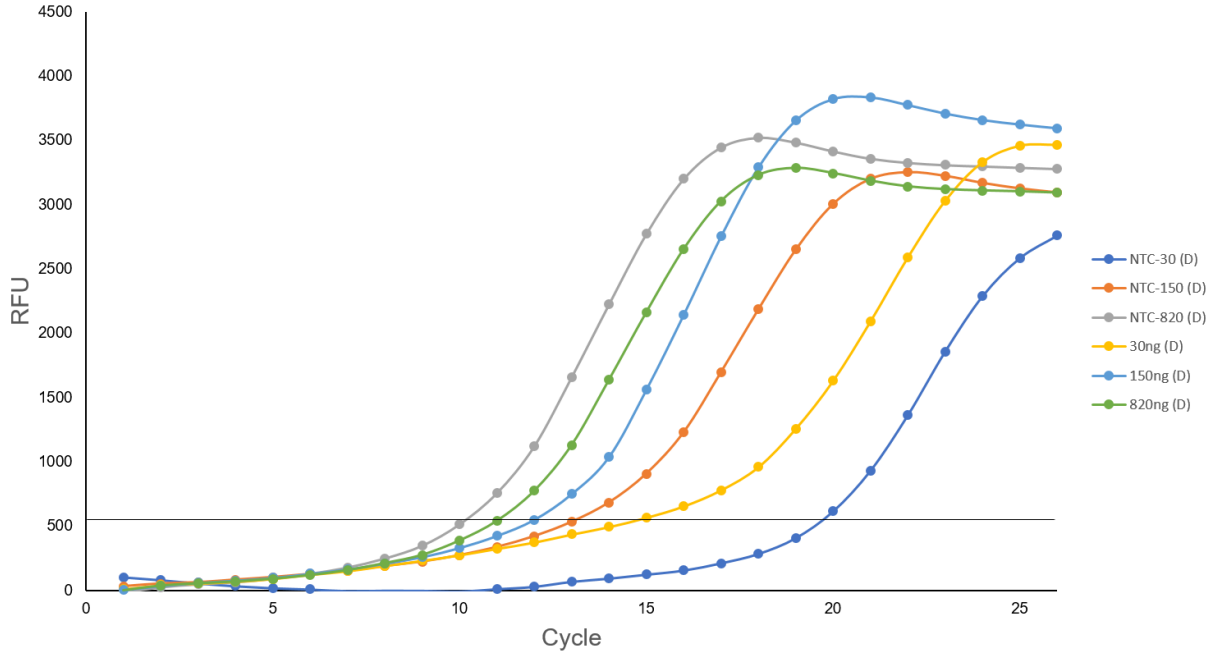


Figure 17: The qPCR curves for SplintR *in vitro* test at various padlock probe and RNA concentrations. The legend labels the curves by amount of RNA in the reaction. NTC means RNA was absent but all other components including padlock probe were added as if RNA were present.

The ligation of padlock probes in NTC samples means SplintR can ligate ssDNA on a DNA splint as well, although literature says it should favor RNA splints (Jin et al., 2016). The improvement of ΔC_t at lower padlock probe concentrations further supports the hypothesis that the padlock probes are using each other as splints. The ΔC_t of 5 cycles for 8 ng padlock probes is approximately a 30-fold difference and we deemed that to be promising. This test highlighted the importance of washing between hybridization and ligation steps *in situ* because padlock probes hybridized to other probes should be washed away, therefore greatly reducing the noise.

Next we tried to improve specificity by using formamide, dimethylformamide, Betaine, or Extreme Thermostable Single-Stranded Binding Protein (ET SSB) in the reaction. Both of these alter the thermodynamics such that it is less favorable to hybridize unless there is significant sequence complement. Again, we used qPCR to measure the quantity of captured

padlock probes and aimed to find a protocol that maximized the amount of captured padlock probes in the template sample and maximize the difference between the sample and NTC. Ampligase and a DNA template were used as a reference. As shown in Figure 18, including 7.5% formamide in the capture reaction had the best combination of specificity and sensitivity. ET SSB is not shown because it had poor results in other experiments.

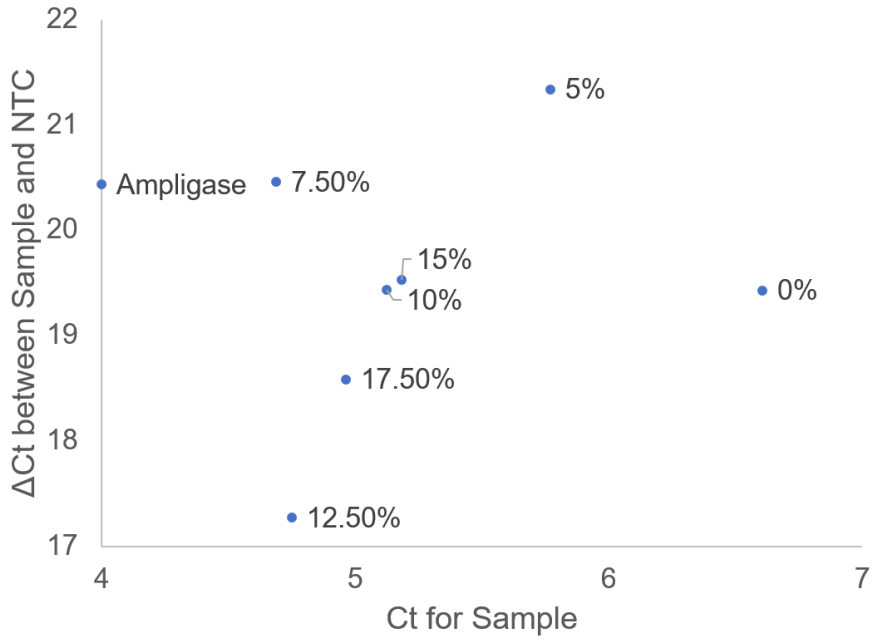


Figure 18: Testing formamide concentrations in padlock probe capture with SplintR to find the best condition for sensitivity and specificity. qPCR was used to quantify amount of ligated padlocks in the sample and NTC. Higher sensitivity is lower on the x-axis and higher specificity is higher on the y-axis.

3.3.2 SplintR *in situ*

Implementing padlock probe capture with SplintR *in situ* required splitting the overnight padlock probe capture reaction into an overnight hybridization at 55 °C and then a 30 minute ligation reaction with SplintR and 7.5% formamide at 37 °C (see Figure S20). In between those two steps, washing with 1X PBS was done to remove padlock probes that weren't hybridized to fixed RNA. Another concern was that if washing was done at a low temperature then the

padlock probes would non-specifically hybridize during the washes and wouldn't end up being removed. So the washes were done with heated 1X PBS and on a hot plate both at 55 °C.

Using SplintR had a greater than five-fold increase in rolonies (see Figure S21) when we tested DARTFISH with SplintR and DARTFISH with Ampligase in parallel on adjacent human cortical sections. However, when looking at genes detected in the cortex versus white matter, only the Ampligase protocol showed a difference in genes detected (see Figure 19). Despite great effort to improve specificity we decided to move on from SplintR.

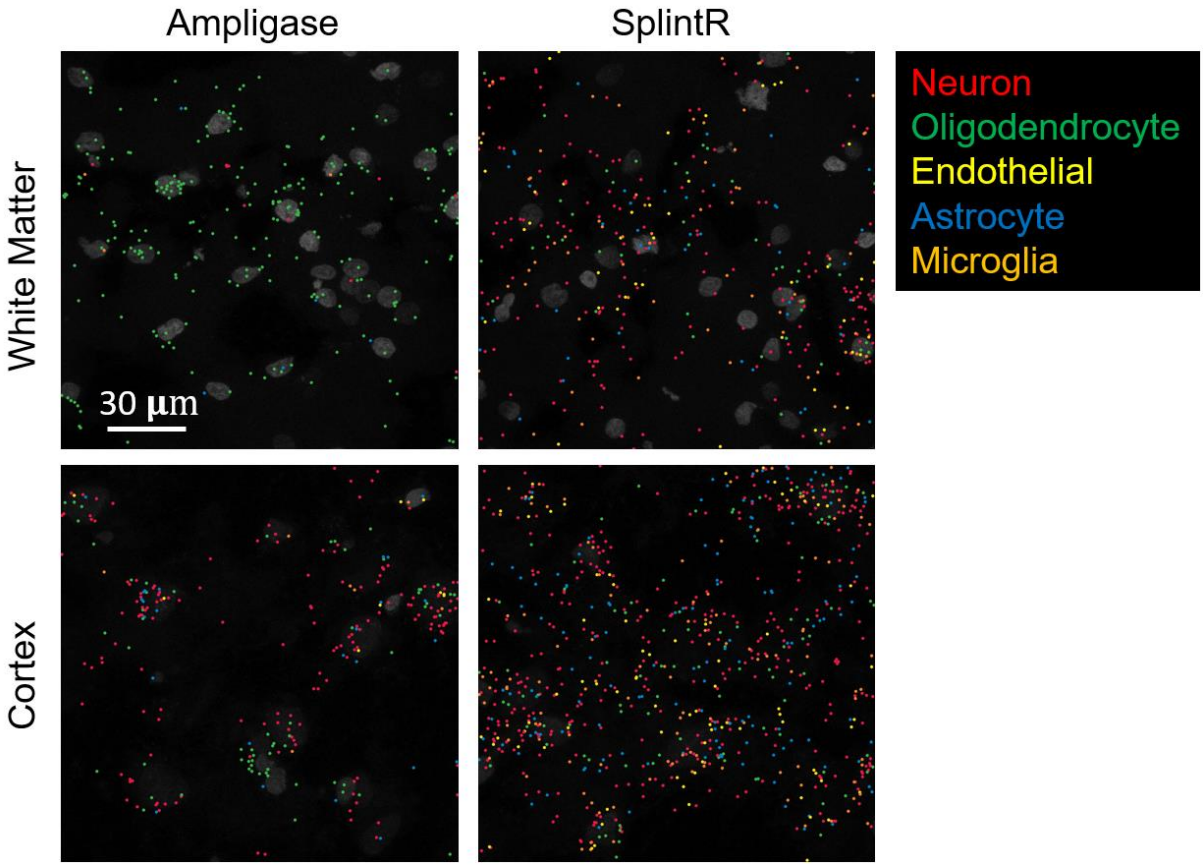
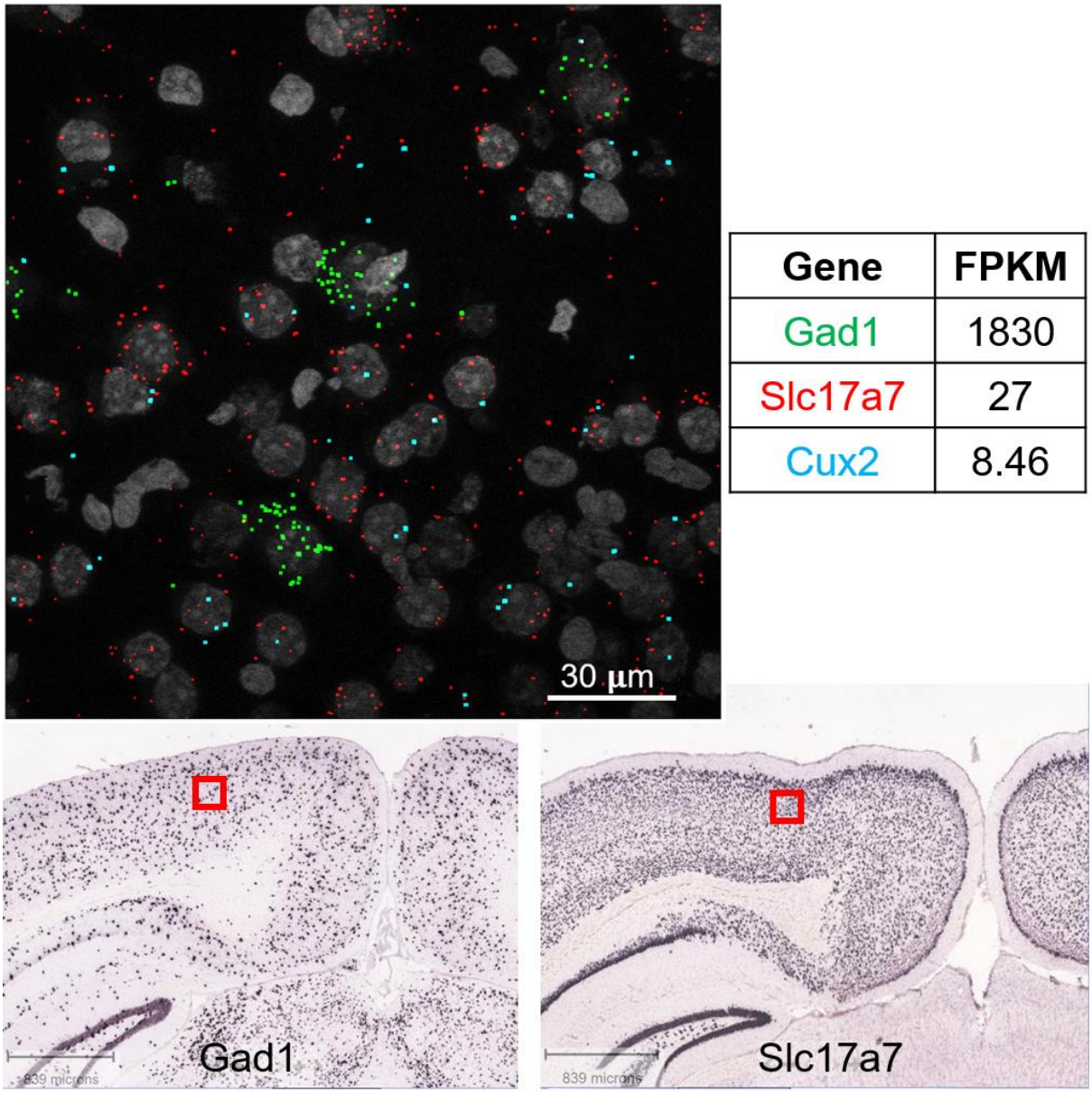


Figure 19: Gene marker rolonies in human brain sections from DARTFISH with SplintR versus Ampligase. DARTFISH with Ampligase shows the majority of rolonies in white matter are gene markers for oligodendrocytes and in cortex they are more mixed. With SplintR, there is no distinction between rolonies in white matter and cortex.

3.3.3 SNAIL Probes

The process of designing SNAIL probes was not dissimilar to padlock probes. SNAIL probes have a skeleton, which include the sequences where the two oligos hybridize to each other, and two variable sequences that hybridize to the RNA target. The variable sequences have 1-3 base pair gap on the RNA target. We also replace the sequencing barcode on the backbone of the circular oligo with our hybridization-based barcode. This significantly increases the length of the circular oligo because the sequencing barcode is 18 bases and a hybridization barcode can be 60 or 80 bases. This both significantly increases the cost of the oligo and could have an effect on circularization efficiency.

The pilot experiment was to test the compatibility of SNAIL probes with longer hybridization-based barcodes. We used mouse probe sequences from the STARmap paper but replaced the sequencing barcode with our 80 base hybridization-based barcode that made the oligo length a total of 119 bases. The genes we chose were Gad1, Slc17a7, and Cux2 because of their well-defined specificity that allows us to identify non-specific rolonies. Gad1 and Slc17a7 are inhibitory and excitatory neuron markers, respectively, and should rarely be seen in the same cells. Cux2 is enriched in upper layers of the cortex and also a gene marker for a subset of excitatory neurons. We fabricated rolonies in a 10 μm coronal mouse brain section following the STARmap protocol and imaged in the upper cortical layer approximately 1.5 mm from the medial longitudinal fissure. As shown in Figure 20, the abundance and location of rolonies matched expectations: Gad1 and Slc17a7 rolonies rarely overlapped in the same cell; Gad1 positive interneurons were sparser than Slc17a7 excitatory neurons, which is also seen in ISH data; The relative abundance of rolonies per cell matched FPKM values from scRNA-seq data; And Cux2 was only found in cells that were Slc17a7 positive.



Gene	FPKM
Gad1	1830
Slc17a7	27
Cux2	8.46

Figure 20: SNAIL probes with 80 base hybridization-based barcode decoded in upper layer of mouse cortex (red box in ISH images). Rolonies for three genes show good specificity with very little overlap of Gad1 and Slc17a7. The abundance of rolonies per cell correlates with FPKM from scRNA-seq and the density of Gad1+ and Slc17a7+ cells matches ISH images.

For SNAIL probes in human tissue we needed to design the variable sequence that binds to the RNA target. We chose the sequences from hybridization arms of the best padlock probes in Chapter 1. Although the ppDesigner algorithm we used for padlock probes is different from

Picky2.2 used by STARmap, the general principle of identifying two sequences specific to the target is the same. They both consider off-target binding, sequence length, GC content, and melting temperature. Four probes were designed for each gene and a 60 base barcode was added. Like in the mouse experiment, GAD1 and SLC17A7 were chosen and the other genes had layer specificity.

The results of experiments in human middle temporal gyrus cortex (MTG) were mixed. For an unknown reason, many areas in the cortex had very few rolonies. But in areas that had rolonies, the number per cell and specificity were better than anything we had seen with padlock probes. Figure 21 shows an example of a field-of-view where a number of cells had rolonies. A couple GAD1+ had greater than 20 GAD1 rolonies, where with padlock probes we saw at most 4 GAD1 rolonies in a single cell. To measure the improvement in detection rate of using SNAIL probes, we compared the rolony count for GAD1, SLC17A7, and RORB in MTG cortex to that of occipital cortex using padlock probes in Chapter 1. We took the average gene rolony count per cell for cells that had at least one rolony of that gene and saw at least five-fold increase in number of rolonies (see Figure 21C). In the MTG we imaged 0.1 mm^2 of the cortex and in the occipital cortex we imaged 0.75 mm^2 across the cortex.

In the MTG section using DARTFISH with SNAIL probes we also saw a GAD1+ cell with 5 PVALB rolonies. PVALB is the gene for paravalbumin and is a marker for a canonical interneuron subtype. We also saw a cell with SLC17A7, RORB, and FOXP2. From single-nuclei RNA-seq of human neurons, RORB and FOXP2 are markers for Ex3d, Ex4, or Ex5 excitatory neuronal subtypes and SLC17A7 confirms that the cell is an excitatory neuron (Lake et al., 2016).

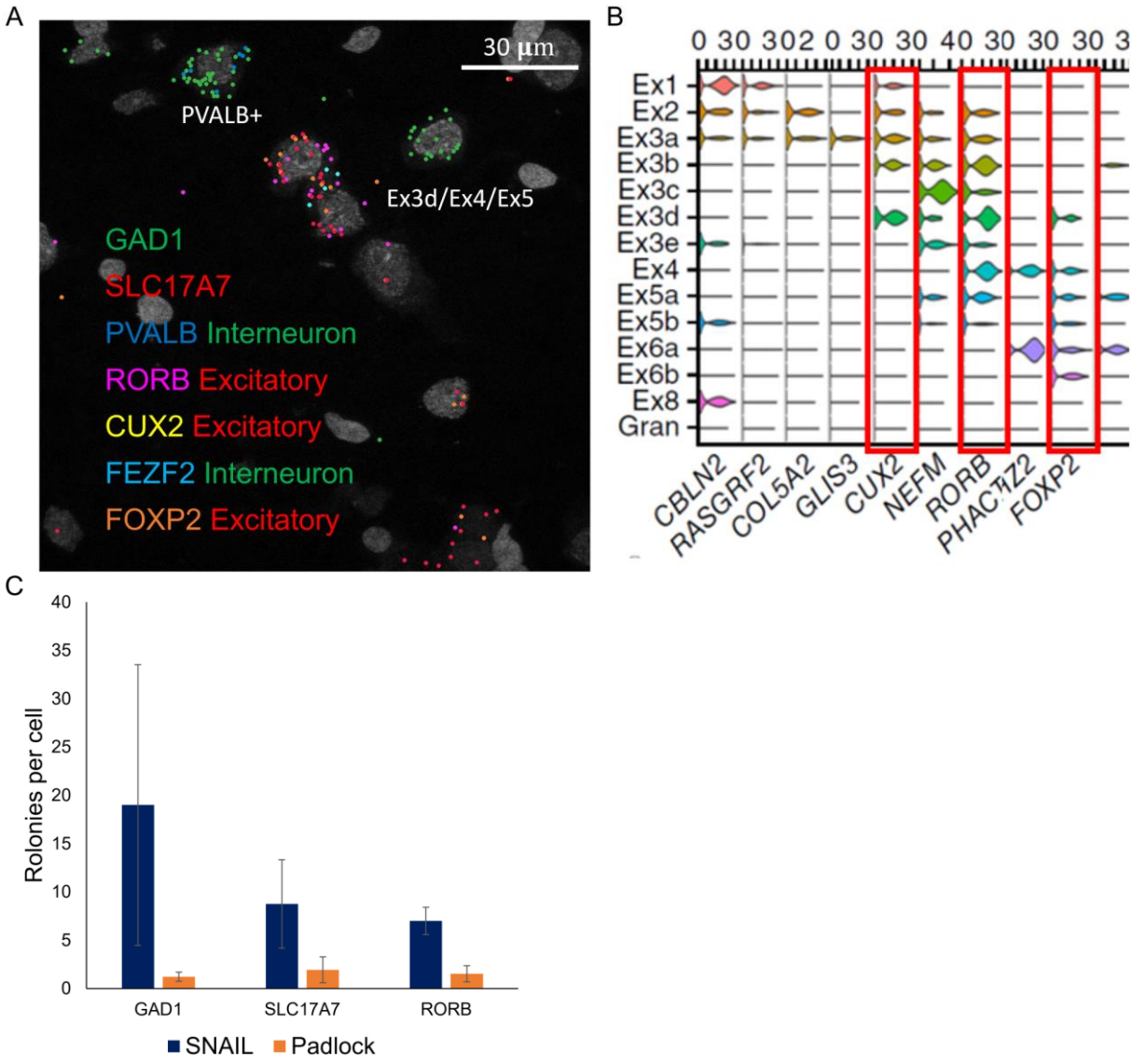


Figure 21: DARTFISH with SNAIL probes for 7 genes in human MTG cortical section can identify neuronal subtypes. (A) The SNAIL probes are highly specific, with only interneuron subtype markers in GAD1+ cells and only excitatory neuron subtype markers in SLC17A7+ cells. In this area we were able to identify a PVALB+ interneuron and an excitatory neuron subtype from the middle layers of the cortex. (B) Violin plots of expression values for excitatory neuron subtype marker genes. (C) A gene-by-gene comparison of average rolony counts in cells with at least one rolony between DARTFISH with SNAIL probes and DARTFISH with padlock probes.

In conclusion, DARTFISH results with SNAIL probes have shown promise in initial experiments targeting a handful of genes. However, more troubleshooting needs to be done to make the method more robust for consistent results. Also, not shown here are results from experiments that used SNAIL probes targeting 52 genes. Those rolonies showed no layer specificity, similar to the SplintR *in situ* experiments. We suspect it is due to some barcodes possibly being used as non-specific binding sites for probes to anneal to and become ligated.

3.4 Appendix to Chapter 3

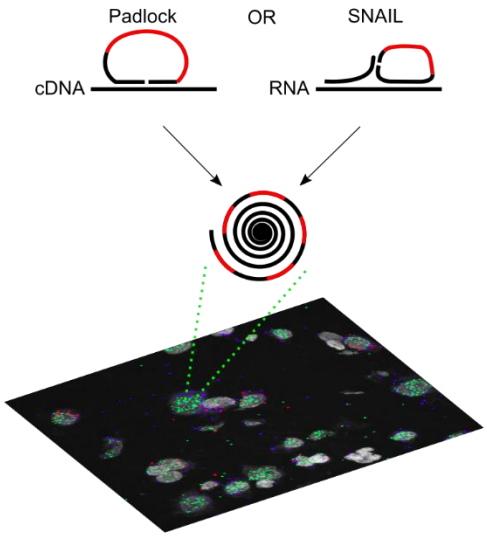


Figure S19: Diagram of padlock probes and SNAIL probes capturing their respective targets. Both can accommodate a barcode (in red) and be rolling circle amplified *in situ*.

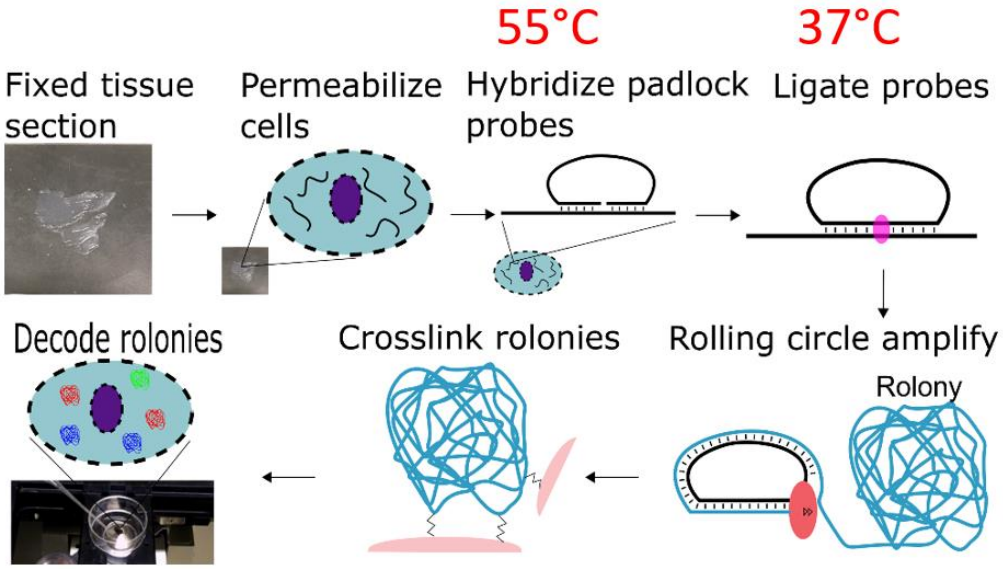


Figure S20: Experimental protocol for DARTFISH using SplintR ligase. Reverse transcription is not needed, shortening the protocol by a day and removing a low efficiency bottleneck in the protocol. Because SplintR is not thermostable the hybridization and ligation must be done in two steps at two different temperatures. The lower temperature of ligation means washing away padlock probes that did not hybridize at 55 °C is critical so they do not non-specifically capture at 37 °C.

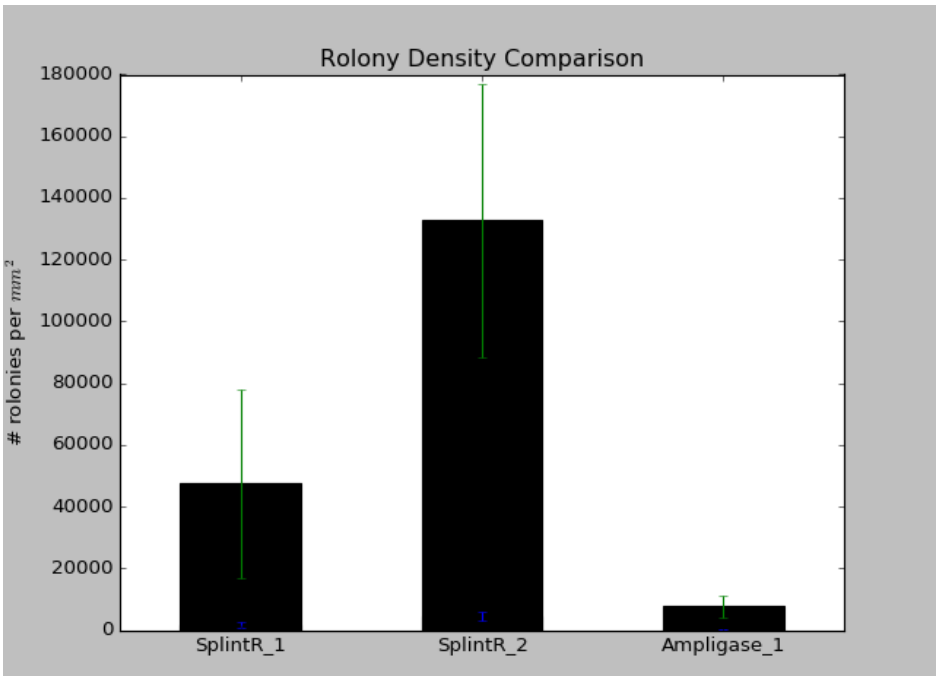


Figure S21: Comparing rolony counts between DARTFISH with Ampligase and DARTFISH with SplintR. SplintR_1 and Ampligase_1 used probe sets that target the same region on genes but are reverse complementary because one captures RNA and the other cDNA. SplintR_2 used a different probe set.

3.6 Acknowledgement for Chapter 3

I am extremely grateful to Dinh Diep and Richard Que for their assistance in SNAIL probe experiments. Dinh was generous to volunteer her time designing and preparing the initial probe set as well as writing the PickInSituProbe software. Richard has been a great partner on this project since he joined. His experiments with SNAIL probes in kidney sections and troubleshooting probe design were invaluable.

Chapter 3 is coauthored with Diep, Dinh; Que, Richard; and Zhang, Kun. The dissertation author was the primary author of this chapter.

REFERENCES

Achim, K., Pettit, J. B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., & Marioni, J. C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, *33*(5), 503–509. https://doi.org/10.1038/nbt.3209

Beliveau, B. J., Kishi, J. Y., Nir, G., Sasaki, H. M., Saka, S. K., Nguyen, S. C., … Yin, P. (2018). OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(10), E2183–E2192. https://doi.org/10.1073/pnas.1714530115

Bhakdi, S., & Thaicharoen, P. (2018). Easy Employment and Crosstalk-Free Detection of Seven Fluorophores in a Widefield Fluorescence Microscope. *Methods and Protocols*, *1*(2), 20. https://doi.org/10.3390/mps1020020

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., … Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, *361*(6409), 1380–1385. https://doi.org/10.1126/science.aau0730

Chen, S., Lake, B. B., & Zhang, K. (2019). Linking transcriptome and chromatin accessibility in nanoliter droplets for single-cell sequencing. *BioRxiv*, 692608. https://doi.org/10.1101/692608

Chung, K., Wallace, J., Kim, S. Y., Kalyanasundaram, S., Andalman, A. S., Davidson, T. J., … Deisseroth, K. (2013). Structural and molecular interrogation of intact biological systems. *Nature*, *497*(7449), 332–337. https://doi.org/10.1038/nature12107

Cui, Y., Zheng, Y., Liu, X., Yan, L., Fan, X., Yong, J., … Tang, F. (2019). Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. *Cell Reports*, *26*(7), 1934-1950.e5. https://doi.org/10.1016/j.celrep.2019.01.079

David, L. E., Fowler, C. B., Cunningham, B. R., Mason, J. T., & O'Leary, T. J. (2011). The effect of formaldehyde fixation on RNA: Optimization of formaldehyde adduct removal. *Journal of Molecular Diagnostics*, *13*(3), 282–288. https://doi.org/10.1016/j.jmoldx.2011.01.010

Dempsey, G. T., Vaughan, J. C., Chen, K. H., Bates, M., & Zhuang, X. (2011). Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nature Methods*, *8*(12), 1027–1040. https://doi.org/10.1038/nmeth.1768

Diep, D., Plongthongkum, N., Gore, A., Fung, H. L., Shoemaker, R., & Zhang, K. (2012). Library-free methylation sequencing with bisulfite padlock probes. *Nature Methods*, *9*(3), 270–272. https://doi.org/10.1038/nmeth.1871

Evanko, D. (2004). Hybridization chain reaction. *Nature Methods*, *1*(3), 186–187.

https://doi.org/10.1038/nmeth1204-186a

Femino, A. M., Fay, F. S., Fogarty, K., & Singer, R. H. (1998). Visualization of single RNA transcripts in situ. *Science*, *280*(5363), 585–590. https://doi.org/10.1126/science.280.5363.585

Fontenete, S., Carvalho, D., Guimarães, N., Madureira, P., Figueiredo, C., Wengel, J., & Azevedo, N. F. (2016). Application of locked nucleic acid-based probes in fluorescence in situ hybridization. *Applied Microbiology and Biotechnology*, *100*(13), 5897–5906. https://doi.org/10.1007/s00253-016-7429-4

Gall, J. G. (2016, April 1). The origin of in situ hybridization - A personal history. *Methods*. Academic Press Inc. https://doi.org/10.1016/j.ymeth.2015.11.026

Gunderson, K. L., Kruglyak, S., Graige, M. S., Garcia, F., Kermani, B. G., Zhao, C., … Chee, M. S. (2004). Decoding randomly ordered DNA arrays. *Genome Research*, *14*(5), 870–877. https://doi.org/10.1101/gr.2255804

Gupta, D., Middleton, L. P., Whitaker, M. J., & Abrams, J. (2003). Comparison of fluorescence and chromogenic in situ hybridization for detection of HER-2/neu oncogene in breast cancer. *American Journal of Clinical Pathology*, *119*(3), 381–387. https://doi.org/10.1309/P40P2EAD42PUKDMG

Harrington, M. G., Fonteh, A. N., Biringer, R. G., Hühmer, A. F. R., & Cowan, R. P. (2006). Prostaglandin D synthase isoforms from cerebrospinal fluid vary with brain pathology. *Disease Markers*, *22*(1–2), 73–81.

Holmes, D. L., & Stellwagen, N. C. (1991). Estimation of polyacrylamide gel pore size from Ferguson plots of normal and anomalously migrating DNA fragments. I. Gels containing 3 % N, N′-methylenebisacrylamide. *Electrophoresis*, *12*(4), 253–263. https://doi.org/10.1002/elps.1150120405

Huber, D., Voith von Voithenberg, L., & Kaigala, G. V. (2018, November 1). Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*. Elsevier B.V. https://doi.org/10.1016/j.mne.2018.10.006

Jilbert, A. R., Burrell, C. J., Gowans, E. J., & Rowland, R. (1986). Histological aspects of in situ hybridization - Detection of poly(A) nucleotide sequences in mouse liver sections as a model system. *Histochemistry*, *85*(6), 505–514. https://doi.org/10.1007/BF00508433

Jin, J., Vaud, S., Zhelkovsky, A. M., Posfai, J., & McReynolds, L. A. (2016). Sensitive and specific miRNA detection method using SplintR Ligase. *Nucleic Acids Research*, *44*(13), e116. https://doi.org/10.1093/nar/gkw399

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., & Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods*, *10*(9), 857–860. https://doi.org/10.1038/nmeth.2563

Kebschull, J. M., Garcia da Silva, P., Reid, A. P., Peikon, I. D., Albeanu, D. F., & Zador, A. M. (2016). High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA. *Neuron*, *91*(5), 975–987. https://doi.org/10.1016/j.neuron.2016.07.036

Kim, J., & Eberwine, J. (2010). RNA: State memory and mediator of cellular phenotype. *Trends in Cell Biology*, *20*(6), 311–318. https://doi.org/10.1016/j.tcb.2010.03.003

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., … Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., … Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, *352*(6293), 1586–1590. https://doi.org/10.1126/science.aaf1204

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Terry, R., … Church, G. M. (2015). Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, *10*(3), 442–458. https://doi.org/10.1038/nprot.2014.191

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., … Church, G. M. (2014). Highly multiplexed subcellular RNA sequencing in situ. *Science*, *343*(6177), 1360–1363. https://doi.org/10.1126/science.1250212

Levesque, M. J., & Raj, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nature Methods*, *10*(3), 246–248. https://doi.org/10.1038/nmeth.2372

Lubeck, E., & Cai, L. (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods*, *9*(7), 743–748. https://doi.org/10.1038/nmeth.2069

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M., & Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*. Nature Publishing Group. https://doi.org/10.1038/nmeth.2892

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, *113*(39), 11046–11051. https://doi.org/10.1073/pnas.1612826113

Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., & Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, *5*(10),

877–879. https://doi.org/10.1038/nmeth.1253

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., … Yosef, N. (2017). The human cell atlas. *ELife*, *6*. https://doi.org/10.7554/eLife.27041

Reyfman, P. A., Walter, J. M., Joshi, N., Anekalla, K. R., McQuattie-Pimentel, A. C., Chiu, S., … Misharin, A. V. (2019). Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine*, *199*(12), 1517–1536. https://doi.org/10.1164/rccm.201712-2410OC

Rigby, P. W. J., Dieckmann, M., Rhodes, C., & Berg, P. (1977). Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *Journal of Molecular Biology*, *113*(1), 237–251. https://doi.org/10.1016/0022-2836(77)90052-3

Rouhanifard, S. H., Mellis, I. A., Dunagin, M., Bayatpour, S., Jiang, C. L., Dardani, I., … Raj, A. (2019, January 1). ClampFISH detects individual nucleic acid molecules using click chemistry–based amplification. *Nature Biotechnology*. Nature Publishing Group. https://doi.org/10.1038/nbt.4286

Sealey, P. G., Whittaker, P. A., & Southern, E. M. (1985). Removal of repeated sequences from hybridisation probes. *Nucleic Acids Research*, *13*(6), 1905–1922. https://doi.org/10.1093/nar/13.6.1905

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*. https://doi.org/10.1038/nbt1486

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., … Frisén, J. (2016, July 1). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.aaf2403

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., … Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Sylwestrak, E. L., Rajasethupathy, P., Wright, M. A., Jaffe, A., & Deisseroth, K. (2016). Multiplexed Intact-Tissue Transcriptional Analysis at Cellular Resolution. *Cell*, *164*(4), 792–804. https://doi.org/10.1016/j.cell.2016.01.038

Titford, M. (2009). Progress in the development of microscopical techniques for diagnostic pathology. *Journal of Histotechnology*, *32*(1), 9–19. https://doi.org/10.1179/his.2009.32.1.9

Trapnell, C. (2015, October 1). Defining cell types and states with single-cell genomics. *Genome Research*. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/gr.190595.115

Tsanov, N., Samacoits, A., Chouaib, R., Traboulsi, A. M., Gostan, T., Weber, C., … Mueller, F. (2016). SmiFISH and FISH-quant - A flexible single RNA detection approach with super-resolution capability. *Nucleic Acids Research*, *44*(22). https://doi.org/10.1093/nar/gkw784

Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., … Deisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, *361*(6400). https://doi.org/10.1126/science.aat5691

Wang, Y. (2018, October 2). Seeing the future of histotechnology through its history. *Journal of Histotechnology*. Taylor and Francis Ltd. https://doi.org/10.1080/01478885.2018.1527590

Weinstein, J. A., Regev, A., & Zhang, F. (2019). DNA Microscopy: Optics-free Spatio-genetic Imaging by a Stand-Alone Chemical Reaction. *Cell*, *178*(1), 229-241.e16. https://doi.org/10.1016/j.cell.2019.05.019

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*. https://doi.org/10.1038/ncomms14049