Highly Multiplexed Single Cell *In Situ* RNA Detection

Thesis by
Sheel Mukesh Shah

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2017
Defended December 12th, 2016

Sheel Mukesh Shah
ORCID: 0000-0002-6321-4669

ACKNOWLEDGEMENTS

I would like to begin by thanking my adviser, Dr. Long Cai, for his guidance, patience, persistence, and support over the course of my time at Caltech. I have learned a great deal from him both personally and professionally and hope that I may have influenced him a bit too. It has been an eventful journey and I hope that he has enjoyed working with me as much as I have with him.

Furthermore, I would like to acknowledge my thesis committee, which includes Dr. Micheal Elowitz, Dr. Viviana Gradinaru, and Dr. John Allman, for their guidance, support, and encouragement over the last few years. Their constant enthusiasm has been a significant motivation over the years.

There are a few people without whom much of this work would not have been possible. I would like to thank Eric Lubeck for his guidance during my early years as a graduate student. Even though most days I felt lost early in my graduate career, he provided a sense of direction and perspective. Also, having the opportunity to work with him on many projects, I have learned not only how to carry out a project, but also finish it. In much the same vein, I would like Zakary Singer for his guidance and mentor-ship during my time here at Caltech. In addition, none of this work would have been possible without the tireless assistance of Wen Zhou and Noushin Koulena. Their influence goes much beyond simple experimental help and have been instrumental in every facet of all the work presented here. Last, but by no means least, I would like to thank all of the Cai lab group members past and present. Their support and late night discussions sessions will not be forgotten. Particularly, I would like to thank Linus Eng for letting me steal all his chocolate.

I have learned throughout my time here at Caltech that the importance of friends and family cannot be overstated. Without their constant support I definitely would not have gotten to this point. First off, I would like to thank my parents Mukesh and Sandhya Shah and my brother Samip Shah. Their constant belief in me has been the single greatest source of motivation through the years. I would like to thank my cousins Secil Shah, Pratik Shah (x2), Priti Shah, Gaurav Shah, and Sona Aurora for always being there when I need someone to talk to. In additon, I would like to acknowledge a few friends specifically, particularly Cecil Patel, Chris Wijekoon, and Kiran Alluri, for always keeping me grounded. I would be remiss not to mention Bryan Lopez, who has been my roommate throughout my entire time at Caltech and

has never once complained about the mess I leave behind or the dishes I do not do.

Finally, I would like to thank the gym crew: Harry Nunns, Constantine Sideris, Mike Abrams, Ty Basinger, and Kim Kibeom for constantly motivating me in the weight room and putting up with my incessant complaining. In retrospect, those few hours a day were so cathartic that I'm not sure I could have survived without it.

ABSTRACT

Identifying the genetic basis of cellular function and identity has become a central question in understanding the functioning of complex biological systems in recent years. Single cell sequencing techniques have provided a great deal of insight into the transcriptional profiles of various cell types. However, single cell RNAseq studies require cells to be removed from their native environments resulting in the loss of spatial relationships between cells and suffer from low detection efficiency. Moving forward, a central question in further understanding large biological systems consisting of many disparate cell types will be how do these cells interact with each other to form functional tissues. To accomplish this goal, a method that keeps the tissue architecture intact is required. Single molecule fluorescence *in situ* hybridization (smFISH) is one such technique, but suffers from a lack of multiplex measurement capability as only a very few genes can be measured in any given sample and has low signal to noise ratio. Here I present a method that overcomes the low signal to noise ratio by using an amplification technique known as single molecule hybridization chain reaction (smHCR). smHCR coupled with the existing sequential FISH (seqFISH) method, which overcomes the inherent multiplexing limit of smFISH, provides a powerful tool to measure the copy numbers of 100's of genes in single cell *in situ*.

The mouse brain contains 100,000,000 cells arranged into distinct anatomical structures. While cell types have been previously characterized by morphology and electrophysiology, single cell RNA sequencing has recently identified many cell types based on gene expression profiles. On the other hand, the Allen Brain Atlas (ABA) provides a systematic gene expression database using *in situ* hybridization (ISH) of the entire mouse brain, but lacks the ability to correlate the expression of different genes in the same cell. Using the smHCR-seqFISH technique to measure the expression profiles of up to 249 genes in single cells in coronal brain sections, we have identified distinct cell clusters based on the expression profiles of 15000 cells and observed spatial patterning of cells in the hippocampus. In the dentate gyrus, we resolved lamina-layered patterns of cell clusters with a clear separation between the granule cell layer and the sub-granular zone. In CA1 and CA3, the data revealed distinct subregions, each with unique combinations of cell clusters. Particularly, we observed that the dorso-lateral CA1 is almost completely cellular homogeneous with increasing cellular heterogeneity on the dorsal to ventral axis.

Together, these results demonstrate the power of highly multiplex *in situ* analysis to the brain, with further application to a wide range of biological systems.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Sheel Shah et al. "NeuroResource In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus". In: *Neuron* 92.2 (2016), pp. 342–357. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2016.10.001. URL: http://dx.doi.org/10.1016/j.neuron.2016.10.001.
SS designed the experiments, carried out the experiments, analyzed the data, and wrote the paper.

[2] Sheel Shah et al. "Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing". In: *Development* 143.15 (2016), pp. 2862–2867. DOI: 10.1242/dev.138560. URL: http://dx.doi.org/10.1242/dev.138560.
S.S participated in the conception of the project, prepared the data, and participated in the writing of the manuscript.

[3] Bin Yang et al. "Resource Single-Cell Phenotyping within Transparent Intact Tissue through Whole-Body Clearing". In: *Cell* 158.4 (Aug. 2014), pp. 945–958. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.07.017. URL: http://dx.doi.org/10.1016/j.cell.2014.07.017.
S.S participated smFISH experiments and helped design the figure.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Single Cell Transcriptomics

Quantitatively exploring the differences between cells at the transcriptional level has provided significant insight into the functioning of cells in multicellular and unicellular organisms. Recent advances in sequencing technology has allowed the development of single cell mRNA sequencing methods that can provide unbiased and high-thoughput measurements of the transcription states of individual cells. These methods have been used to reveal new biological insights into the transcriptional dynamics of cells, the organizations of tissues, and relationships between genes.

The study of cellular transcriptomics has been a cornerstone of cell biology for over 3 decades. Until recently, the prevailing methodology to study transcription at the cellular level was expression microarrays[23]. While this technology enabled significant scientific progress, such as discovery of new genes, insight into diseases processes, drug discovery, and toxicology research, there persisted some inherent limitations. These limitations included mis-hybridizations, detection artifacts due to fluorescence based readouts, the need to know RNA sequences a priori in a time where such data was not readily available, high detection threshold, and completely lack of quantification. These concerns limited the applicability of expression microarrays to comprehensively investigate the transcriptome[1, 6, 18].

Alternative were available to assuage some of these limitations, such as serial analysis of gene expression (SAGE)[27]. The SAGE method, and those related to it, leveraged the recent advances in sequencing technology at the time and resulting cost-effectiveness by probing the transcriptome by DNA sequencing of cloned tags. By mapping these tagged sequences back onto a reference genome, the mRNA could be identified with some level of certainty and the sequencing data could be used to extract an estimate of the copy number of a specific mRNA species. While, the SAGE method did provide a level of quantitative power, the method proved to be extremely labor intensive, the sequences of the mRNA or gene still needed to be known a priori, and sequencing efficiency remained a concern.

Finally, a more efficient, comprehensive, and less complicated method to measure the transcriptome was developed based on direct sequencing of cDNA, named

RNA-seq[15]. Briefly, this method works by first selecting RNA based on the characteristic poly(A) tail of mRNA. The RNA is then fractured to get fragments of about 200 nucleotides. The resulting fragments are converted to a cDNA library by random primers. The cDNA library is then sequenced and individual reads are mapped onto the reference genome. The reads can also be counted and normalized to obtain a quantitative measure of the of the amount of the mRNA present. Also, as this method is based on direct sequencing of cDNA, new genes can be found and gene isoform variability can be investigated.

RNA-seq, while providing significant insight into the transcriptional differences between biological samples, could not be used to compare individual cells in the same sample. Due to the large amount of RNA required and the cellular complexity of biological tissues, RNA-seq could only provide course bulk averages of transcriptional differences between tissue and not the actual cellular difference between tissues. Towards this goal, single cell methods to address were developed taking advantage of advances in sequencing and whole transcriptome amplification technologies. The first implementations of single cell RNA-seq used oligo-dT primers and subsequent ligation adapter PCR or linear transcription with T7[24]. These methods however suffered resulted in strong 3 prime bias due to the inefficiency of the reverse transcriptase enzyme and PCR based amplification bias. To alleviate the 3 prime bias, a method based on template switching was developed, called Smart-seq, using a Moloney Murine Leukemia Virus reverse transcriptase[21, 19]. Further advances were made to reduce the PCR amplification bias but using unique molecular identifiers (UMIs) to tag mRNA molecules with unique barcodes before whole transcriptome amplification[8].

At this point, the single greatest barrier to large scale adoption of single cell sequencing in biology was the lack of throughput available to sequence enough cells to get statistically relevant data and the requisite labor involved in cell isolation. These technical challenges are being addressed in many ways. The first attempt to solve this challenge used a microfluidics approach to capture cells into micro-wells. Currently, the most widely used platform using this approach is the Fluidigm C1 chip[22]. This method first relies on dissociating the cells in the tissue and then flowing the resulting solution into the integrated fluidic circuit (IFC) of the C1 device which can capture up to 800 cells. Cell are then stained and imaged to measure cell count and viability. The IFC is then placed in the C1 device which lyses the cells and amplifies the RNA content of each cell. The output of the C1 is then

sent off for sequencing to get single cell sequencing data. While the the Fluidigm paradigm does increase the throughput of single cell sequencing, it suffers from low efficiency of amplification and the throughput is not nearly high enough to capture meaningful data from highly heterogeneous tissue or extremely rare cell populations in a cost-effective manner.

Recently, an alternative group of methods to increase the throughput of single cell RNA sequencing have arisen that use microfluidic droplets containing a bead coated with barcoded primers to capture cells and identify them[14, 9]. Current estimates suggest that these methods may be 100 times cheaper. The cost of a single cell using the C1 method has been estimated to be 10 dollars, while droplet based methods are estimated to cost about 10 cents a cell. Two competing droplet based methods have arisen, named Drop-seq and in-Drop. These methods allow for the sequencing of thousands of cells in a relatively short period of time with unparalleled cost-efficiency.

The single greatest drawback to all these single cell methods is ultimately the lack of mRNA capture efficiency and spatial resolution. Currently, these methods can only be used to reliably quantify high expression genes or genes with a high degree of variability between samples and provide no spatial resolution as cells need to be isolated from the tissue. However, despite these limitations single cell RNA-seq has provided great insight in to the functioning of biological systems.

1.2 Single Cell Measurements are Necessary as Transcriptional Heterogeneity Defines Cell States and Types

Identifying functionally distinct cells in a complex biological requires the ability to quantitatively classify cells based on some measure or canonical marker. However, cells may be functionally distinct and have no known reliable markers that can identify them for further studies. Even cells with known markers may contain significant diversity that will be ignored if cells were classified based on a single or few known markers. With the advent of single cell sequencing, a concerted effort has been made to identify cell types in different tissues based on a quantitative measure of RNA expression. Using RNA profile with subsequent analysis with high dimensional data analysis techniques has yielded new insights into everything from neuroscience to stem cells[16, 25].

Single-cell measurements for cell type identification is necessary due to the fact that bulk measurements average data from many cells together resulting in the loss of

individual cell identity in the measurements. In addition, bulk measurements cannot distinguish between differences arising from gene regulation and differences arising from changes in cell types. Finally, bulk measurements of mRNA will also average out any dynamical processes occurring the in the sample, which can be essential for the understanding of developmental processes and disease progression[10].

Neuroscience, particularly, has recently been a focal point in the effort towards validating current and finding new cell types. Due to the number of tasks the mammalian brain has to accomplish and the amount of information it has to integrate, the cell diversity of the brain must be equally complex. Particularly, the cerebral cortex supports functions such as sensorimotor sensing and control, memory integration, and social learning and behaviors. The normal functioning of these pathways requires a large of broad cell types such as neurons, glia, and vasculature. Recently, Zeisel et al. used single cell RNA-seq to classify cells in the mouse somatosensory cortex and the CA1 of the hippocampus[29]. By using the quantitative power of single cell sequencing along with novel data analysis methods, they found 47 previously unknown molecularly distinct subclasses, that made up the largely known broad cell types. Furthermore, single cell analysis allowed the identification of known marker genes, which allowed them to align the sequenced cells with known known cell types, and infer the rough morphology and location of the sequenced cells. Putting all of this data together they managed to find a novel new layer I interneuron expressing Pax6 and a postmitotic oligodendrocyte subclass marked by Itpr2. By using single cell sequencing, Zeisel et al. were able to provide significant insight into the cellular building blocks of neural circuits and potentially identify previously unknown components of these circuits.

Single cell sequencing can also be used to accomplish completely de novo cell type discovery as shown by Macosko et al. using their Drop-seq method[14]. By using their massively parallel approach to sequencing, they were able to analyze transcriptomes from almost 45,000 mouse retinal cells and identified 39 transcriptionally distinct cell populations. This level of throughput allowed them to create a molecular atlas of gene expression for known retinal cell classes along with the identification of potentially novel new candidate cell subtypes.

By using a similar method to increase the cellular throughput of single cell sequencing, Klein et al. were able to find rare cell states in wild type embryonic stem cells (mESCs) as well as properties of cellular differentiation[9]. Their analysis of mESCs revealed the details of the population structures within mESC cultures

and that the onset of differentiation seems to be stochastic after leukemia inhibitory factor withdrawal.

While cell type identification of single cells has become the primary application of single cell transcriptomics, high dimensional transcription measurements can also be used to investigate the dynamics of cellular transcription states. One example of such an application is the work done by Treutlein et al. in investigating the dynamics of direct lineage reprogramming[26]. In certain cases, with the correct stimuli, primed fibroblasts have been shown to be able to directly differentiate into a terminal cell type without specifically having to convert them into induced pluripotentent stem cells first. This process is known as direct lineage reprogramming and is a prime example of a system in which a fundamental shift in transcription occurs. While the protocol to directly differentiate mouse embryonic fibroblasts into an induced neural cell type is available, the intermediate stages through which the cells progress to reach the final neuronal fate is mostly unknown. Treutlein et al used single-cell RNA sequencing at multiple time points to determine the cellular heterogeneity at each time point. Using this single cell transcriptional data to order cells based on transcriptional similarity, they found that the reprogramming path is mostly continuous with no intermediate stem cell fate. Furthermore, their transcription data seems to suggest that Ascl1 is responsible for initialization of reprogramming by causing cells to exit the cell cycle and inducing changes in gene expression to produce more neural transcription factors. Their data also seems to suggest that the 20 percent conversion rate is not due to reprogramming initiation but results from competition with a myogenic program that seems to also be initiated during this reprogramming.

While these studies show the prodigious insight that can be gained by single cell sequencing into transcriptional regulation in complex tissue and transcriptional heterogenity of cells in a tissue, the cells are taken out of their biological context to accomplish these measurements. The next leap forward in single cell transcriptomics will be the ability to get high dimensional transcription information while preserving the spatial context of the cells.

1.3 Efforts to Multiplex smFISH

Identifying the spatial organization of distinct cell types within their native environment is essential in understanding the function of many biological systems, especially in neuroscience. Some efforts have been made to classify cells based on

gene expression. In particular, the Allen Brain Atlas (ABA) provides a systematic gene expression database using in situ hybridization (ISH) of the entire mouse brain one gene at a time. This comprehensive reference provides regional gene expression information but lacks the ability to correlate the expression of different genes in the same cell and provides little quantitative power. More recently, single-cell RNA sequencing has identified many cell types based on gene expression profiles. However, while single-cell RNA-seq methods provide useful information on multiple genes in individual cells, it has relatively low detection efficiencies and requires cells to be removed from their native environment, resulting in the loss of spatial information. These different approaches can lead to contradictory descriptions of cellular organization in the brain and other biological systems.

Single molecule RNA imaging approaches can overcome this barrier by providing transcription data while preserving the spatial relationships between cells. One such approach is known as single molecule fluorescence in situ hybridization (smFISH)[7, 20]. This method works by first designing many oligonucleotide probes with sequences complementary to the mRNA of interest. These probes are then labeled with flourophores and added to the sample of choice. The probes are allowed to hybridize to the mRNA of interest and then the excess probes are washed away. Standard fluorescence microscopy can then be used to image the sample resulting in bright fluorescent points indicating the location of individual mRNA molecules in the sample. While a very powerful approach to probe the transcriptome of cells, smFISH suffers from two major drawbacks. The first limitation is the ability to image transcripts in thick tissue sections. Due to the thickness of the sample, the fluorescent signals tend to be obfuscated by scattering of light through the sample, increased autoflorescence, and poor hybridization efficiency. The second major limitation of smFISH is the number of RNA species that can be uniquely identified. The number of mRNA species that can be quantified is generally limited to the number of distinct fluorescence channels available for imaging, which in most cases is about 3-5. These limitations make it challenging to perform high dimensional, unbiased spatial transcriptomic measurements.

Although, smFISH provides a convenient avenue to investigate the transcriptional profile of single cells in their native enviornment, the applications of smFISH beyond cell culture has proved to be difficult. A limited number of studies have managed to use smFISH in the context of tissue sections, but these studies have generally required either extremely thin slices or prohibitively large probe sets to

get reliable results[13, 5, 17]. The reasons for the lack of application to tissue sections are three fold: first, the thickness of the sample reduces the permeability of the probes resulting in low hybridization inefficiencies; second, the thickness of the sample increases light scattering; and third, the thickness of the sample results in increased auto-fluorescence signal. These three problems can partially be solved by using hydrogel tissue clearing techniques. Yang et al. used their PACT clearing method to visualize single actin mRNA transcripts in 100 micron thick mouse brain section[28]. Actin transcripts were shown to be retained in the cytoplasm of neurons with a reasonable signal to noise ratio in thick brain tissue sections (Figure 1). This increased signal to noise ratio is due to the fact that the PACT treated tissue showed remarkable decrease in scattering and tissue autofluorescence.

However, even with the greatly reduced scattering of the tissue section due to clearing, smFISH signal still remained somewhat poor and unreliable. Along with tissue clearing, a method was needed that would increase the specific signal from individual mRNA molecules. In order to accomplish this signal amplification, the available hybridization chain reaction (HCR) method was adapted for single molecule detection[4]. The details of this extension of HCR can be found in chapter 2.

A great deal of progress has been made recently in developing highly quantitative methods to profile the transcriptome of single cells. Building upon single-molecule fluorescence in situ hybridization, Lubeck and Cai devised a general method to highly multiplex single-molecule in situ mRNA imaging regardless of transcript density using super-resolution microscopy[11]. However, the spectral barcoding methods used in these previous works is difficult to scale up beyond 20–30 genes because of the limited number of fluorophores.

To overcome the scalability problem, a temporal barcoding scheme was developed that uses a limited set of fluorophores and scales exponentially with time. Specifically, sequential probe hybridizations on the mRNAs in fixed cells impart a unique pre-defined temporal sequence of colors, generating in situ mRNA barcodes[12]. The multiplex capacity scales as F^n, where F is the number of fluorophores and n is the number of rounds of hybridization. Thus, one can increase the multiplex capacity by increasing the number of rounds of hybridization with a limited pool of fluorophores. This approach is called sequential fluorescence in situ hybridization (seqFISH) (Figure 2). In parallel, in situ sequencing methods were developed to directly sequence transcripts in tissue sections, but these methods

suffer from low detection efficiency (<1%)[10]. Recently, Chen et al. expanded the error correction method in the original seqFISH demonstration by using a Hamming distance 2 based error correcting barcode system called MERFISH. However, this implementation requires larger transcripts (>6 kb) and many more rounds of hybridization than the method described here[3]. Furthermore, seqFISH and its variants have only previously been applied in cell culture systems due to the difficulty of smFISH detection in tissue. In chapter 3, I demonstrate an improved version of seqFISH in complex tissues by including signal amplification and a time-efficient error correction scheme, allowing investigation of the structural organization of the hippocampus with single-cell resolution.

1.4 Future Directions for Highly Multiplexed *In Situ* mRNA Detection

Multiplexing in situ mRNA detection has been of great interest to the single cell community for almost 10 years now. Many methodological variants have been devised to accomplish this goal. This work provides one such method with the primary utility being the ease of application to thick tissue sections. However, few problems still persist in the general applicability of this method along with some technical limitations.

The primary scientific limitation at this point for seqFISH is the inability to multiplex to even larger numbers of transcripts. Up to now, the largest number of transcripts have been reliable detected in a efficient combinatorial fashion in thick tissue sections is 249 genes. For more general applicability and scientific interests, this number will need to be expanded potentially to thousands and ideally to be able to measure the entire transcriptome. Two major limitations exist that make increased multiplexing difficult: first, is the limitation of available optical space in a cell and second is the increased non-specific binding noise that results from increasing the number of probes. The first limitation may be overcome by using expansion microscopy methods to increase the size of the cells allowing more optical voxels that can be unique resolved within the cell[2]. The second limitation will require significant improvements in probe design or an alternative method for signal amplification that is less prone to false positive signals.

The next challenge in the general applicability of seqFISH is the ability to design specific probes with minimal cross-hybridizations and non-specific binding. Some work has already been done on this front with probes being designed to detect all known transcript variants and quantitative minimization of cross-hybridization

in large probe pools.

One additions area for significant improvement is the generalization of cell segmentation. For seqFISH to reach maximum applicability and generalization, a standardized cell segmentation method will need to be developed. While different samples will invariably require slightly different segmentation strategies, a generalized software pipeline needs to be established that can accurately segment cells in complex tissue sections. Again, work has begun in this direction with the application of machine learning image segmentation methods applied to membrane antibody stain images. These methods seem to be robust and general in their application.

There also needs to be a concerted effort in the future to improve the image decoding software. Many aspects of the existing software pipeline can be improved to make it more user friendly and more accurate. The dot finding can be improved with point spread function modeling of the dots and subsequent iterative dot finding. This extension of dot finding method will allow the resolution of mRNA dots that are closely spaced together resulting in fewer ambiguous barcodes and greater detection efficiency. The barcode calling itself could be done in an iterative fashion such that points that match to create a single barcode are removed with every round and the remaining points are re-matched.

As seqFISH data is almost completely a new data type, more effort needs to be devoted to the theoretical framework around how this data is interpreted. The analysis methods (i.e. hierarchical clustering, bi-clustering, etc.) that have been applied to RNA-seq may not be applicable in this case as the data obtained is of a much higher resolution picture of the transcriptome. While RNA-seq data relies mostly on defining cell types between large varying blocks of genes, seqFISH can provide detailed quantitative measures such that all measured genes can be of equal importance. Also, greater attention needs to be payed into how to cluster the spatial data obtained through the seqFISH method.

As a final note, the experimental labor can be significantly reduced by automation of the experimental protocol. This can and will be done in short order.

1.5 Figures

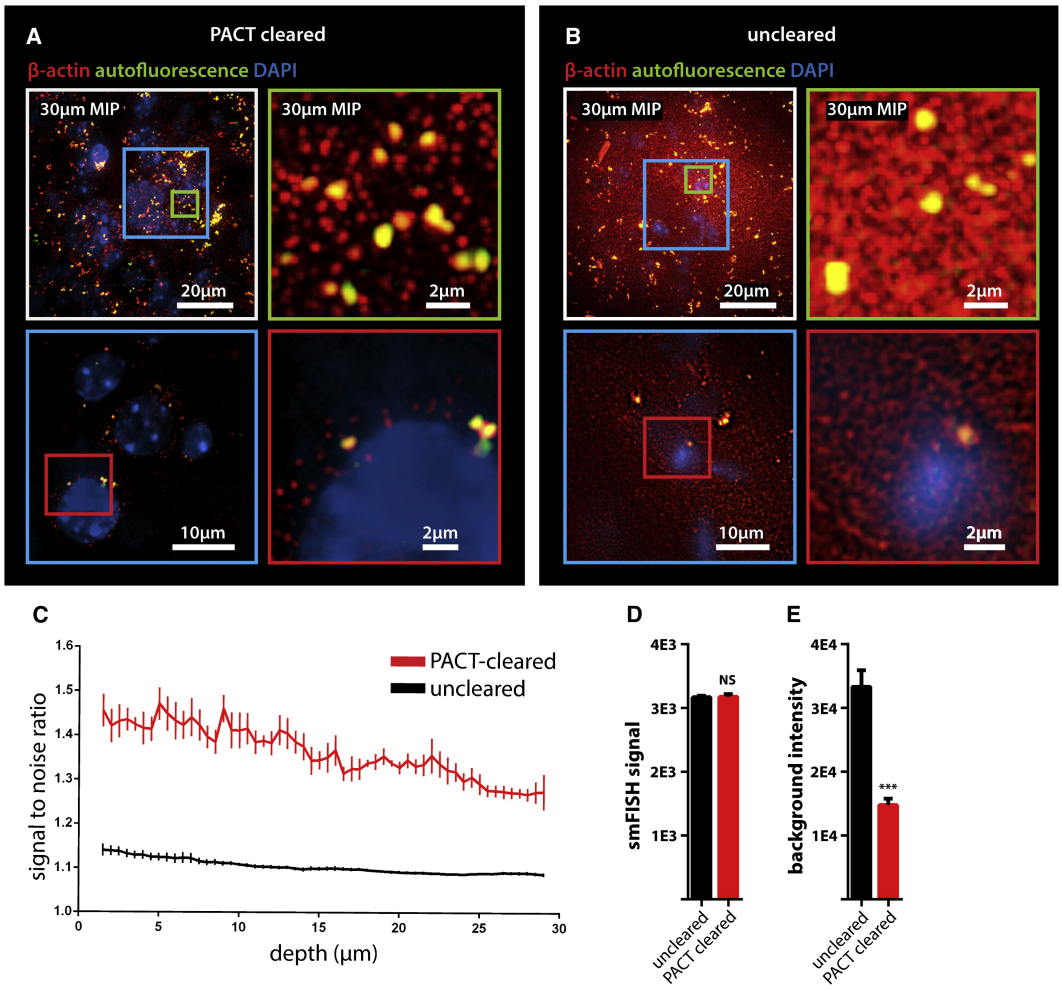

A
PACT cleared
β-actin autofluorescence DAPI
30µm MIP
30µm MIP
20µm
2µm
10µm
2µm
B
uncleared
β-actin autofluorescence DAPI
30µm MIP
30µm MIP
20µm
2µm
10µm
2µm
C
PACT-cleared
uncleared
signal to noise ratio
depth (µm)
D
smFISH signal
NS
uncleared
PACT cleared
E
background intensity
***
uncleared
PACT cleared

Figure 1 *(previous page)*: Detection of Individual mRNA Transcripts in PACT Tissue Sections by smFISH. **(A and B)** 100-micron-thick mouse brain slices were hybridized with twenty-four 20-mer oligonucleotide probes toward actin mRNA labeled with Alexafluor 594. **(A)** PACT-cleared smFISH brain slices. Top shows 30 micron maximum intensity projection. An abundant number of diffraction limited spots corresponding to single actin mRNAs (red) were readily detected up to 30 micron in depth under 589 nm illumination. Note bright amorphous granules (yellow) are background lipofuscin vesicles that show up in both 589 nm (red) and 532 nm autofluorescence (green) channels, whereas smFISH signals are in the red channel only.**(B)** Compared to PACT-cleared slices, smFISH in uncleared brain slices showed significantly decreased contrast. (Bottom in A and B show single slices of 0.5 micron at 12 micron depth; the images were processed from raw data using the same contrast scale and Laplacian of Gaussian filtering). **(C)** Signal to noise ratio as a function of depth shows PACT-clearing tissue increases the signal to noise ratio of smFISH throughout the thickness of the sample as compared to uncleared tissue. **(D)** smFISH intensities show no appreciable differences between uncleared and PACT-cleared tissue. p = 0.8722; 2-tailed Student's t test. **(E)** Comparison of background intensity between uncleared and PACT-cleared tissue illustrates the significant reduction of background fluorescence in PACT-cleared tissue.p = 0.0006; 2-tailed Student's t test. All graphs are shown in mean ± SEM. Imaging set-up: **(A and B)** samples imaged on a Nikon Ti Eclipse microscope with an Andor Ikon-M camera and a 60×/1.4NA Plan Apo λ objective with an additional 1.5× magnification. Images acquired as Z-stacks with a 0.5 micron step size over 30 micron. Samples were excited by a 640 nm Coherent Cube, and 532 nm (SDL-532-200TG) and 405 nm(SDL-405-LM-030) lasers.

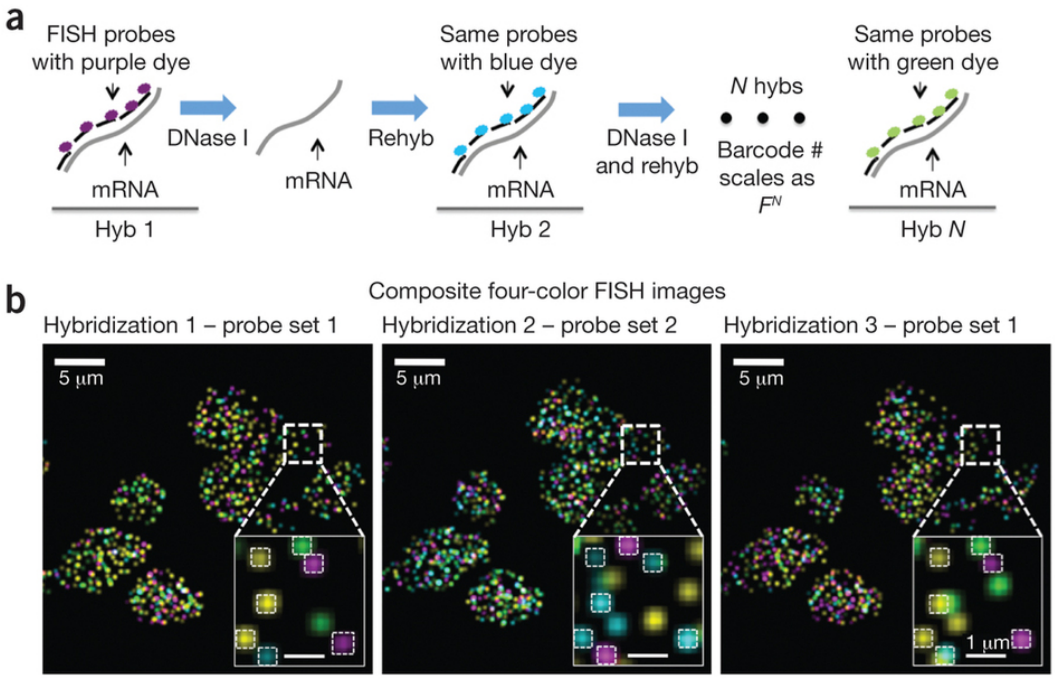


Figure 2: Sequential barcoding. **(A)** Schematic of sequential barcoding. In each round of hybridization, 24 probes are hybridized on each transcript, imaged and then stripped by DNase I treatment. The same probe sequences are used in different rounds of hybridization (hyb), but probes are coupled to different fluorophores. **(B)** Composite four-color FISH data from three rounds of hybridizations on multiple yeast cells. Twelve genes are encoded by two rounds of hybridization, with the third hybridization using the same probes as hybridization 1. The boxed regions are magnified in the bottom right corner of each image. Spots colocalizing between hybridizations are detected (as outlined in insets) and have their barcodes extracted. Spots without colocalization are due to nonspecific binding of probes in the cell as well as mishybridization. The number of instances of each barcode can be quantified to provide the abundances of the corresponding transcripts in single cells.

1.6 References

[1] Tineke Casneuf et al. “In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation”. In: *BMC Bioinformatics* 8.1 (2007), p. 461. ISSN: 14712105. DOI: 10.1186/1471-2105-8-461. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-461.

[2] Fei Chen et al. “Nanoscale imaging of RNA with expansion microscopy”. In: *Nature Methods* 13.8 (July 2016), pp. 679–684. ISSN: 1548-7091. DOI: 10.1038/nmeth.3899. URL: http://www.nature.com/doifinder/10.1038/nmeth.3899.

[3] K. H. Chen et al. “Spatially resolved, highly multiplexed RNA profiling in single cells”. In: *Science* 348.6233 (Apr. 2015), aaa6090–aaa6090. ISSN: 0036-8075. DOI: 10.1126/science.aaa6090. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.aaa6090.

[4] Harry M. T. Choi, Victor A. Beck, and Niles A. Pierce. “Next-Generation in Situ Hybridization Chain Reaction: Higher Gain, Lower Cost, Greater Durability”. In: *ACS Nano* 8.5 (May 2014), pp. 4284–4294. ISSN: 1936-0851. DOI: 10.1021/nn405717p. URL: http://pubs.acs.org/doi/abs/10.1021/nn405717p.

[5] Jonathan R Chubb and Tanniemola B Liverpool. “Bursts and pulses: insights from single cell studies into transcriptional mechanisms”. In: *Current Opinion in Genetics & Development* 20.5 (2010), pp. 478–484. ISSN: 0959437X. DOI: 10.1016/j.gde.2010.06.009.

[6] Aron C Eklund et al. “Replacing cRNA targets with cDNA reduces microarray cross-hybridization”. In: *Nature Biotechnology* 24.9 (Sept. 2006), pp. 1071–1073. ISSN: 1087-0156. DOI: 10.1038/nbt0906-1071. URL: http://www.nature.com/doifinder/10.1038/nbt0906-1071.

[7] Andrea M. Femino et al. “Visualization of Single RNA Transcripts in Situ”. In: *Science* 280.5363 (1998).

[8] Saiful Islam et al. “Quantitative single-cell RNA-seq with unique molecular identifiers”. In: *Nature Methods* 11.2 (Dec. 2013), pp. 163–166. ISSN: 1548-7091. DOI: 10.1038/nmeth.2772. URL: http://www.nature.com/doifinder/10.1038/nmeth.2772.

[9] Allon M Klein et al. “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.” In: *Cell* 161.5 (May 2015), pp. 1187–201. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.04.044. URL: http://www.ncbi.nlm.nih.gov/pubmed/26000487%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4441768.

[10] Je Hyuk Lee et al. “Highly Multiplexed Subcellular RNA Sequencing in Situ”. In: *Science* 343.6177 (2014).

[11] Eric Lubeck and Long Cai. "Single-cell systems biology by super-resolution imaging and combinatorial labeling". In: *Nature Methods* 9.7 (June 2012), pp. 743–748. ISSN: 1548-7091. DOI: 10.1038/nmeth.2069. URL: http://www.nature.com/doifinder/10.1038/nmeth.2069.

[12] Eric Lubeck et al. "Single-cell in situ RNA profiling by sequential hybridization". In: *Nature Methods* 11.4 (Mar. 2014), pp. 360–361. ISSN: 1548-7091. DOI: 10.1038/nmeth.2892. URL: http://www.nature.com/doifinder/10.1038/nmeth.2892.

[13] Anna Lyubimova et al. "Single-molecule mRNA detection and counting in mammalian tissue". In: *Nature Protocols* 8.9 (Aug. 2013), pp. 1743–1758. ISSN: 1754-2189. DOI: 10.1038/nprot.2013.109. URL: http://www.nature.com/doifinder/10.1038/nprot.2013.109.

[14] Evan Z. Macosko et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5 (2015), pp. 1202–1214. ISSN: 00928674. DOI: 10.1016/j.cell.2015.05.002.

[15] Ali Mortazavi et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5.7 (July 2008), pp. 621–628. ISSN: 1548-7091. DOI: 10.1038/nmeth.1226. URL: http://www.nature.com/doifinder/10.1038/nmeth.1226.

[16] Nicholas E Navin. "The first five years of single-cell cancer genomics and beyond." In: *Genome research* 25.10 (Oct. 2015), pp. 1499–507. ISSN: 1549-5469. DOI: 10.1101/gr.191098.115. URL: http://www.ncbi.nlm.nih.gov/pubmed/26430160%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4579335.

[17] Yuma Oka et al. "Whole-mount single molecule FISH method for zebrafish embryo". In: *Scientific Reports* 5 (Feb. 2015), p. 8571. ISSN: 2045-2322. DOI: 10.1038/srep08571. URL: http://www.nature.com/articles/srep08571.

[18] Michał J Okoniewski et al. "Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations". In: *BMC Bioinformatics* 7.1 (2006), p. 276. ISSN: 14712105. DOI: 10.1186/1471-2105-7-276. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-276.

[19] Simone Picelli et al. "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature Protocols* 9.1 (Jan. 2014), pp. 171–181. ISSN: 1754-2189. DOI: 10.1038/nprot.2014.006. URL: http://www.nature.com/doifinder/10.1038/nprot.2014.006.

[20] Arjun Raj et al. "Imaging individual mRNA molecules using multiple singly labeled probes". In: *Nature Methods* 5.10 (Oct. 2008), pp. 877–879. ISSN: 1548-7091. DOI: 10.1038/nmeth.1253. URL: http://www.nature.com/doifinder/10.1038/nmeth.1253.

[21] Daniel Ramsköld et al. “Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells”. In: *Nature Biotechnology* 30.8 (July 2012), pp. 777–782. ISSN: 1087-0156. DOI: 10.1038/nbt.2282. URL: http://www.nature.com/doifinder/10.1038/nbt.2282.

[22] Veronica Sanchez-Freire et al. “Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns”. In: *Nature Protocols* 7.5 (Apr. 2012), pp. 829–838. ISSN: 1754-2189. DOI: 10.1038/nprot.2012.021. URL: http://www.nature.com/doifinder/10.1038/nprot.2012.021.

[23] Mark Schena et al. “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray”. In: *Science* 270.5235 (1995).

[24] Fuchou Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nature Methods* 6.5 (May 2009), pp. 377–382. ISSN: 1548-7091. DOI: 10.1038/nmeth.1315. URL: http://www.nature.com/doifinder/10.1038/nmeth.1315.

[25] Cole Trapnell. “Defining cell types and states with single-cell genomics.” In: *Genome research* 25.10 (Oct. 2015), pp. 1491–8. ISSN: 1549-5469. DOI: 10.1101/gr.190595.115. URL: http://www.ncbi.nlm.nih.gov/pubmed/26430159%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4579334.

[26] Barbara Treutlein et al. “Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq”. In: *Nature* 534.7607 (June 2016), pp. 391–395. ISSN: 0028-0836. DOI: 10.1038/nature18323. URL: http://www.nature.com/doifinder/10.1038/nature18323.

[27] Victor E. Velculescu et al. “Serial Analysis of Gene Expression”. In: *Science* 270.5235 (1995).

[28] Bin Yang et al. “Resource Single-Cell Phenotyping within Transparent Intact Tissue through Whole-Body Clearing”. In: *Cell* 158.4 (Aug. 2014), pp. 945–958. ISSN: 0092-8674. DOI: 10.1016/j.cell.2014.07.017. URL: http://dx.doi.org/10.1016/j.cell.2014.07.017.

[29] Amit Zeisel et al. “Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq”. In: *Science* 347.6226 (2015).

Chapter 2

SINGLE-MOLECULE RNA DETECTION AT DEPTH VIA HYBRIDIZATION CHAIN REACTION AND TISSUE HYDROGEL EMBEDDING AND CLEARING

[1] Sheel Shah et al. "Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing". In: *Development* 143.15 (2016), pp. 2862–2867. DOI: 10.1242/dev.138560. URL: http://dx.doi.org/10.1242/dev.138560.

2.1 Summary

Accurate and robust detection of mRNA molecules in thick tissue samples can reveal gene expression patterns in single cells within their native environment. Preserving spatial relationships while accessing the transcriptome of selected cells is a crucial feature for advancing many biological areas, from developmental biology to neuroscience. However, because of the high autofluorescence background of many tissue samples, it is difficult to detect single-molecule fluorescence *in situ* hybridization (smFISH) signals robustly in opaque thick samples. Here, we draw on principles from the emerging discipline of dynamic nucleic acid nanotechnology to develop a robust method for multi-color, multi-RNA, imaging in deep tissues using single-molecule hybridization chain reaction (smHCR). Using this approach, single transcripts can be imaged using epifluorescence, confocal or selective plane illumination microscopy (SPIM) depending on the imaging depth required. We show that smHCR has high sensitivity in detecting mRNAs in cell culture and whole-mount zebrafish embryos, and that combined with SPIM and PACT (PAssive CLARITY Technique) tissue hydrogel embedding and clearing, smHCR can detect single mRNAs deep within thick (0.5 mm) brain slices. By simultaneously achieving ~20-fold signal amplification and diffraction-limited spatial resolution, smHCR offers a robust and versatile approach for detecting single mRNAs *in situ*, including in thick tissues where high background undermines the performance of unamplified smFISH.

2.2 Introduction

Imaging gene expression levels with single-cell resolution in intact tissues is essential for understanding the genetic programs in many systems, such as developing embryos and dynamic brain circuits. Single-molecule fluorescence in situ hybridization (smFISH) has been the standard tool for detection of individual RNAs in cells (1-5). Using smFISH, an mRNA is detected by a probe set containing 20-40 DNA probes, each carrying one or more fluorophores, and each complementary to a different short subsequence (20-50 nt) along the mRNA target. This approach ensures that multiple probes bind the mRNA, generating bright puncta that can be discriminated from background staining resulting from non-specific binding of individual probes. However, background due to sample autofluorescence is significantly higher in tissue samples than in cell culture, making it difficult to robustly detect smFISH signals in tissue. In addition, light scattering caused by deep tissue imaging necessitates probes with higher photon counts than for thin section imaging. While we have shown that tissue clearing by PACT (PAsive CLARITY Technique) can alleviate autofluorescence and light scattering problems (6-8) while preserving RNA molecules (8), a more robust signal amplification strategy is needed to enable multi-color mapping of single mRNAs in deep tissues.

Attempts have been made to specifically amplify a single mRNA signal, but these tend to suffer from low efficiency and complex protocols (9). We herein describe a simple and efficient method for multiplexed single-molecule signal amplification based on the mechanism of hybridization chain reaction (HCR) (21,10,11). With this approach, short DNA probes complementary to mRNA targets trigger chain reactions in which metastable fluorophore-labeled DNA hairpins self-assemble into tethered fluorescent amplification polymers (Figure 1a). As with smFISH, each target mRNA is addressed by 20-40 probes complementary to different subsequences along the target to enable discrimination between mRNAs with multiple probes bound and dots resulting from non-specific binding of individual probes. In contrast to previous in situ HCR methods (10,11), we limit the HCR amplification time to achieve a mean polymer length of ~20-40 hairpins, generating puncta that are bright enough for high sensitivity yet small enough for diffraction-limited resolution (Supplementary Fig. 2). Using orthogonal HCR amplifiers programmed to operate independently, straightforward multiplexing is achieved for up to five channels simultaneously (Figure 1b). We term this method single-molecule HCR (smHCR).

2.3 Results and Discussion

To characterize the sensitivity and selectivity of smHCR in cultured cells, we performed a colocalization experiment in which a low-copy target mRNA (Pcdha constant region) was simultaneously detected using three probe sets of 22 probes each (one smFISH set and two smHCR sets), with the probes alternating between the three sets along the target. Dots were identified in each channel by applying a threshold following standard methods for smFISH data analysis (Supplementary Fig. 3a) (2). We define true mRNA signals as those dots that are colocalized in at least two of the three channels (Figure 1c). We calculate a true positive rate for a given channel as the percentage of true mRNA signals detected as dots in that channel. We calculate a false positive rate for a given channel as the percentage of dots in that channel that are not true mRNA signals. For the two smHCR channels and the smFISH channel, the true positive rates are all approximately 88% (Figure 1d), and the false positive rates are approximately 36%, 27% and 20%, respectively (Supplementary Fig. 3b). Comparing dot intensities, smHCR provides signal amplification of approximately a factor of 15-35 relative to smFISH (Supplementary Figs 3cd and 4), a feature that will become crucial in detecting single transcripts in tissues that have higher levels of autofluorescence.

Notably, in experimental designs where two channels can be allocated to each target mRNA, near-quantitative single-molecule mapping can be achieved (Supplementary Figs 5). Using this approach, the threshold for dot identification is lowered in each channel to achieve a higher true positive rate (> 95%) at the cost of a higher false positive rate (> 60%). Dot colocalization between channels can then be used to identify the subset of dots that represent true mRNA signals. Alternatively, for the standard situation where each target is detected in only a single channel, and hence colocalization cannot be used for dot classification, the threshold must be raised to reject false positives at the cost of also rejecting some true positives, as is the case for smFISH (Supplementary Fig 5) (2).

To examine the performance of smHCR within the more challenging imaging setting of an intact vertebrate embryo, we repeated the 3-channel colocalization study in a whole-mount zebrafish embryo with confocal imaging. The target mRNA, kdrl (a medium-copy target expressed in the endothelial cells of blood vessels), is detected using three smHCR probe sets of 39 probes each, with probes alternating between the three sets along the target mRNA. For the three smHCR channels, we observe true positive rates of 86%, 84%, and 86% (Figure 2b), and the false positive

rates of 36%, 21%, and 31% (Supplementary Fig 6).

To compare the properties of smHCR and smFISH in zebrafish embryos, we disabled HCR amplification in the Alexa 546 channel by introducing only the first HCR hairpin species, enabling only one hairpin carrying one fluorophore to bind to each probe (so-called smFISH* of Figure2c; compare the dots in the middle row of panels a and c). Comparing the signal intensities of smHCR and smFISH*, we find that the ratio of median dot intensities for smHCR and smFISH* is approximately 15 (Figure 2d). While the autofluorescence in zebrafish embryos is low enough that unamplified smFISH remains viable (12,20; Supplementary Figures 6-8), automated signal detection is facilitated by the greater signal to background ratio of smHCR.

In the higher background of adult mouse brain sections, smHCR signal amplification becomes essential for robust detection of individual transcripts, and is also important when mapping bulk expression in cleared tissue (22). To minimize autofluorescence, PACT clearing turns tissues optically transparent and macromolecule-permeable by removing light-scattering lipids and replacing them with a porous hydrogel (7,8). We hypothesized that PACT combined with smHCR should enable reliable single molecule imaging at depth within brain samples. As a first test, we used confocal microscopy to image PACT-cleared brain slices at depths up to 84 μm (Figure 3), performing a 3-channel colocalization study using two smHCR probe sets and one smFISH probe set. Using smFISH, few dots are visible at a depth of 10 μm, and no signal is evident at a depth of 37 μm (Figure 3a bottom). This situation contrasts with the two smHCR channels, where dots remain bright at a depth of 70 μm (Figure 3a middle and top). For each of the two smHCR channels, the true positive rate is approximately >90% (Figure 3b left) and the false positive rate is approximately 20% (Figure 3b right) across the full range of depths. For the smFISH channel, the true positive rate is dramatically lower and the false positive rate is dramatically higher at all depths (Figure 3b).

To further examine the role of tissue clearing, we calculated the absolute number of colocalized dots per imaging voxel for pairs of channels with and without PACT (Figure 3c). Due to the high and ubiquitous nature of Pgk1 expression, two samples when imaged in the same relative locations (layer I-layer II\III of parietal cortex) should show similar transcript numbers per unit volume. With PACT, the two smHCR channels show no measurable decline in colocalized dot count as a function of depth, but without PACT, the dot count decreases at depths beyond ~15 μm (Figure 3c left). PACT also significantly reduces the background dot count

resulting from autofluorescence (Figure 3d). Comparisons between smHCR and smFISH emphasize the lack of signal using smFISH (Figure 3c right).

While confocal microscopy rejects out-of-focus background, image acquisition is slow, out-of-plane excitation can photobleach the sample, and the imaging depth is limited as compared to SPIM. SPIM (13-17) offers a fast alternative (~100 times faster than confocal) that rejects out-of-focus noise by illuminating and capturing images only from a thin selective plane, typically on the order of 1-10 μm. As smFISH signal is undetectable with SPIM, if SPIM, PACT, and smHCR are compatible, it would become feasible to efficiently perform phenotypical studies with single molecule resolution while preserving the natural long-range architecture of thick samples. To examine the performance of SPIM, PACT, and smHCR, we first mapped the expression patterns for two mRNAs (Ctgf and Gfap) in 250-μm brain slices, recapitulating the large-scale reference patterns in the Allen Brain Atlas (ABA) (Figure 4a and Movie S1), but now with single-molecule resolution (Figure 4b). To further characterize SPIM performance, we mapped single Scg10 mRNAs (a medium to high-copy number target) at depths up to 0.5 mm in PACT-cleared brain slices (Figure 4c and Movie S2). Examining true positive (Figure 4d) and false positive (Supplementary Figure 10) rates for three smHCR channels reveals that SPIM extends the sensitivity and selectivity achieved with confocal microscopy to significantly greater depths (see Supplementary Figures 11 and 12 for an illustration of image analysis and dot classification for smHCR/PACT/SPIM data in thick samples). Additional studies mapping a high-copy transgenic mRNA in 1-mm brain slices from Thy1-EYFP mice revealed strong and selective HCR signal at depth (Supplementary Figures 13-16 and Movie S3), although in this case the expression level of the target was too high to resolve individual dots. Notably though, as PACT-cleared tissue retains endogenous YFP fluorescence, we were able to directly test the selectivity of HCR staining without the need for parallel antibody staining; we observe a one-to-one correspondence between cells stained by YFP protein fluorescence and cells stained for YFP mRNA by smHCR.

In conclusion, we have shown that smHCR provides a robust method to map single mRNAs of varying abundance in diverse samples. In combination with PACT and SPIM, smHCR enables efficient mapping of single transcripts in thick autofluorescent brain slices, allowing the spatial architecture of the tissue to be preserved. Noting that whole bodies and a wide range of tissues, including human, have been successfully cleared (7,8), we expect the combination of smHCR, PACT,

and SPIM to enable molecular profiling of a wide variety of samples with single-cell and, if desired, single-transcript resolution while preserving geometry and connectivity information. As smHCR is compatible with sequential hybridization methods that we have previously developed (18,19), it should be possible to perform highly multiplexed studies within thick autofluorescent samples, mapping single mRNAs at depth.

2.4 Figures

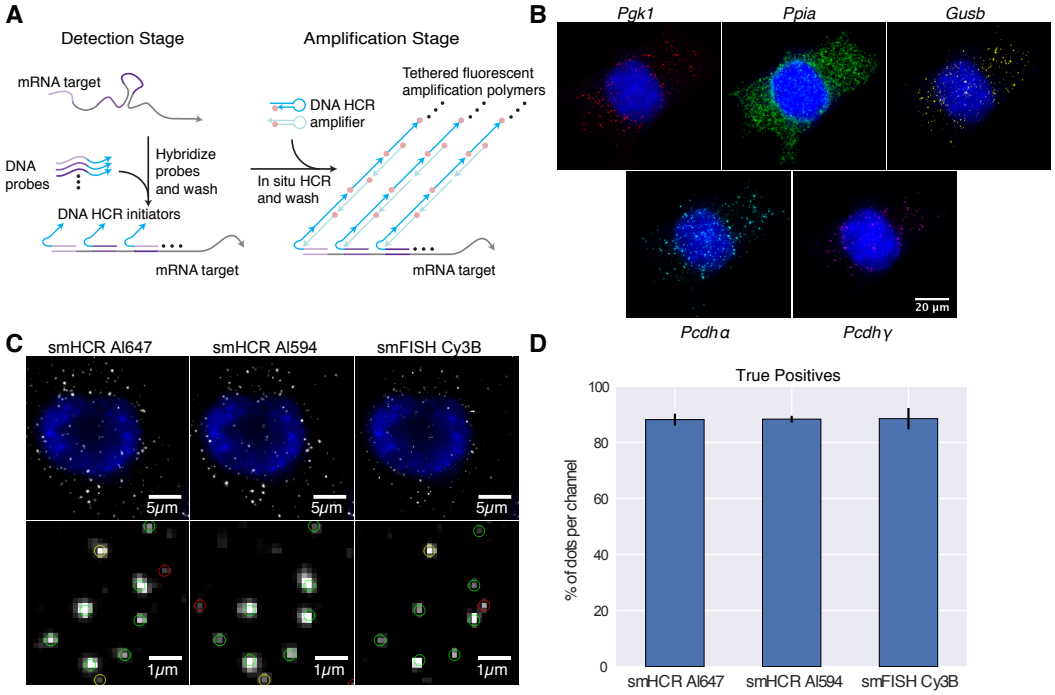


Figure 1: Single-molecule hybridization chain reaction (smHCR). **(A)** smHCR protocol. Detection stage: an mRNA target is detected by a probe set containing 20-40 short DNA probes, each binding a 20-30 nt subsequence of the target; each probe in the probe set carries an initiator for the same HCR amplifier. Amplification stage: metastable fluorophore-labeled DNA HCR hairpins penetrate the sample and self-assemble into fluorescent amplification polymers tethered to their initiating probes. The same two-stage protocol is used for multiplexed studies: during the detection stage, all probe sets are introduced simultaneously, each carrying an initiator for an orthogonal HCR amplifier; during the amplification stage, all HCR amplifiers are introduced simultaneously, each labeled with spectrally distinct fluorophores. **(B)** Simultaneous mapping of five target mRNAs in cultured CAD cells using five spectrally distinct HCR amplifiers (DAPI in blue): Pgk1 (Cy 7), Ppia (Alexa 647), Gusb (Alexa 594), Pcdhα(Cy3b), Pcdhγ(Alexa 488). **(C)** Comparison of smHCR and smFISH for detection of Pgk1 via dot colocalization in three channels: smHCR (Alexa 647), smHCR (Alexa 594), smFISH (Cy3B). Dots are classified as triple-detected true positives (present in all 3 channels; green circles), double-detected true positives (present in 2 out of 3 channels; yellow circles), or false positives (present in only one channel; red circles). **(D)** True-positive rates for each channel in panel (c) (median ± median absolute deviation; N = 10 wells). Microscopy: epifluorescence. Probe sets for panels (b-d): 22 probes per set, each addressing a 20-nt target subsequence. See Supplementary Figures 2-5 for additional data.

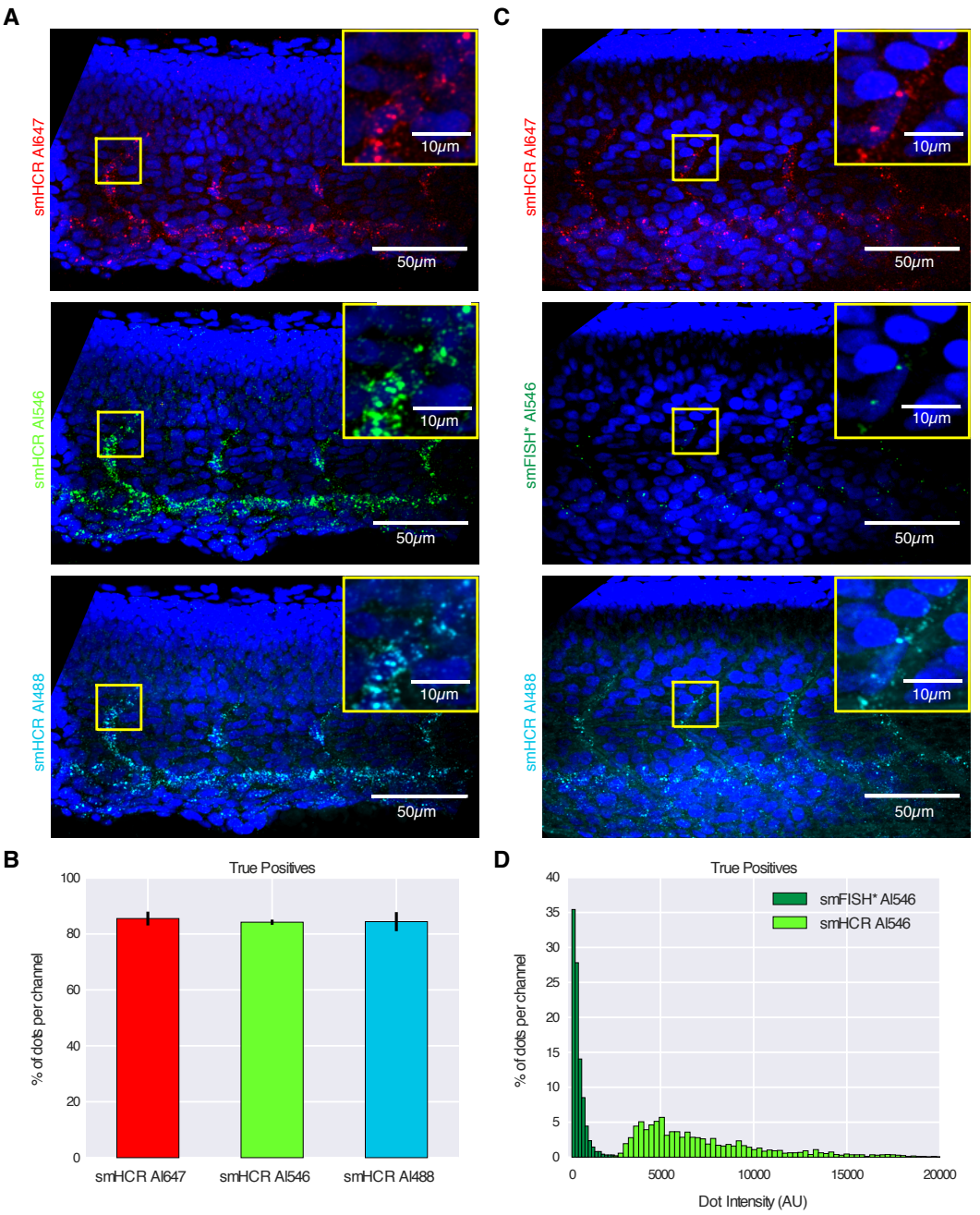


Figure 2: Imaging single mRNAs within whole-mount zebrafish embryos using smHCR. **(A)** Dot colocalization in 3-channels (DAPI in blue): smHCR (Alexa 647),: smHCR (Alexa 546), smHCR (Alexa 488). **(B)** True-positive rates for each channel in panel (a) (median ± median absolute deviation, N = 6 embryos). **(C)** Comparison of smHCR and smFISH* via dot colocalization in three channels: smHCR (Alexa 647), smHCR (Alexa 546), smHCR (Alexa 488). Channel pairs between panels (a) and (c) are shown with the same contrast; the Alexa 546 images illustrate the difference in intensity between amplified smHCR dots and unamplified smFISH* dots. **(D)** True-positive dot intensities for smHCR (Alexa 546; N=6 embryos) and smFISH* (Alexa 546; N=3 embryos). Target mRNA: kdrl (expressed in the endothelial cells of blood vessels). Microscopy: spinning disk confocal. Probe sets: 39 probes per set, each addressing a 30 nt target subsequence. Embryos fixed: 27 hpf. See Supplementary Figures 6-8 for additional data.

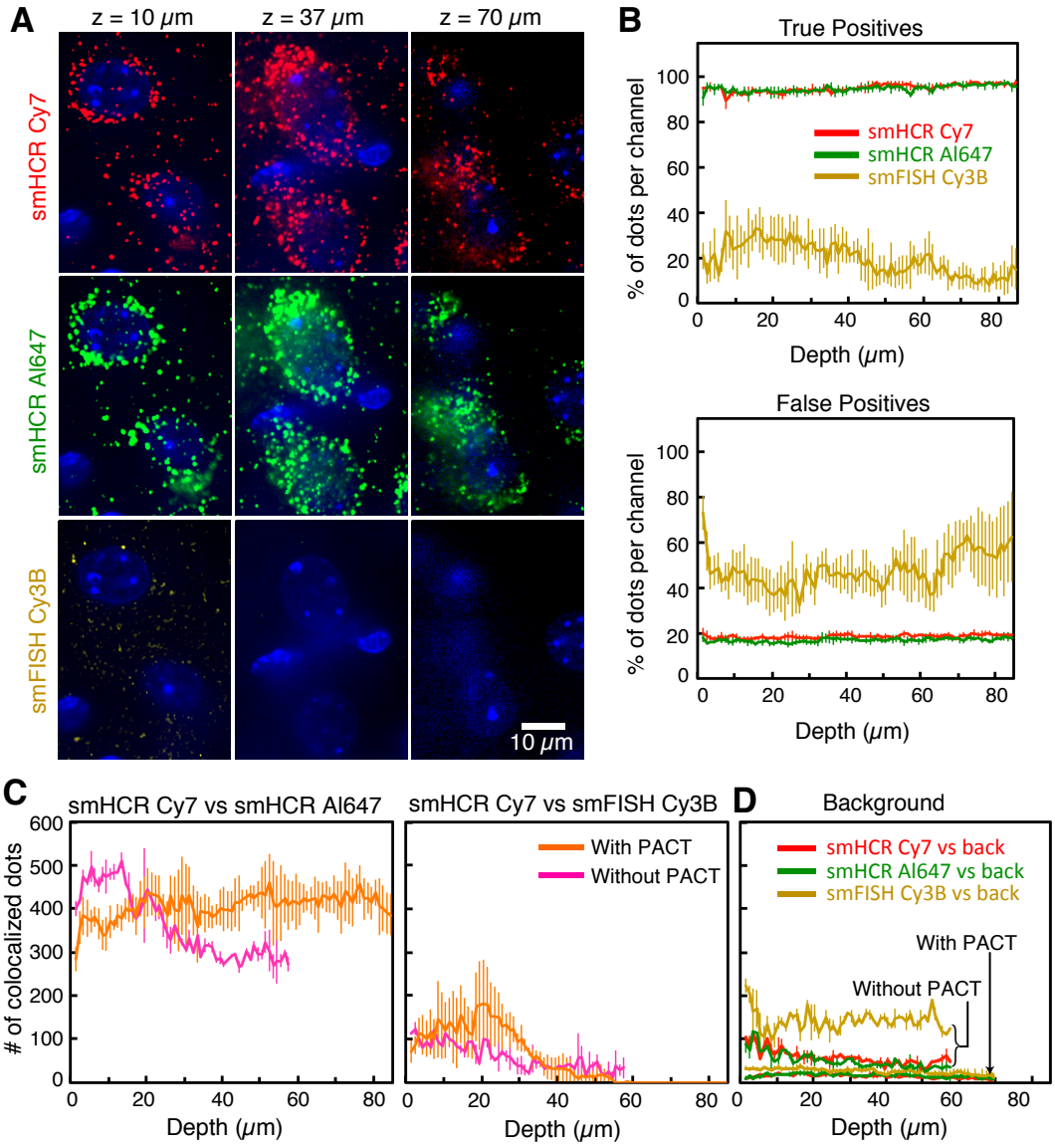


Figure 3: Imaging single mRNAs in adult mouse brain sections using smHCR and PACT. **(A)** Dot colocalization in three channels at three depths (DAPI in blue): smHCR Cy7, smHCR Al647, smFISH Cy3B. Images are displayed with the same contrast within each row. **(B)** True positive and false positive rates as a function of depth (median ± median absolute deviation, N = 8 sections). **(C)** Effect of PACT clearing on the absolute number of colocalized dots for pairs of channels as a function of depth within a 110 μm x 110 μm x 1 μm voxel (median median absolute deviation, N = 8 sections with PACT, N = 3 sections without PACT). **(D)** Characterization of background with and without PACT via colocalization of dots in any of three channels with dots due solely to autofluorescence in a fourth channel (excitation at 589 nm). (median ± median absolute deviation, N = 6 sections with PACT, N = 3 sections without PACT). Target: Pgk1. Microscopy: spinning disk confocal. Probe sets: 22 or 23 probes per set, each addressing a 20-nt target subsequence. See Supplementary Figure 9 for additional data.

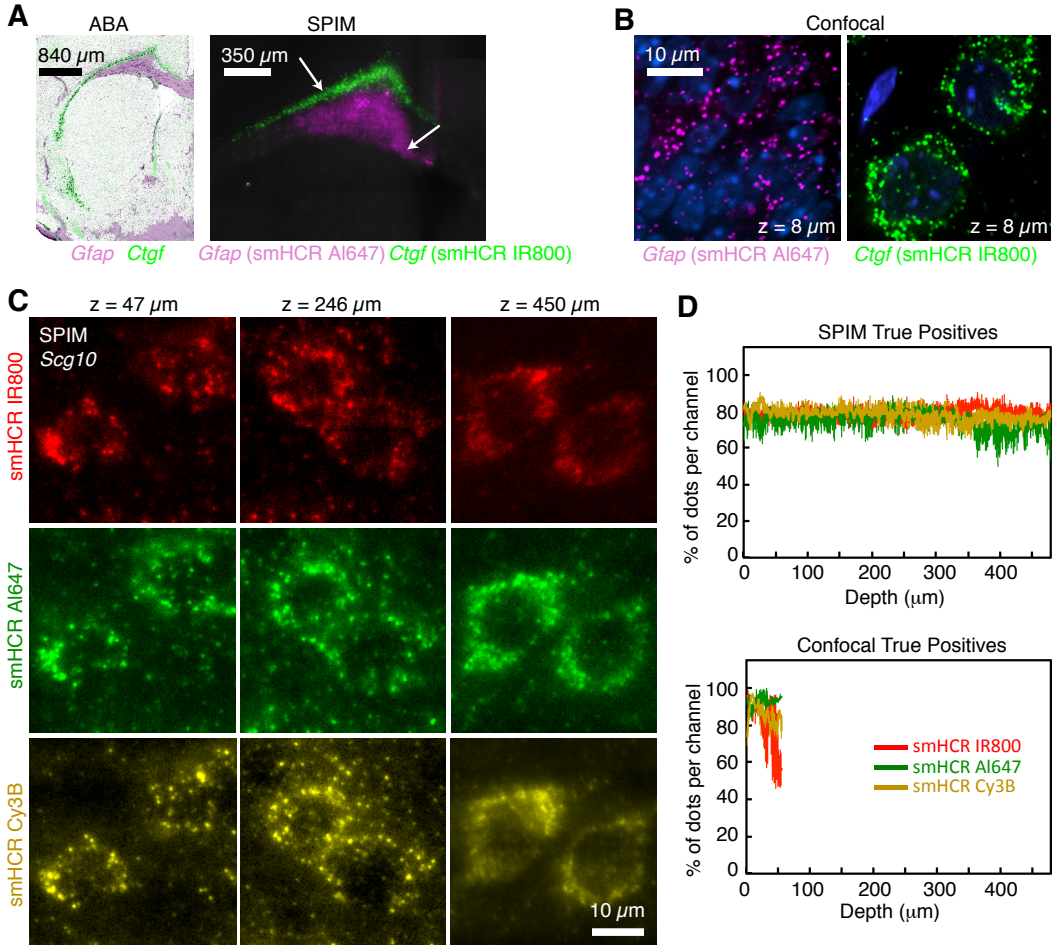


Figure 4: Imaging single mRNAs in thick adult mouse brain sections using smHCR, PACT, and SPIM. **(A)** Ctgf and Gfap expression based on: (left) reference Allen Brain Atlas (ABA) composite (Ctgf-RP_040407_02_H07-coronal slice #16 and Gfap-RP_Baylor_253913 coronal slice #16, sections were selected and overlaid based on the ventricle outline) and (right) 2-channel smHCR with PACT and SPIM. The SPIM image shows a maximum intensity projection of a 250-μm stack of images. Consistent with the ABA-based overlay, the SPIM images show Ctgf highly expressed in the deepest cortical layer and Gfap expressed in the white matter astrocytes. **(B)** High-magnification confocal images at two locations within the same sample (approximate locations denoted by arrows in panel a). DAPI in blue. **(C)** Scg10 mRNAs imaged at three depths using smHCR, PACT, and SPIM. **(D)** True positive rate as a function of depth (median ± median absolute deviation, N = 3 sections from different brains) using SPIM or confocal imaging of Scg10. Probe sets: 20 probes per set, each addressing a 20-nt target subsequence. See Supplementary Figure 10 and Supplementary Movie S2 for additional data.

2.5 Supplemental Data and Figures

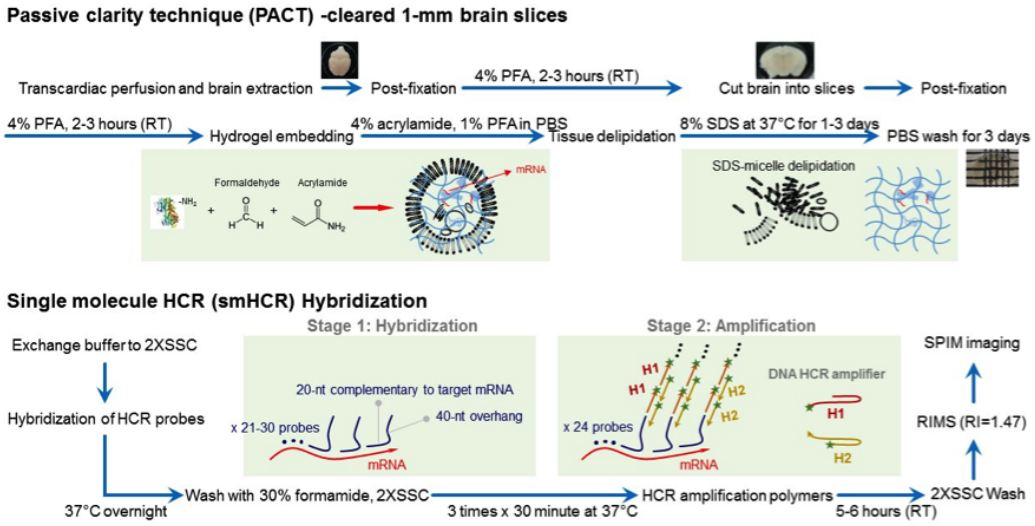


Figure S1: PACT-clearing and smHCR protocol overview. **(Top)** PACT flow chart. The main deviation from the standard PACT protocol (Yang et al., 2014) is an additional fixation step after brain slicing, in order to better preserve mRNA transcripts. **(Bottom)** smHCR flow chart. The PACT-cleared brain slices were first hybridized with DNA HCR probes (probe concentration: 1-5 nM) at 37 °C overnight, and washed in high-stringency conditions (30% formamide, 2× SSC at 37 °C for 1.5 hours) to remove the free and nonspecifically-bound probes. During HCR amplification, hairpins (120 nM) were added to the sample and incubated for 6 hours at RT. The HCR brain slices were then imaged in RIMS (refractive index 1.47; Yang et al, 2014) using SPIM microscope (Treweek et al., 2015).

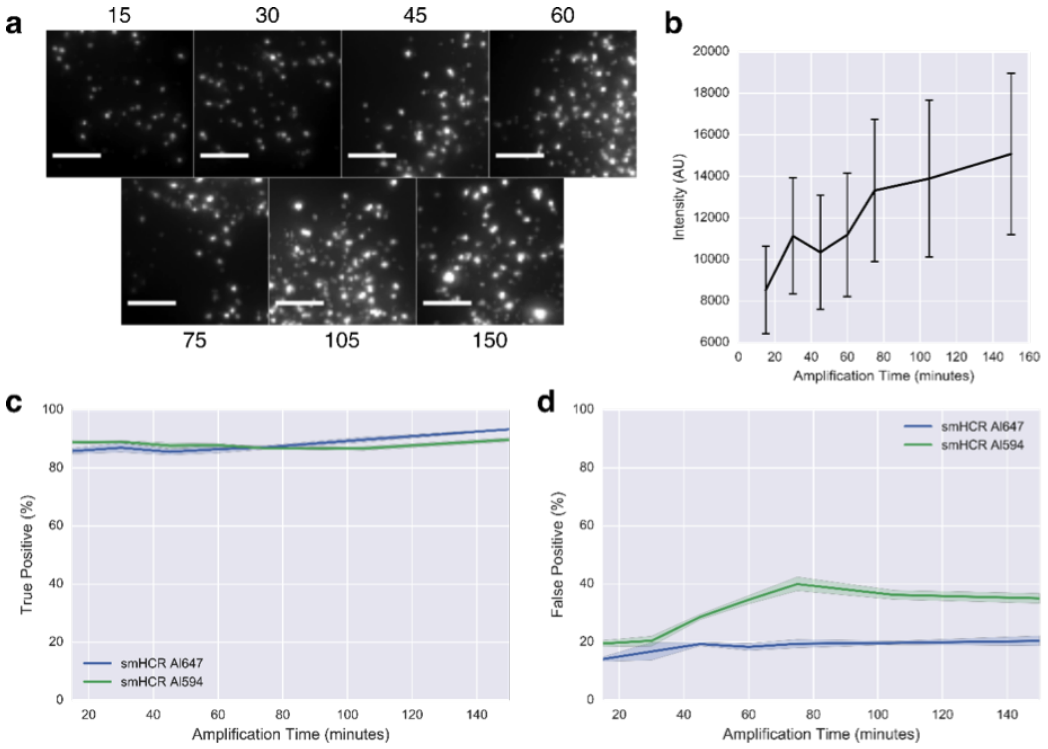


Figure S2: Single molecule HCR behavior as a function of time. **(a)** Images of mRNA transcript signals for Pgk1 transcripts at various amplification times. Images at 15, 30 and 45 minutes show mostly diffraction-limited dots. Images at 60 mins and longer show mostly non-diffraction limited dots. **(b)** Median true positive dot intensity as a function of time. **(c)** True positive rate as function of amplification time. **(d)** False positive rate as function of amplification time. Target mRNA: Pgk1. Sample: cultured CAD cells. N = 5 wells.

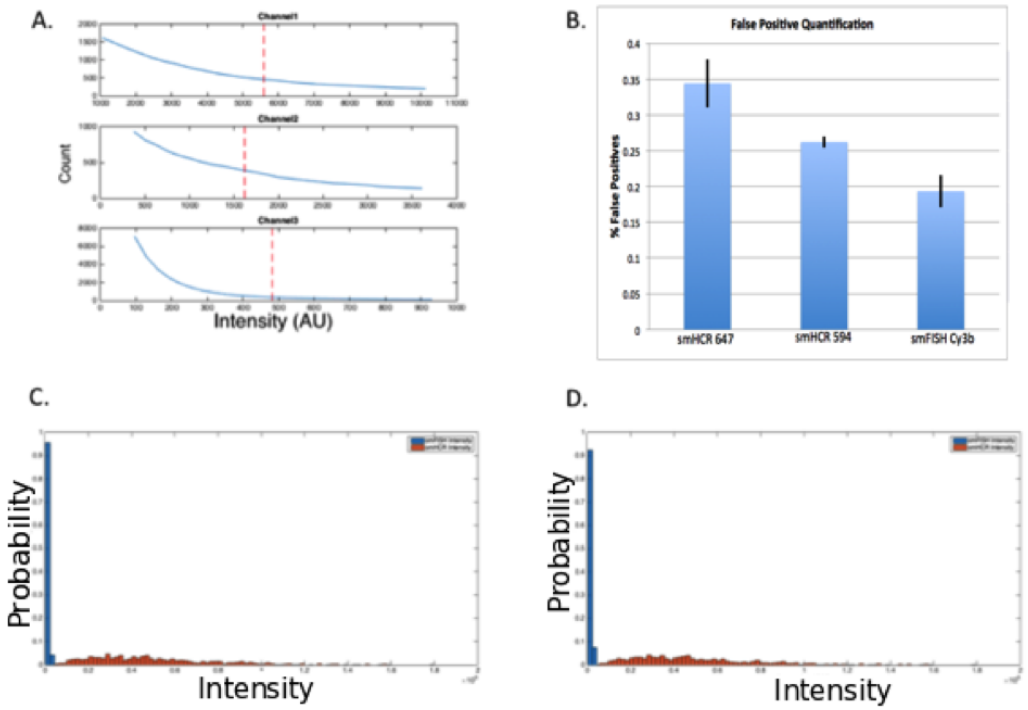


Figure S3: **A.** Thresholding for the 3-channel colocalization studies of Figure 1cd. Channel1 = smHCR B1 Alexa647, Channel2 = smHCR B3 Alexa594, Channel3 = smFISH Cy3B. Dashed lines depict thresholds selected for Figure 1cd. **B.** Quantification of false positives for the three-channel colocalization studies of Figure 1cd. A dot that is present in only one channel is interpreted as a false positive in that channel. **C.** Comparison of smFISH (Alexa 647) and smHCR (DNA HCR B1-Alexa 647) true-positive dot intensities. The ratio of median signal intensities is 36. Single molecule HCR intensity distribution is shown in orange and smFISH is shown in blue. **D.** Comparison of smFISH (Alexa 594) and smHCR (DNA HCR B3-Alexa 594) true-positive dot intensities. Single molecule HCR intensity distribution is shown in orange and smFISH is shown in blue. The ratio of median signal intensities is 13. Target mRNA: Pcdhα. Sample: cultured CAD cells.

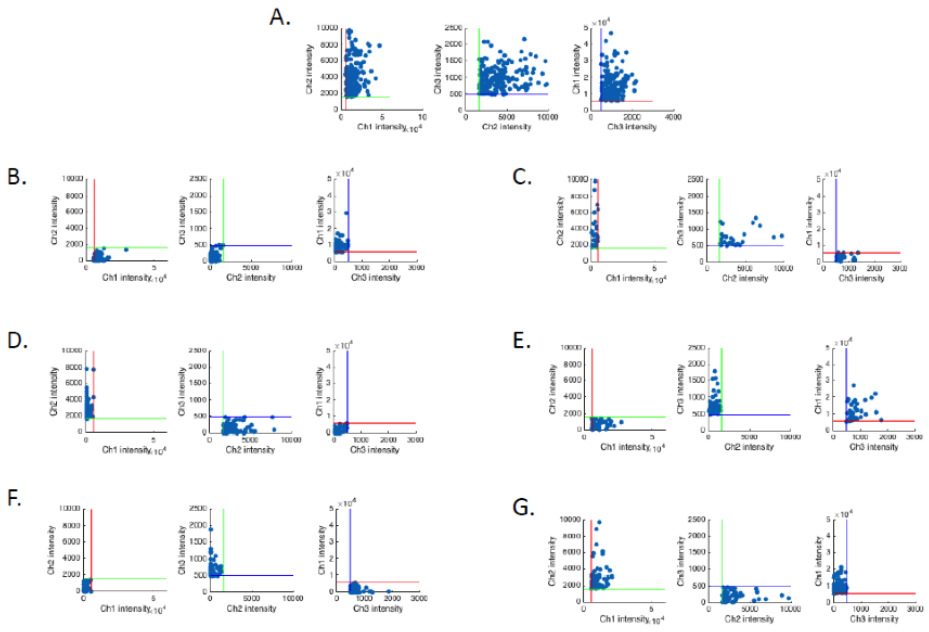


Figure S4: **A.** Pairwise intensity of true-positive dots colocalized in all three channels in the study of Figure 1cd. **B.** Pairwise intensity of false-negative dots in Channel 1 for colocalization study of Figure 1cd. A dot that is absent in only one channel is interpreted to be a false negative in that channel. **C.** Pairwise intensity of false-positive dots in Channel 1 for colocalization study of Figure 1cd. A dot that is present in only one channel is interpreted to be a false positive in that channel. **D.** Pairwise intensity of false-negative dots in Channel 2 for colocalization study of Figure 1cd. **E.** Pairwise intensity of false-positive dots in Channel 2 for colocalization study of Figure 1cd. **F.** Pairwise intensity of false-negative dots in Channel 3 for colocalization study of Figure 1cd. **G.** Pairwise intensity of false-positive dots in Channel 3 for colocalization study of Figure 1cd. Plots B, D, and F show that all transcripts are mostly detected via smHCR and smFISH, but the false positives are due to the fact that the intensities of these dots are set below the threshold value for true positive detection. The transcripts only detected in 2 out of three channels have a pair in the third channel, but that dot is below the determined noise floor. Also, plots C, E, and G show that some false positive dots also may be true positives with corresponding intensities in the other two channels being below the noise threshold. Dot intensities depicted as the maximum pixel intensity. Channel thresholds depicted as colored lines: Red line = Ch1 threshold, Green line = Ch2 threshold, Blue line = Ch3 threshold. Ch1 = Alexa 647 smHCR, Ch2 = Alexa 594 smHCR, Ch3 = Cy3B smFISH. Target mRNA: Pcdhα. Sample: cultured CAD cells.

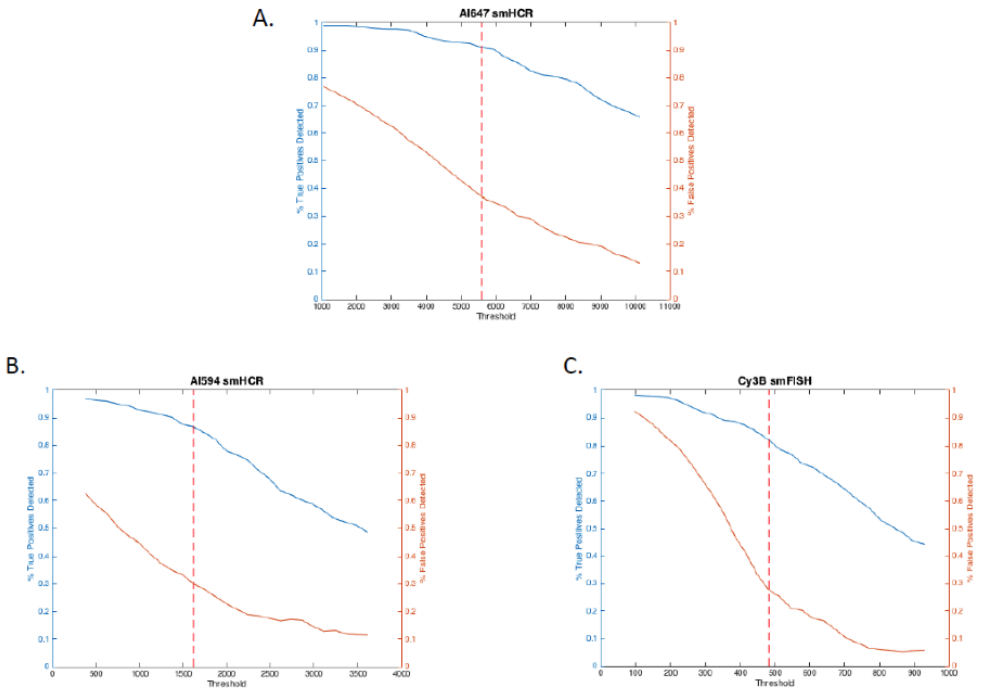


Figure S5: **A.** Tradeoff between sensitivity and selectivity (Channel 1; Alexa 647 smHCR) as a function of pixel intensity threshold for the three-channel colocalization study of Figure 1cd. **B.** Tradeoff between sensitivity and selectivity (Channel 2; Alexa 594 smHCR) as a function of pixel intensity threshold for the three-channel colocalization study of Figure 1cd. **C.** Tradeoff between sensitivity and selectivity (Channel 3; Cy3B smFISH) as a function of pixel intensity threshold for the three-channel colocalization study of Figure 1cd. Dots that are present in at least two of the three channels are classified as true positives. A dot that is absent in only one channel is interpreted as a false negative in that channel; a dot that is present in only one channel is interpreted as a false positive in that channel. Threshold used for Figure 1cd depicted as a dashed line. Decreasing the threshold increases true positives (improving sensitivity) at the cost of increasing false positives (damaging selectivity). If each target is detected in two channels, false positives can be discarded based on colocalization information, enabling near-quantitative sensitivity. Target mRNA: Pcdhα. Sample: cultured CAD cells.

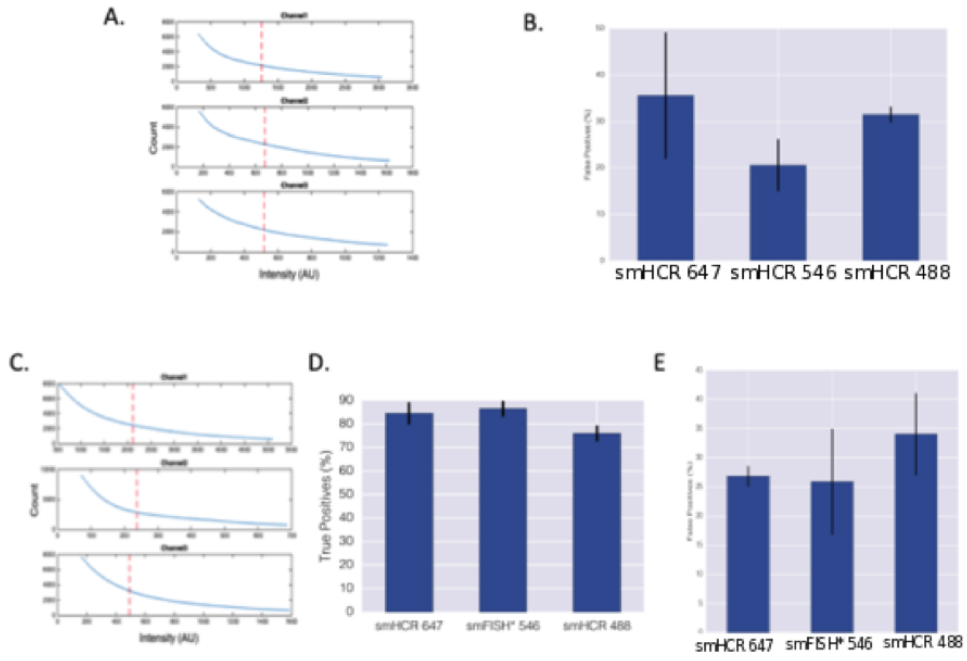


Figure S6: **A.** Thresholding for the 3-channel colocalization studies of Figure 2a. Channel1 = smHCR Alexa647, Channel2 = smHCR Alexa546, Channel3 = smHCR Alexa488. Dashed lines depict thresholds selected for Figure 2a. **B.** Quantification of false positives for the three-channel colocalization studies of Figure 2a. A dot that is present in only one channel is interpreted as a false positive in that channel. **C.** Thresholding for the 3-channel colocalization studies of Figure 2c. Channel1 = smHCR Alexa647, Channel2 = smFISH* Alexa546, Channel3 = smHCR Alexa488. Dashed lines depict thresholds selected for Figure 2c. **D.** Quantification of true positives for the three-channel colocalization studies of Figure 2c. Dots that are present in at least two of the three channels are classified as true positives. A dot that is absent in only one channel is interpreted as a false negative in that channel and as a true positive in the other two channels. **E.** Quantification of false positives for the three-channel colocalization studies of Figure 2c. A dot that is present in only one channel is interpreted as a false positive in that channel. Target mRNA: kdrl. Sample: whole-mount zebrafish embryo.

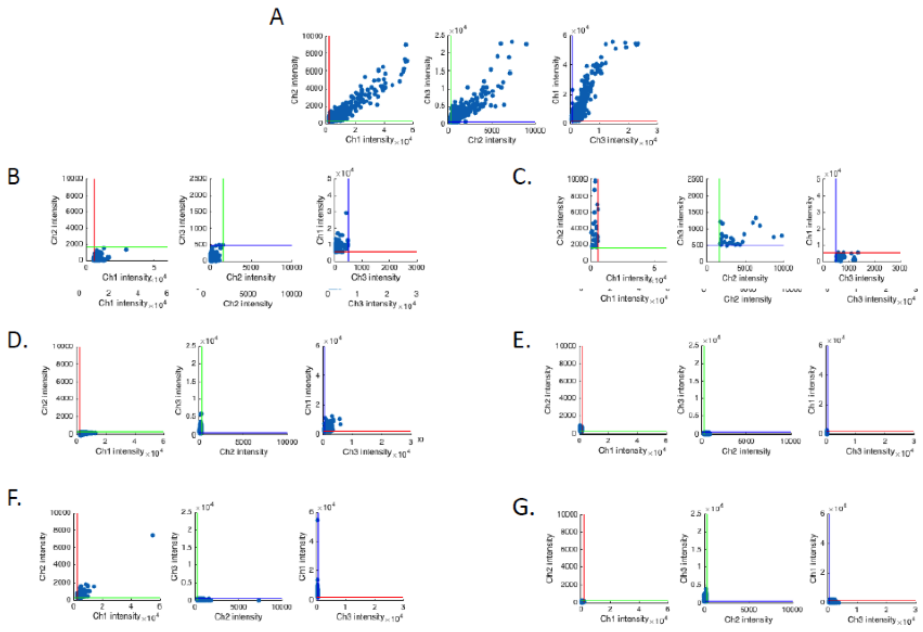


Figure S7: **A.** Pairwise intensity of true-positive dots colocalized in all three channels in the study of Figure 2c. A dot that is present in only one channel is interpreted to be a false positive in that channel. **B.** Pairwise intensity of false-negative dots in Channel 1 for colocalization study of Figure 2c. A dot that is absent in only one channel is interpreted to be a false negative in that channel. **C.** Pairwise intensity of false-positive dots in Channel 1 for colocalization study of Figure 2c. **D.** Pairwise intensity of false-negative dots in Channel 2 for colocalization study of Figure 2c. **E.** Pairwise intensity of false-positive dots in Channel 2 for colocalization study of Figure 2c. **F.** Pairwise intensity of false-negative dots in Channel 3 for colocalization study of Figure 2c. **G.** Pairwise intensity of false-positive dots in Channel 3 for colocalization study of Figure 2c. Plots B, D, and F show that all transcripts are mostly detected via smHCR in all channels, but the false positives are due to the fact that the intensities of these dots are set below the threshold value for true positive detection. The transcripts only detected in 2 out of three channels have a pair in the third channel, but that dot is below the determined noise floor. Also, plots C, E, and G show that some false positive dots also may be true positives with corresponding intensities in the other two channels being below the noise threshold. Dot intensities depicted as the maximum pixel intensity. Channel thresholds depicted as colored lines: Red line = Ch1 threshold, Green line = Ch2 threshold, Blue line = Ch3 threshold. Ch1 = Alexa647 smHCR, Ch2 = Alexa546 smFISH*, Ch3 = Alexa488 smHCR. Target mRNA: kdrl. Sample: whole-mount zebrafish embryo.

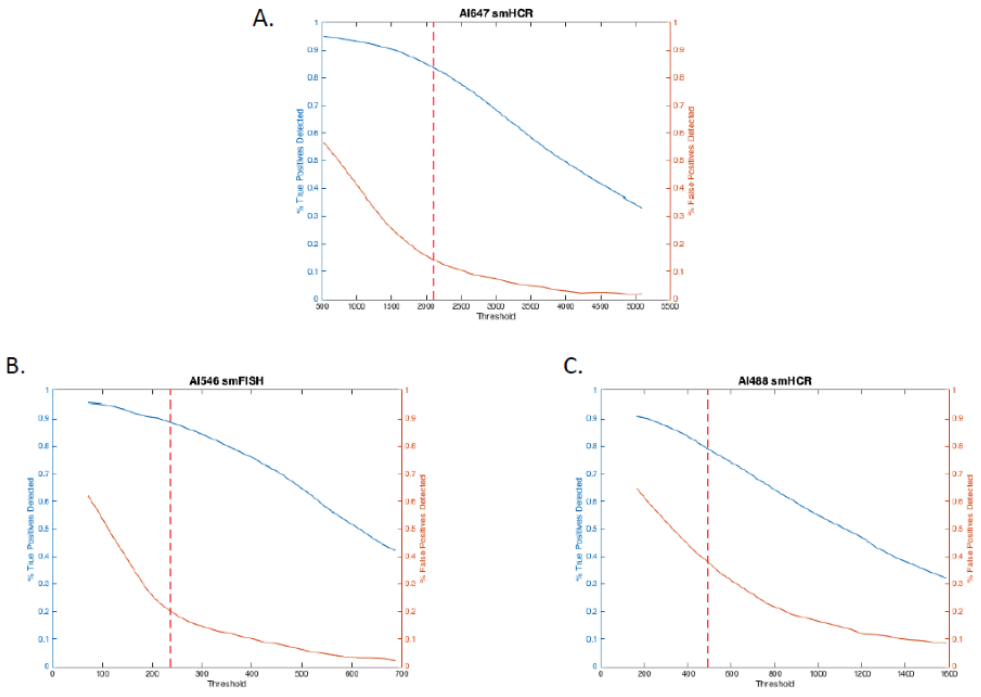


Figure S8: **A.** Tradeoff between sensitivity and selectivity (Channel 1; Alexa 647 smHCR) as a function of pixel intensity threshold for the three-channel colocalization study of Figure 2c. **B.** Tradeoff between sensitivity and selectivity (Channel 2; Alexa 546 smFISH*) as a function of pixel intensity threshold for the three-channel colocalization study of Figure 2c. **C.** Tradeoff between sensitivity and selectivity (Channel 3; Alexa 488 smHCR) as a function of pixel intensity threshold for the three-channel colocalization study of Figure 2c. Dots that are present in at least two of the three channels are classified as true positives. A dot that is absent in only one channel is interpreted as a false negative in that channel; a dot that is present in only one channel is interpreted as a false positive in that channel. Threshold used for Figure 2c depicted as a dashed line. Decreasing the threshold increases true positives (improving sensitivity) at the cost of increasing false positives (damaging selectivity). If each target is detected in two channels, false positives can be discarded based on colocalization information, enabling near-quantitative sensitivity. Target mRNA: kdrl. Sample: whole-mount zebrafish embryo.

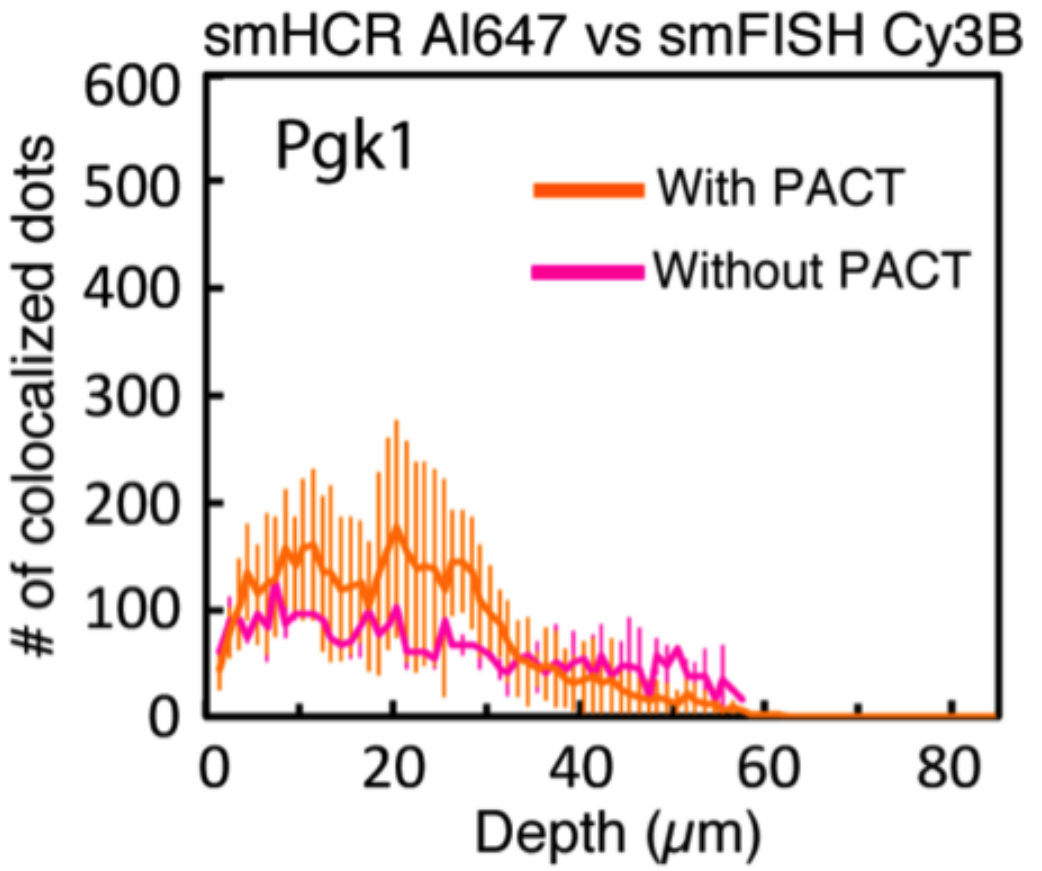


Figure S9: Imaging single mRNAs in adult mouse brain sections using smHCR and PACT. Effect of PACT clearing on the absolute number of colocalized dots for Alexa 647 (smHCR) and Cy3B (smFISH) channels as a function of depth within a 110 μm x 110 μm x 1 μm voxel (median median absolute deviation, N = 8 sections with PACT, N = 3 sections without PACT). Target: Pgk1. Microscopy: spinning disk confocal. Probe sets: 22 or 23 probes per set, each addressing a 20-nt target subsequence.

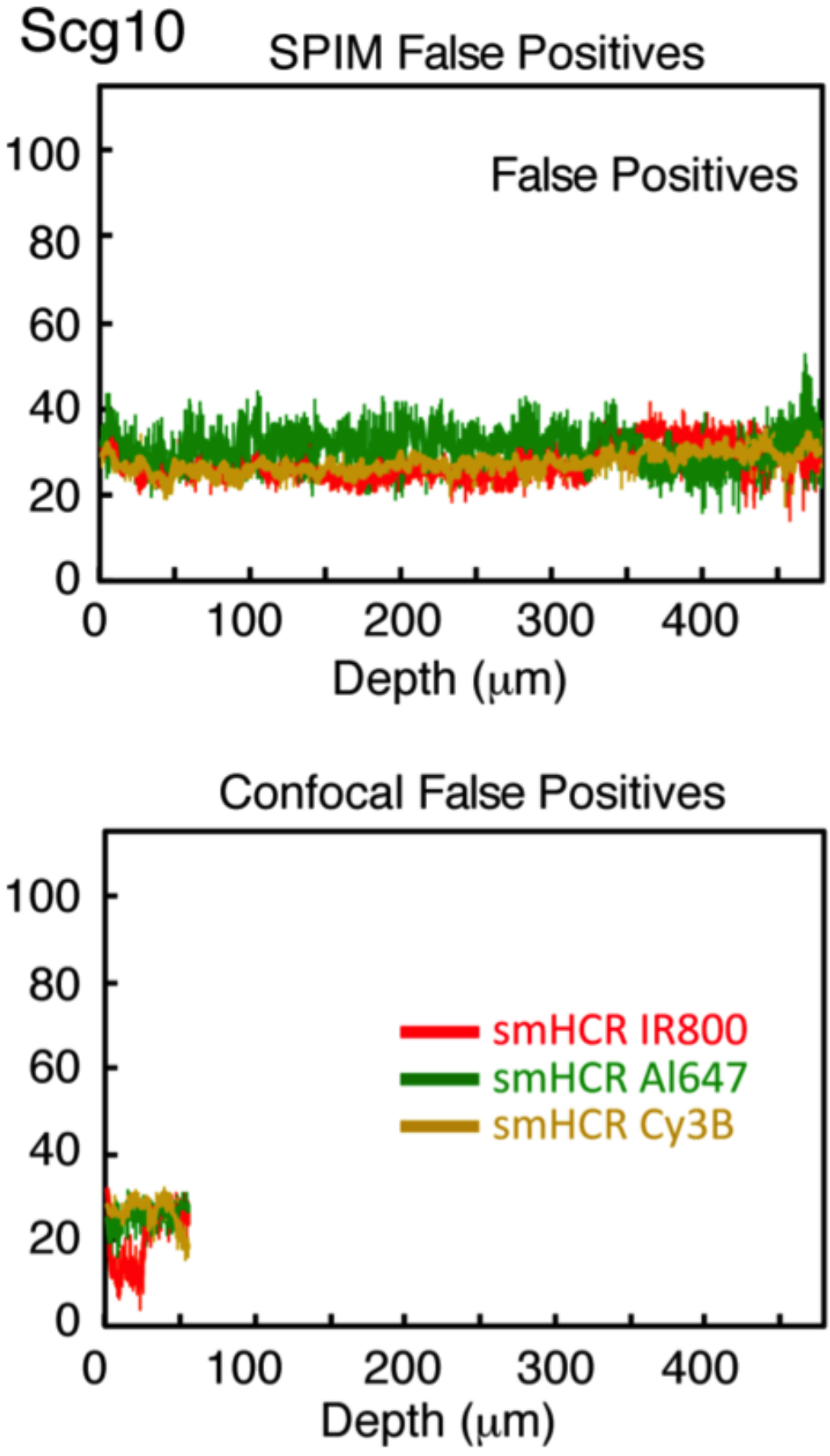

Scg10
SPIM False Positives
False Positives
100
80
60
40
20
0
0
100
200
300
400
Depth (μm)
Confocal False Positives
smHCR IR800
smHCR Al647
smHCR Cy3B
100
80
60
40
20
0
0
100
200
300
400
Depth (μm)

Figure S10 *(previous page)*: Imaging single mRNAs in thick adult mouse brain sections using smHCR, PACT, and SPIM. False positive rate as a function of depth (median median absolute deviation, N = 3 sections from different brains) using SPIM or confocal imaging of Scg10. Probe sets: 20 probes per set, each addressing a 20-nt target subsequence.

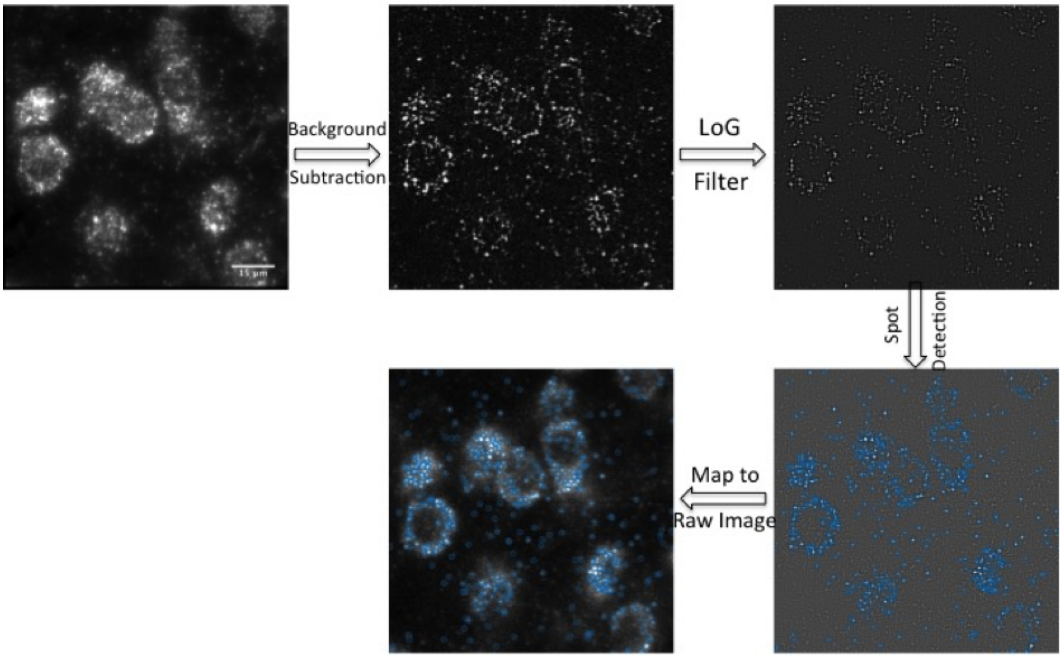


Figure S11: Processing workflow for dot finding in smHCR/PACT/SPIM images. A rolling ball background subtraction algorithm is first used to remove background noise. Next, a Laplacian of Gaussian filter is used to amplify diffraction limited dot signal. Local maxima are found in the resulting image. The locations of the local maxima are mapped back onto the raw image. Processing of the SPIM images in this way finds putative mRNA signals that would otherwise not be obvious to the eye.

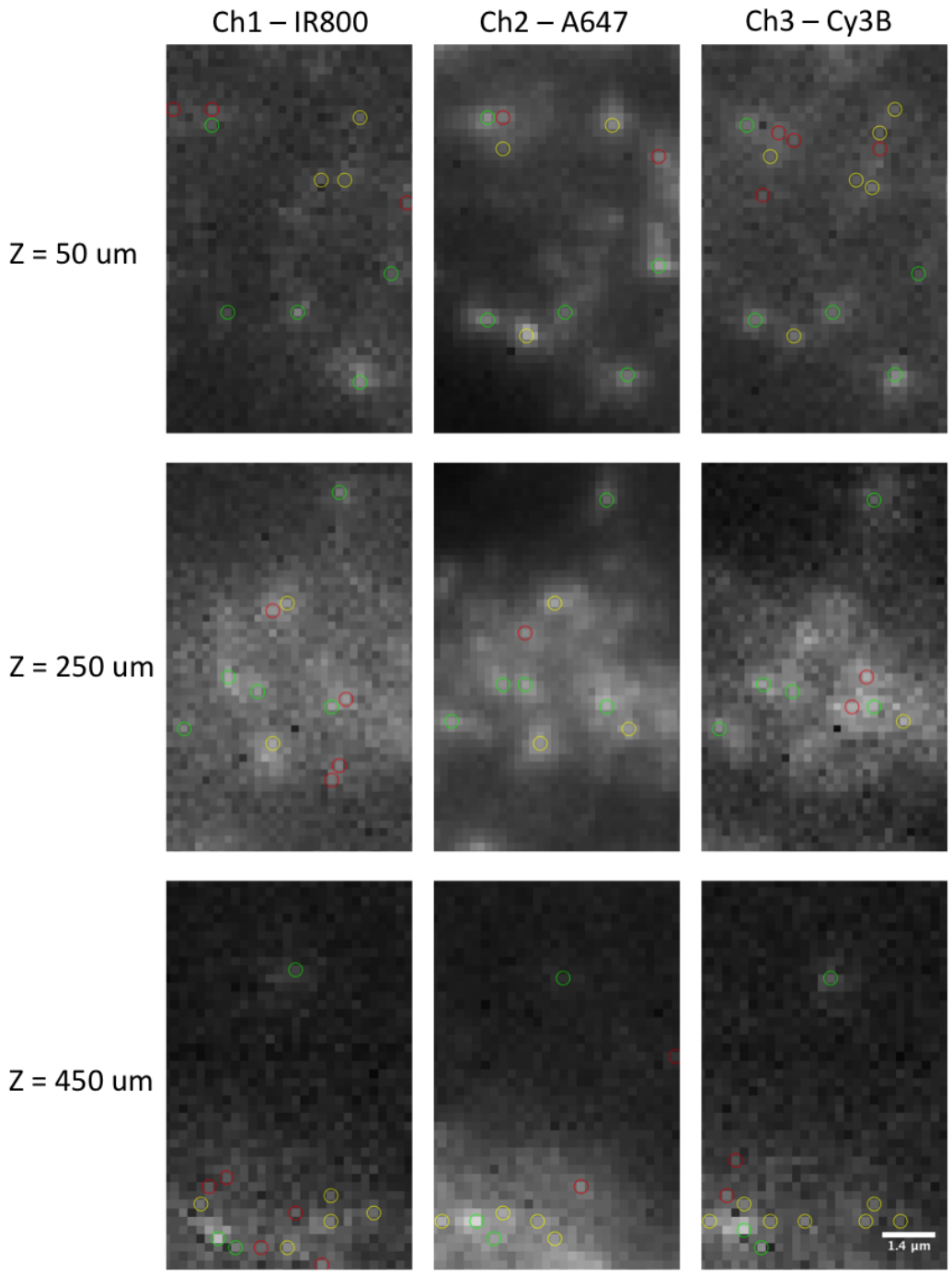


Figure S12: Illustration of dot classification for smHCR/PACT/SPIM data. Dots are classified as triple-detected true positives (present in all 3 channels; green circles), double-detected true positives (present in 2 out of 3 channels; yellow circles), or false positives (present in only one channel; red circles). Circles are superimposed on raw images.

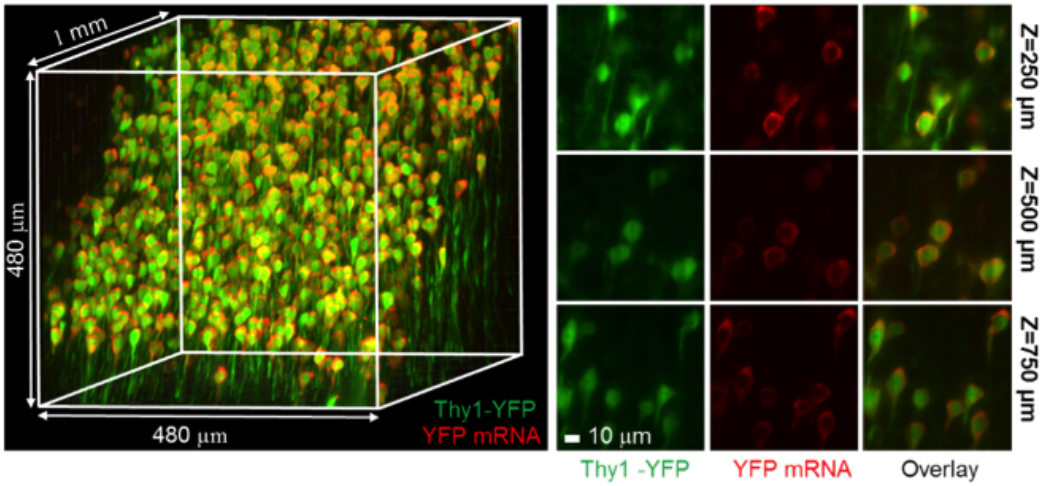


Figure S13: Mapping mRNAs in 1-mm PACT-cleared mouse brain sections using smHCR and SPIM. (Left) Three-dimensional reconstruction depicting colocalization of Thy1-YFP protein (expressed in ~5% of cells) and Thy1-YFP mRNA (smHCR: Cy3B). (Right) Representative optical sections demonstrating that smHCR signal is detected throughout the brain slice (z = 250 µm, 500 µm, 750 µm).

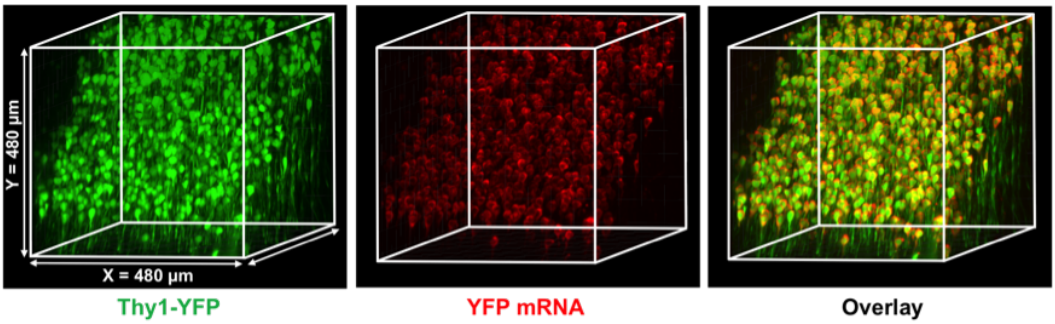


Figure S14: 3D reconstruction of 1-mm PACT-cleared Thy1-YFP mouse brain slice. (Left) Endogenous YFP Protein. (Middle) Cy3B labeling of YFP mRNA. (Right) Overlay shows the specificity of the HCR for YFP transcripts throughout the thick 1-mm slice.

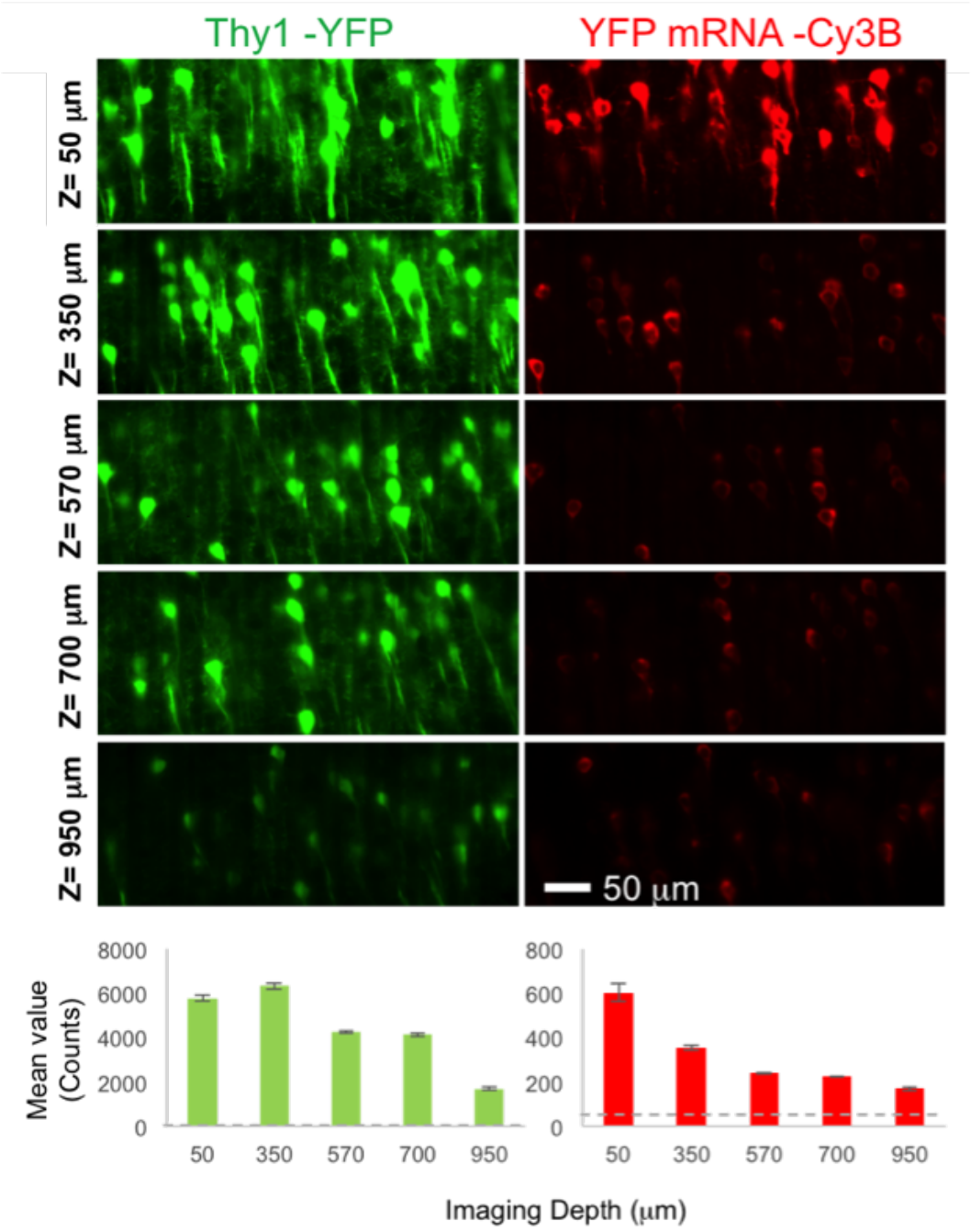


Figure S15: Fluorescence intensity decreases as a function of tissue depth. An attenuation in the fluorescence signal of the YFP and smHCR-Cy3B channels was observed as a function of tissue depth. However, even at 1-mm depth, and without light power compensation, the HCR hybridization still shows a detectable signal that co-localizes with the YFP signal. The dashed line in the lower graph indicates the camera dark count.

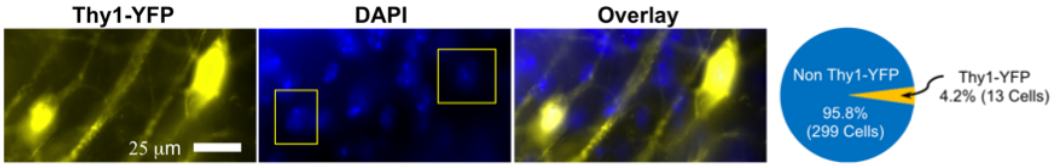


Figure S16: Across a 1-mm cleared Thy1-YFP mouse brain slice, only a small subset of brain cells show YFP expression. (Left) Widefield microscope image (YFP, DAPI, and overlay); yellow boxes in the DAPI image indicate the location of YFP expressing cells. (Right) YFP expressing cells account for less than 5% (13 cells) of the total number of cells (312 cells). Methods: 11 field-of-views of cortical areas (Olympus UPLSAPO 60x, NA 1.35, w.d. 0.15) were processed from a 1-mm Thy1-YFP brain slice. First, a group of YFP expressing cells was located and an image stack (~50 images with 1 µm spacing) was captured (DAPI and YFP). The volume of each stack was 74 µm × 74 µm × ~50 µm. Then the number of DAPI stained cells and YFP expressing cells was counted manually.

2.6 Supplemental Movies

Movie S1: ctgf-mousebrain.mov[1]

Three-dimensional reconstruction of a thick (0.25 mm) PACT-cleared mouse brain. The samples were hybridized with HCR probes against Ctgf mRNA and amplified with HCR amplifier B1-IR800. To generate a 3D reconstruction, ~200 optical sections with 1 μm spacing were taken (25 frames per second), and the intensity was normalized between the different sections.

Movie S2: scg10-mousebrain.mov

Three-dimensional reconstruction of a thick (0.5 mm) PACT-cleared mouse brain for Figure 3C, representing the colocalization of three smHCR channels (blue, green, and red are the channels of Cy3B, Al647, and IR800 respectively) for target Scg10. The movie pauses at z = 33 μm and at z = 445 μm with circles representing true mRNA signals colocalized in two or more of the three channels. About 500 optical sections spaced with 1 μm were taken (10 frames per second), and the intensity was normalized between the different sections. In each channel, the maximum intensity projection of every 5 sections (~5 μm) was used to construct the movie.

Movie S3: yfp-mousebrain.mov

Three-dimensional reconstruction of a thick (1 mm) PACT-cleared mouse brain for Supplementary Figures 13-15. To generate a 3D reconstruction, ~1,000 optical sections with 1 μm spacing were taken (25 frames per second), and the intensity was normalized between the different sections.

[1] All movies can be found on the Dryad Digital Content repository at the following link https://datadryad.org/resource/doi:10.5061/dryad.31r0d

2.7 Supplemental Tables

All supplementary tables can be downloaded as Microsoft Excel files from the following link:

http://dev.biologists.org/content/develop/suppl/2016/07/26/dev.138560.DC1/DEV138560supp.pdf

2.8 References

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

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

3. Levsky, J.M., Shenoy, S.M., Pezo, R.C., Singer, R.H. Single-cell gene expression profiling. Science 297, 836-840 (2002).

4. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., Tyagi, S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 4, e309 (2006).

5. Fan, Y., Braut, S.A., Lin, Q., Singer, R.H., Skoultchi, A.I. Determination of transgenic loci by expression FISH. Genomics 71, 66-69 (2001).

6. Chung, K., Wallace, J., Kim, S.Y., Kalyanasundaram, S., Andalman, A.S., Davidson, T.J., Mirzabekov, J.J., Zalocusky, K.A., Mattis, J., Denisin, A.K., Pak, S., Bernstein, H,, Ramakrishnan, C., Grosenick, L., Gradinaru, V., Deisseroth, K. Structural and molecular interrogation of intact biological systems. Nature 497, 332-337 (2013).

7. Treweek, J.B., Chan, K.Y., Flytzanis, N.C., Yang, B., Deverman, B.E., Greenbaum, A., Lignell, A., Xiao, C., Cai, L., Ladinsky, M.S., Bjorkman, P.J., Fowlkes, C.C. and Gradinaru, V. Whole-Body Tissue Stabilization and Selective Extractions via Tissue-Hydrogel Hybrids for High Resolution Intact Circuit Mapping and Phenotyping. Nat Prot, 2015; doi:10.1038/nprot.2015.122

8. Yang, B., Treweek, J.B., Kulkarni, R.P., Deverman, B.E., Chen, C.-K., Lubeck, E., Shah, S., Cai, L., and Gradinaru, V. Single-Cell Phenotyping within Transparent Intact Tissue through Whole-Body Clearing. Cell 158, 1-14 (2014).

9. Player, A.N., Shen, L.P., Kenny, D., Antao, V.P., Kolberg, J.A. Single-copy gene detection using branched DNA (bDNA) in situ hybridization. J Histochem Cytochem. 49, 603-612 (2001).

10. Choi, H.M.T., Chang, J.Y., Trinh, L.A., Padilla, J.E., Fraser, S.E., and Pierce, N.A. Programmable in situ amplification for multiplexed imaging of mRNA expression Nature Biotechnol. 28,1208-1212 (2010).

11. Choi, H.M.T., Beck, V.A., Pierce, N.A. Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. ACS Nano. 8, 4284-4294 (2014).

12. Oka, Y., Sato, T.N. Whole-mount single molecule FISH method for zebrafish embryo. Sci Rep. 5, 8571 (2015).

13. Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J., and Stelzer, E. H. "Optical sectioning deep inside live embryos by selective plane illumination microscopy". Science 305, 1007-1009 (2004).

14. Dodt, H.U., Leischner, U., Schierloh, A., Jährling, N., Mauch, C.P., Deininger, K., Deussing, J.M., Eder, M., Zieglgänsberger, W., and Becker, K. Ultramicroscopy: three-dimensional visualization of neuronal networks in the whole mouse brain. Nature Method 4, 331-336 (2007).

15. Keller, P.J., Schmidt, A.D., Santella, A., Khairy, K., Bao, Z., Wittbrodt, J., and Stelzer, E.H. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. Nature methods 7, 637-642 (2010).

16. Tomer, R., Ye, L., Hsueh, B., and Deisseroth, K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. Nat. Prot. 9, 1682-1697 (2014).

17. Baumgart, E., and Kubitscheck, U. Scanned light sheet microscopy with confocal slit detection. Optics express 20, 21805- 21814 (2012).

18. Lubeck, E., Coskun, A., Zhiyentayev, T., Ahmed, M., Cai, L. Single cell in situ RNA profiling by sequential hybridization. Nature Methods 11, 360–361 (2014).

19. Lubeck, E., Cai, L. Single cell systems biology by super-resolution imaging and combinatorial labeling. Nature Methods 9, 743–748 (2012).

20. Stapel, L.C., Lombardot, B., Broaddus, C., Kaimueller, D., Jug, F., Myers, E.W., Vastenhouw, N.L., Automated detection and quantification of single RNAs at cellular resolution in zebrafish embryos. Development 143, 540-546, 2016.

21. Dirks, R.M., Pierce, N.A., Triggered amplification by hybridization chain reaction, Proc Natl Acad Sci USA, 101(43):15275-15278, 2004.

2.9 Methods

Cell Culture

Sample Preparation

CAD cells were obtained from Sigma Aldrich (Sigma, Cat. # 08100805) cultured as described previously (Suri et al., 1993). CAD cells were grown in DMEM/F-12 medium (ThermoFisher, Cat. # 12634), supplemented with 10% FBS (Gibco, Cat. # 16000) and 1% penicillin-streptomycin (Gibco, Cat. # 15140) on standard tissue culture flasks in a humidified 5% C02 incubator. CAD cells were passaged every 3 to 4 days and cells were plated at a 1:10 dilution. For microscopy experiments, 22 mm × 40 mm no. 1.5 coverslips (EMS, Cat. # 72204) were coated with 0.01% poly-D-lysine (Sigma, Cat. # P7280) for 2 hours at 37 °C in standard tissue culture dishes. Coverslips were washed with UltraPure water (ThermoFisher, Cat. # 10977) and cells were plated on the coverslip at a 1:6 dilution. Cells were grown overnight and fixed the following day with 4% formaldehyde in 1× PBS for 15 minutes at room temperature. Formaldehyde was then washed out with RNAse-free 2× SSC and stored in 70% ethanol and stored at -20oC.

In situ hybridization

Probe design and synthesis

DNA 20-nt probes complementary to the target mRNA were designed using Stellaris Designer (Supplementary Tables 1 and 2). For smFISH probes, 5' amine modified oligos were purchased and were directly coupled to a fluorophore and purified as described in Lubeck et al. 2012. smHCR probes are 60-nt long with a 20-nt long mRNA recognition sequence, 4-nt spacer and a 36-nt HCR initiator on the 3‘ end.

Probe hybridization

Coverslips coated with cells were removed from 70% EtOH storage at -20 °C. Coverslips were air-dried to remove all traces of EtOH and rehydrated in 2× SSC for 5 minutes. Samples were then hybridized overnight at 1 nM probe concentration in 10% hybridization buffer at 37 °C. 10% hybridization buffer is composed of 10% formamide (Ambion, Cat. # AM9342), 10% Dextran Sulfate (Sigma-Aldrich, Cat. # D6001) and 2× SSC (Sigma-Aldrich, Cat. # 93017) in RNAse-Free H2O (Life Technologies, Cat. # 10977-015). While higher concentrations of probe often generate more signal per mRNA molecule, the corresponding increase in background generally negates any advantages.

Probe wash

Following hybridization, samples were washed in a solution of 30% formamide, 2× SSC, 0.1% Triton-X 100 (Sigma-Aldrich, Cat. # T8787) for at least 15 minutes. Increasing the wash duration appeared to have little effect on signal or background. All traces of wash buffer and unbound probe were removed by multiple rinses in 2× SSC.

HCR amplification

During the wash step above, fluorophore-labeled HCR hairpins purchased from Molecular Instruments (molecularinstruments.org) were thawed out of -20 °C storage and snap-cooled (heat at 95 °C for 2 minutes and cool to room temperature for 30 minutes) before use. Amplification was performed for 45 minutes at room temperature at a concentration of 120 nM per hairpin in an amplification buffer consisting of 10% Dextran Sulfate and 2× SSC in RNAse-Free water. Amplification times can be varied to modify polymer length. Following amplification, samples were washed in the wash buffer described above for at least 10 minutes to remove unbound hairpins.

Imaging

After amplification, samples were briefly stained with DAPI (Life Technologies, Cat. # D1306), rinsed in 2× SSC and placed in an enzymatic anti-bleaching buffer described in Lubeck et al. 2014. Samples were imaged on Nikon Ti-Eclipse microscopes. Due to the high gain of smHCR, samples could be easily imaged on microscopes equipped with either a solid state laser or a xenon arc lamp. All publication data was obtained using laser illumination. All microscopes used high quantum efficiency (QE) Andor iKon-M cameras, but in principle, lower QE cameras can be used to image smHCR.

Image Processing and Analysis

All image analysis was performed on three-dimensional image stacks in MATLAB. Before image analysis, the background in each image was subtracted using the ImageJ rolling ball background subtraction algorithm with a radius of 3 pixels. Next, regions of interest were selected to avoid issues of non-uniform illumination and spherical aberrations on the edges on images. The images were then corrected for chromatic aberrations by using multi-spectral beads to determine a geometric transformation to align all channels. Once the images were sufficiently aligned, local maxima within the image were found after first thresholding the images based on

previously published smFISH thresholding methods (Raj et al., 2008). Transcripts for smFISH images were found using previously published LOG filter and local maxima finding methods (Raj et al., 2008).

Transcripts for smHCR images were found by setting all pixels values below a threshold to zero and then finding local maxima. A mask of shape [1 1 1; 1 0 1; 1 1 1] is used to dilate (MATLAB function "imdilate") the grayscale image. The center of the mask is positioned at each pixel, and the pixel value is replaced by the maximum value found at the neighboring pixels where there is a value of 1 in the 3-by-3 matrix. The resulting image is such that every pixel value is replaced by the maximum value of its 8 neighbors. The resulting image is compared to the original image such that every pixel that is greater in the original image than in the dilated imaged is a local maxima of the image. Due to the differing profiles of the smHCR dots from the smFISH dots, LOG filtering to find HCR dots often results in errors in dot positions. Once dots were found in all images, the local maxima were matched using a maximum 1-pixel tolerance symmetric nearest neighbor search (SNNS). SNNS only matches dots that are the closest to each other and not to any other point. Analysis software can be obtained upon request.

We define true mRNA signals as those dots that are colocalized in at least two of the three channels. We calculate a true positive rate for channel x as the percentage of true mRNA signals detected as dots in channel x:

$$\%TP_x = 100N_x^{true}/N_{mRNA} \tag{2.1}$$

where N_x^{true} is the number of dots representing true mRNAs detected in channel x, and N_{mRNA} is the total number of true mRNAs detected across the three channels. We calculate a false positive rate for channel x as the percentage of dots in channel x that are not true mRNA signals:

$$\%FP_x = 100N_x^{false_x}/(N_x^{true} + N_x^{false}) \tag{2.2}$$

where N_x^{false} is the number of dots detected in channel x that do not represent true mRNAs (i.e., dots in channel x that do not colocalize with dots in either of the other channels).

Whole Mount Zebrafish

The following protocol is adapted from Choi et al. 2014.

Sample Preparation

1. Collect zebrafish embryos and grow in a petri dish with egg H2O at 28 °C until 27 h post-fertilization (27 hpf).
2. Dechorionate embryos, transfer to an eppendorf tube and remove excess egg H2O.
3. Fix zebrafish embryos in 1mL freshly prepared 4% paraformaldehyde (PFA) for 24 hr at 4 °C.
4. Wash embryos 3 x 5 min with PBS to stop the fixation. Fixed embryos can be stored at 4 ºC at this point.
5. Dehydrate and permeabilize with a series of methanol (MeOH) washes:
   a. 100% MeOH for 4 x 10 min
   b. 100% MeOH for 50 min.
6. Rehydrate with a series of graded MeOH / PBST washes for 5 min each:
   a. 75% MeOH / 25% PBST
   b. 50% MeOH / 50% PBST
   c. 25% MeOH / 75% PBST
   d. 5 x 100% PBST
7. Store embryos at 4C before use.

In situ hybridization

Probe design and synthesis

smHCR probes for the mRNA target (kdrl) are 71 nt long (30-nt mRNA recognition sequence, 5-nt spacer, 36-nt initiator). Each probe set contains 39 1-initiator DNA probes (Supplementary Table 3; sequences listed 5'to 3'). Probes were designed to have $\geqslant$ 5 nt gaps between each other along the target mRNA . The order of probe binding is alternating between probe sets along the mRNA. Within a given probe set, each probe initiates the same DNA HCR amplifier. Probes were purchased from Molecular Instruments.

Probe hybridization

1. Pre-hybridize approximately 6 embryos in pre-hybridization buffer for 30 min at 65 °C. Pre-hybridization buffer is composed of 50% formamide (Ambion,

Cat. # AM9342), 5× SSC (Invitrogen, Cat. # 15557-044), 9 mM citric acid pH 6 (Mallinckrodt Chemicals, Cat. # 0627-12), 0.1% Tween-20 (Bio-Rad, Cat. # 161-0781), 50 µg/mL heparin (Sigma, Cat. # H3393), 1× Denhardt's solution (Invitrogen, Cat. # 750018) and 10% dextran sulfate (Sigma, Cat. # D6001).

2. Meanwhile, prepare the probe solution:

   a. Each probe should be at a concentration of 2 nM.

   b. Create a probe mixture in probe hybridization buffer of the desired volume. Hybridization buffer composition is identical to the pre-hybridization buffer with 30% formamide instead of 50%.

   c. Warm the mixture to 37 °C.

3. Remove the pre-hybridization solution from samples and add the probe solution.

4. Incubate overnight (>12 hrs) at 37 °C.

Probe wash

Following hybridization, samples were washed at 37 °C in a series of graded probe wash buffer / 5× SSCT. Probe wash buffer is composed of 30% formamide, 5× SSC, 9 mM citric acid pH 6, 0.1% Tween-20, and 50 µg/mL heparin.

1. 75% probe wash buffer / 25% 5× SSCT for 15 minutes
2. 50% probe wash buffer / 50% 5× SSCT for 15 minutes
3. 25% probe wash buffer / 75% 5× SSCT for 15 minutes
4. 100% 5× SSCT for 15 minutes
5. 100% 5× SSCT for 30 minutes

HCR amplification

1. Pre-amplify samples in amplification buffer (5x SSC, 0.1% Tween-20, 10% dextran sulfate) for 30 min at room temperature.

2. Meanwhile, prepare the hairpin solution:

   a. Hairpins should first be snap cooled individually at the provided concentration – 3 μM. Only snap cool the amount of hairpin required for this experiment: hairpins are used at a final concentration of 60 nM.

   b. Snap cool each hairpin separately by heating to 95 °C for 90 seconds and cooling to room temperature in a dark drawer for 30 min.

   c. Mix the two hairpins together in amplification buffer to achieve a final concentration of 60 nM for each hairpin.

3. Remove the pre-amplification solution from samples and add the hairpin solution.

4. Incubate the samples for 1 hr at room temperature.

Hairpin washing

Remove excess hairpins by washing at room temperature:

1. 2 x 5 min of 5x SSCT
2. 2 x 30 min of 5x SSCT
3. 1 x 5 min of 5x SSCT

Imaging

Prior to imaging, the 5× SSCT was replaced with SlowFade Gold Antifade Mountant with DAPI (ThermoFisher Scientific, Cat. # S36937). The head and yolk sac were resected and fish tails were mounted sandwiched between two no. 1 coverslips (VWR, Cat. # 48393-106). Images were collected on an Andor CSU-W1 spinning disk confocal on a Nikon Ti-E with Perfect Focus System microscope equipped with a Plan Apo λ 60×/1.4 and an Andor iXon ULTRA 888BV camera.

Image Processing and Analysis

Image processing and analysis is performed as described above for cell culture.

Adult mouse brain slices

Sample Preparation

All procedures were done in an RNase-free environment. (Early) adult mice (36-100 days old, C57BL/6N-Tac1-IRES2-Cre-D or C57BL/6J Thy1-yfp, female) were subjected to standard transcardiac perfusion with ice-cold 4% PFA in 1x PBS followed by brain extraction and post-fixation in 4% PFA for 2-3 hours at room temperature (RT). Using a vibratome (VT1200S, Leica) or a mouse brain matrix (Kent Scientific Corp., Cat. # RBMS-200C), the brain was cut into 0.25-1 mm thick slices, and further post-fixed for an additional 2-4 hours at RT in order to better preserve mRNA transcripts.

PACT clearing

The above PFA-fixed brain slices were then PACT cleared following the previously reported protocol (Yang et al. 2014, Treweek et al. 2015). Briefly, the slices were incubated in hydrogel monomer solution containing either 4% acrylamide (A4P0) or 4% acrylamide with 1% PFA (A4P1) in 1x PBS supplemented with 0.25% photoinitiator 2,2'-Azobis[2-(2-imidazolin-2-yl) propane] dihydrochloride (VA-044) (Wako Chemicals USA, Cat. # 011-19365) at 4 °C overnight. The following day the samples were rigorously degassed using nitrogen gas for 8-10 minutes, and the tissue-gel hybrid was constructed at 37 °C for ~3 hours. After multiple rounds of 1x PBS washes, the samples were transferred to 50 mL conical tubes pre-filled with detergent solution (8% SDS in 1x PBS, pH 7.6), and incubated at 37 °C for 1-3 days until the samples were optically transparent. Before HCR in situ hybridization, the brain slices were extensively washed with 1x PBS for 2-3 days in order to remove residual SDS in the sample.

In situ hybridization

Probe design and synthesis

Probes for the Pgk1 transcripts hybridized in mouse brain slices are the same probe set as used for cultured CAD cells. For all smHCR probes, each probe contains 20-nt mRNA recognition sequence designed using Stellaris Designer, 4-nt spacer, and 36-nt initiator. The probe set contains 21-32 DNA probes (Supplementary Table 4).

Probe hybridization

smHCR was performed in a 24 well plate, and during hybridization each well

was filled with 2-ml buffer solutions and constantly shaking by a 3D rotating mixer. Each slice of brain samples was transferred to an individual well, and first pre-incubated in 2× SSC for 1 hour followed by the incubation of hybridization buffer (10% formamide, 10% dextran sulfate, 2× SSC) containing DNA HCR probes (1-5 nM per probe) overnight (> 12 hrs) at 37 °C.

Probe wash

The following day, the samples were washed by three rounds of wash buffer (30% formamide, 2× SSC at 37 °C for 30 minutes in each round) to remove free and nonspecifically-bound probes. Samples were then rinsed with 2 more rounds of 2× SSC solution at RT.

HCR amplification

The HCR amplification stage was performed according to the next-generation HCR protocol (Choi et al., 2014). At the amplification stage, in a 24-well plate each well was filled with 500 µl of amplification buffer (minimum amount of buffer volume sufficient to cover the entire brain slice) consisting of 10% Dextran Sulfate, 2× SSC, and the amplifier hairpins at a concentration of 120 nM. Amplification in 0.25-1 mm brain slices was performed for 5-6 hours (RT) and incubated on a rotating shaker (~70 rpm). Samples were then washed with 2x SSC for 2-3 times.

Imaging

The sample slices were stained with DAPI for 10 minutes, rinsed with 2x SSC for 2-3 times, and then placed in an enzymatic anti-bleaching buffer consisting of 10 mM Tris-HCl, 0.8% Glucose, and Pyranose oxidase (Sigma P4234) prepared either in 2x SSC for confocal microscopy or in RIMS (final refractive index of 1.47) for SPIM imaging.

Spinning-disk confocal microscopy

Brain slices were mounted and sandwiched between two No. 1 coverslips (EMS, Cat. # 63768-01). Images were collected on an Andor CSU-X spinning disk confocal on an Olympus IX81 microscope equipped with Andor iKon-M camera. The images were acquired by an oil immersion objective (Olympus UPLSAPO 60x, NA 1.35, w.d. 0.15).

SPIM

SPIM images were acquired using a custom built microscope for tissue clearing applications (Tomer et al. 2014, Keller et al. 2010, Treweek et al. 2015). Prior

to mounting, the brain slices in RIMS solution containing the anti-bleaching buffer were degassed by nitrogen gas for ~30 minutes. After degassing, the sample and the degassed solution were quickly transferred to in a Spectrosil Quartz cuvette (Fisher scientific, Cat. # NC9520499), and the brain slices were positioned on the cuvette wall by inserting another thin glass slide inside the cuvette. The cuvette was then attached to a translation stage that can lower the sample to the imaging chamber pre-filled with 100% glycerol or RIMS (refractive index of 1.47). The images were acquired by our custom built microscope, and the SPIM design and the optical components can be found in Treweek et al. (2015). Briefly, the channels of YFP, Cy3B, alexa 647 and IR800 fluorophores were excited using 473 nm, 561 nm, 640 nm, and 730 nm lasers, respectively. The emitted light was collected using a 10x CLARITY objective (NA 0.6, Olympus) in figure 4a and Movie S1, or a 25x CLARITY objective (NA 1.0, Olympus) in all other SPIM images. The images were digitized using a 4-megapixel cMOS camera (Andor Zyla 4.2), at a frame rate of 25 frames per second. To speed up the processing time, the pixels were binned (2x2) for images in figure 4a and Movie S1. When multiple tiles were acquired, TeraStitcher was used to stitch the images (Bria 2012).

Image Processing and Analysis

For SPIM dot analysis, the images were first registered using MATLAB command imregdemons, which calculates a 3D displacement field to maximally align individual channels. Once, the channels are aligned, the background is then subtracted as explained previously. A Laplacian of Gaussian (LoG) filter was then applied to the resulting images. Dots were then identified using MATLAB's local maxima finder (imregionalmax).

Supplementary References

Bria, A. and Lannello, G. TeraStitcher - A tool for fast automatic 3D-stitching of teravoxel-sized microscopy images. BMC bioinformatics 13, 316 (2012).

Choi, H.M.T., Beck, V.A. and Pierce, N.A. Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. ACS Nano 8(5), 4284-4294 (2014).

Keller, P.J., Schmidt, A.D., Santella, A., Khairy, K., Bao, Z., Wittbrodt, J., and Stelzer, E.H. Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. Nature Methods 7, 637-642 (2010).

Lubeck, E. and Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nature Methods 9, 743–748 (2012).

Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. and Cai, L. Single-cell in situ RNA profiling by sequential hybridization. Nature Methods 11, 360–361 (2014).

Raj A, van den Boogard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nature Methods 5, 877-879 (2008).

Suri C, Fung BP, Tischler AS, Chikaraishi DM. Catecholaminergic cell lines from the brain and adrenal glands of tyrosine hydroxylase-SV40 T antigen transgenic mice. J Neurosci. 13(3), 1280-91 (1993).

Tomer, R., Ye, L., Hsueh B., and Deisseroth K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. Nature Protocols 9, 1682-1697 (2014).

Treweek, J.B., Chan, K.Y., Flytzanis, N.C., Yang, B., Deverman, B.E., Greenbaum, A., Lignell, A., Xiao, C., Cai, L., Ladinsky, M.S., Bjorkman, P.J., Fowlkes, C.C. and Gradinaru, V. Whole-Body Tissue Stabilization and Selective Extractions via Tissue-Hydrogel Hybrids for High Resolution Intact Circuit Mapping and Phenotyping. Nature Protocols 10, 1860-1896 (2015).

Yang, B., Treweek, J.B., Kulkarni, R.P., Deverman, B.E., Chen, C.-K., Lubeck, E., Shah, S., Cai, L., and Gradinaru, V. Single-Cell Phenotyping within Transparent Intact Tissue through Whole-Body Clearing. Cell 158, 1-14 (2014).

Chapter 3

IN SITU TRANSCRIPTION PROFILING OF SINGLE CELLS REVEALS SPATIAL ORGANIZATION OF CELLS IN THE MOUSE HIPPOCAMPUS

[1] Sheel Shah et al. "NeuroResource In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus". In: *Neuron* 92.2 (2016), pp. 342–357. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2016.10.001. URL: http://dx.doi.org/10.1016/j.neuron.2016.10.001.

3.1 Summary

Identifying the spatial organization of tissues at cellular resolution from single cell gene expression profiles is essential to understanding biological systems. Using an in situ 3D multiplexed imaging method, seqFISH, we identify unique transcriptional states by quantifying and clustering up to 249 genes in 16,958 cells to examine whether the hippocampus is organized into transcriptionally distinct subregions. We identified distinct layers in the dentate gyrus corresponding to the granule cell layer and the subgranular zone and contrary to previous reports, discovered that distinct subregions within the CA1 and CA3 are composed of unique combinations of cells in different transcriptional states. In addition, we found that the dorsal CA1 is relatively homogenous at the single cell level, while ventral CA1 is highly heterogeneous. These structures and patterns are observed using different mice and different sets of genes. Together, these results demonstrate the power of seqFISH in transcriptional profiling of complex tissues.

3.2 Introduction

The mouse brain contains 10^8 cells arranged into distinct anatomical structures. While cells in these complex structures have been traditionally classified by morphology and electrophysiology, their characterization has been recently aided by gene expression studies. In particular, the Allen Brain Atlas (ABA) provides a systematic gene expression database using in situ hybridization (ISH) of the entire mouse brain one gene at a time (Dong et al., 2009; Fanselow and Dong, 2010; Thompson et al., 2008). This comprehensive reference provides regional gene expression information, but lacks the ability to correlate the expression of different

genes in the same cell. More recently, single cell RNA sequencing (RNA-seq) has identified many cell types based on gene expression profiles (Darmanis et al., 2015; Tasic et al., 2016; Zeisel et al., 2015). However, while single cell RNA-seq provides useful information on multiple genes in individual cells, it has relatively low detection efficiencies and requires cells to be removed from their native environment resulting in the loss of spatial information. These different approaches can lead to contradictory descriptions of cellular organization in the brain and other biological systems.

In the hippocampus, recent RNA-seq data suggests that the CA1 region is composed of cells with a continuum of expression states (Cembrowski et al., 2016, Zeisel et al 2015), while ABA analysis indicates that sub-regions within the CA1 have distinct expression profiles (Thompson et al, 2008). To resolve the two conflicting descriptions of hippocampal organization, a method to profile transcription in situ in the hippocampus with single cell resolution is needed. Here, we demonstrate a general technique that enables the mapping of cells and their transcription profiles with single molecule resolution in tissue, allowing an unprecedented resolution of cellular transcription states for molecular neuroscience (Fig 1A). A great deal of progress has been made recently in developing highly quantitative methods to profile the transcriptome of single cells. Building upon single molecule fluorescence in situ hybridization (smFISH) (Femino et al., 1998; Raj et al., 2006;) , Lubeck et al. devised a general method to highly multiplex single molecule in situ mRNA imaging irrespective of transcript density using super-resolution microscopy (Betzig et al., 2006; Rust et al., 2006; Lubeck and Cai, 2012;). However, the spectral barcoding methods used in these previous works is difficult to scale up beyond 20-30 genes because of the limited number of fluorophores (Fan et al., 2001; Lubeck and Cai, 2012).

To overcome the scalability problem, a temporal barcoding scheme was developed that uses a limited set of fluorophores and scales exponentially with time (Lubeck et al., 2014). Specifically, sequential probe hybridizations on the mRNAs in fixed cells impart a unique pre-defined temporal sequence of colors, generating an in situ mRNA barcodes. The multiplex capacity scales as

$$F^N$$

, where F is the number of fluorophores and N is the number of rounds of hybridization. Thus, one can increase the multiplex capacity by increasing the number of

rounds of hybridization with a limited pool of fluorophores. We called this approach Sequential barcoded Fluorescence in situ Hybridization (seqFISH) (Lubeck et al., 2014). In parallel, in situ sequencing methods were developed to directly sequence transcripts in tissue sections, but these methods suffer from low detection efficiency (<1%) (Ke et al., 2013; Lee et al., 2014). Recently, Chen et al. expanded the error correction method in the original seqFISH demonstration by using a Hamming distance 2 based error correcting barcode system, called merFISH. However, this implementation requires larger transcripts (>6kb) and many more rounds of hybridization than the method described here (Chen et al., 2015b). Furthermore, seqFISH and its variants have only been applied in cell culture systems due to the difficulty of smFISH detection in tissue. Here, we demonstrate an improved version of seqFISH in complex tissues by including signal amplification and a time-efficient error correction scheme (Fig 1A-D, Table S1), allowing us to resolve the structural organization of the hippocampus with single cell resolution.

3.3 Results

Signal amplification and error correction enable robust detection of mRNAs in tissues.

To overcome the autofluorescence and scattering inherent to brain tissues, we used an amplified version of smFISH, called single molecule Hybridization Chain Reaction (smHCR) (Fig 1E, S1A) (Choi et al., 2014, Shah et al., 2016). Single molecule HCR amplified signal 22.1 ± 11.5 (mean ± s.d., n=1338, Fig S1B) fold compared to smFISH, enabling robust and rapid detection of individual mRNA molecules in tissues and facile alignment of spots between hybridizations (Fig 2A). Single transcripts can be detected and localized in 3D with just 24 probes in tissues, enabling detection of transcripts <1kb in size, with a fidelity comparable to the smFISH gold standard (Fig S1C-D) but with signals 20-fold brighter (Shah et al., 2016). Single molecule HCR DNA polymers can also be digested by DNAse and re-hybridized in brain slices, allowing HCR-seqFISH to be robustly implemented (Fig 2A). We note the smHCR enables true 3D imaging in tissues, whereas the previous sequential FISH demonstrations (Lubeck et al., 2014, Chen et al., 2015) were performed only in flat cell cultures.

Furthermore, we improved upon our existing barcode system by implementing a time-efficient error correction scheme. The major source of error in seqFISH is the loss of signal due to mis-hybridization, which increases with the numbers of hybridization. We introduced an extra round of hybridization to correct loss

of signal during any round of hybridization (Fig 1D) (Supplementary Text). By minimizing the number of hybridizations, this error correction scheme is efficient to implement. For example, using 5 fluorophores and 4 rounds (instead of 3 rounds) of hybridization to code for 125 genes, we can still uniquely assign barcodes to genes even when signal from any single round of hybridization is missing. Although merFISH can tolerate 2 errors in the barcodes, it requires 16 rounds of hybridization to code 140 genes (Chen et al. 2015). As increasing the number of hybridizations can potentially lead to more experimental error and analysis complexity, our simple error correction method corrects for the most common error, dropped signal. Also, the fewer rounds of hybridizations decrease the total imaging and experimental time, which is rate-limiting for tissue experiments. HCR-seqFISH with simpler error-correction scheme allows efficient and accurate quantification of transcription profiles in tissues.

Using this HCR-seqFISH method, we surveyed the regional and sub-regional transcriptional heterogeneity within the temporal and parietal cortex and hippocampus of the mouse brain by imaging similar coronal sections collected from 3 different animals. Two similar sections from separate mice were profiled with probes for 125 genes, while one additional brain slice was imaged for 249 genes. In each of the coronal slices, between 60-80 fields of view were imaged, each 216 μm x 216μm x 15μm, in the cortex and hippocampus (Fig 1A and S1E). For the 125 gene set, 56 of the genes (Fig 1D, Table S1) were selected because they showed spatially heterogeneous expression based on the ABA (Lein et al., 2007), another 44 were selected from a list of transcription factors, and 25 marker genes were selected from single cell RNA-seq datasets (Zeisel et al., 2015). One hundred of these genes were barcoded by 4 rounds of hybridization (Fig 1B). The remaining 25 high abundance genes were measured individually using 5-color smHCR in 5 serial rounds of hybridizations (Fig 1C). This hybrid approach of measuring medium expression genes with barcoding seqFISH and high copy number genes serially in subsequent hybridizations allows a large dynamic range of transcripts to be profiled in the same cell.

seqFISH is an accurate and efficient method to multiplex RNA in situ.

To determine the accuracy of the seqFISH method in quantifying mRNA levels in single cells in tissue, we compared the copy number of 5 of the 100 target genes measured by barcoding to the copy number found by smHCR detection in the same cell (Fig 2B, S2A) in 15 μm brain sections. We found that the copy number of the

RNAs per cell as measured by barcoding and smHCR agreed with an R-value of 0.85 and a slope of 0.84 (N=3851). As smHCR matches smFISH transcript quantitation (Shah et al., 2016), the barcoded seqFISH method can quantify mRNA molecules in single cells with 84% efficiency compared to the gold standard of smFISH. In comparison, single cell RNA-seq measurements are 5-20% efficient based on spike-in controls and in situ sequencing is less than 1% efficient (Darmanis et al., 2015; Klein et al., 2015; Lee et al., 2014; Macosko et al., 2015; Tasic et al., 2016; Zeisel et al., 2015; Ståhl et al., 2016). This high efficiency of detection results from a low transcript drop rate and a high barcode recovery rate due to the error correction round of hybridization. In our experiment, 78.9% of barcodes (N=2,115,477 barcodes) were found in all 4 hybridization rounds and 21.1% were identified in 3 out of the 4 hybridizations (Fig 2C), indicating that the probability of detecting a given mRNA molecule is 94% in each round of hybridization (Fig S2B).

To quantify the amount of false positive signal due to misalignment of barcodes and nonspecific binding of probes, we measured the amount of off-target barcodes that were detected. With four rounds of hybridizations and 5 fluorophores, there were 54=625 unique codes. We assigned 100 of these barcodes to measure mRNAs detected at 914.8 ± 570.5 counts per cell (mean ± s.d., N=3439). In comparison, the 525 remaining off-target barcodes that were not used were detected at 4.6 ± 4.7 (mean ± s.d., N =3439) counts per cell (Fig 2D). False positives, due to chance alignment of nonspecifically bound spots, contributed minimally to the barcode readouts because of this three order of magnitude difference in detected barcodes (on target vs. off target). The false positives we observe fall only on barcodes hamming distance one away from on-target barcodes, yet minimally contribute to undercounting on-target barcodes (Fig 2E). Furthermore, even the most frequent off-target barcode was observed 65.57 times less frequently than the most infrequent mRNA coding barcode (Fig 2E, S2). Even though during each round of hybridization, 24.8 ± 0.4% (mean ± s.e., N=4 rounds of hybridization) of the spots were nonspecifically bound probes, barcode miss-assignments did not occur frequently because non-specifically bound probes do not reappear in the same location after digestion with DNAse and re-hybridization (Fig 2A). Together the quantifications of false positive and false negative barcodes demonstrate that this method is highly efficient and accurate at detecting RNAs in situ in single cells within tissues.

Cell clusters are based on combinatorial expression profiles.

We imaged the expression of 125 genes in coronal sections from two mice for a total of 14,908 cells (Fig S1E). Cortical and hippocampal cells were segmented based on DAPI and Nissl staining. A tessellation algorithm was developed to accurately segment densely packed cells in the hippocampus. To avoid capturing mRNA from neighboring cells, we contracted by 10% the borders of cells determined by the segmentation algorithm.

To group the single cell data into distinct transcriptional states, we Z-score normalized the copy number of each transcript in every cell (Fig 3A) and hierarchically clustered the cells to identify cells with similar expression patterns (Table S2, Fig S3). While these clusters do not necessarily represent canonical cell types, many of these clusters contain clear transcriptional markers of known cell types previously identified by single cell RNA-seq (Fig 3B) (Zeisel et al., 2015, Tasic et al 2016). Cell clusters 12 and 13 contained clear expression of Gja1 which marks out astrocytes (Zeisel et al., 2015, Tasic et al 2016). Cluster 12 also expresses Mfge8 while cluster 13 did not, indicating two distinct subpopulations of astrocytes (Fig 3B). There are further subclusters within each of the astrocyte populations with different spatial localization patterns (Fig S3C-E). Cluster 11 cells expressed Laptm5, a known microglia marker (Zeisel et al., 2015, Tasic et al 2016). Cluster 3 expressed interneuron genes while cluster 1-2 and 4-5 expressed genes associated with pyramidal neurons (Zeisel et al., 2015, Tasic et al 2016). The major clusters were robust to down-sampling the number of cells used in clustering (Fig S4), with some of the hippocampal pyramidal and glial clusters robustly defined even with 400 cells. Similarly, principal component analysis (PCA) visualization of the data (Fig S3H) recapitulated the major clusters that correspond to astrocyte, microglia, cortical pyramidal, hippocampal pyramidal, dentate gyrus (DG) granule, and interneuron cells.

As the cluster distance between different cells is proportional to the number of differentially expressed genes in the target list, an unbiased clustering of the 125 gene data without weighting specific genes should not be interpreted directly as canonical “cell types,”, but rather as grouping cells with different patterns of genes expression based on the current target list. We will refer to some of these clusters as pyramidal neurons or astrocytes for ease of notation, but strictly speaking, they are cells clusters with similar expression patterns as neurons or astrocytes.

Cell clusters show distinct regional localization

Many neuronal clusters mapped to distinct regions in the brain (Fig 3B). Several classes of pyramidal cells (cluster 1-2) showed exclusive localization to the hippocampus, while other classes (4-5) showed predominantly cortical localization. There were also a class of cells (cluster 7) that were almost exclusively present in the DG. Interestingly, these clusters segregated based solely on gene expression profiles without adding any spatial information into the clustering algorithm. These differences in transcriptional states of neurons could be due to intrinsic differences in the cells or due to different local environment and activity patterns.

In contrast, astrocyte, microglia and other non-neuronal cell clusters were generally uniformly present in all areas of the brain (Fig 3B). However, subclusters of astrocytes did localize to different regions of the brain preferentially (Fig S3E), with subcluster 12.3 localized preferentially to the cortex, while 12.1 subcluster was uniformly distributed. Similarly, cluster 9 cells contain subclusters (9.3, 9.5 and 9.6) that localize exclusively to the DG, while other subcluster (9.1) localize almost exclusively to the cortex. The regional localization of neurons are especially pronounced with cluster 1 and 2 localized almost exclusively to the hippocampus, with some of the subclusters localized predominantly to the CA3. Furthermore, while pyramidal cell clusters 4 and 5 are preferentially cortically localized, the few hippocampal cells in these clusters form their own subclusters (4.4 and 5.4) (Fig S3E). In cluster 6 cells, many subclusters with distinct expression profiles are localized almost exclusively in the CA1, CA3 or the DG (Fig S3C). In contrast, cluster 7 cells show a relatively homogenous regionalization pattern, but further subdivide based on combinatorial expression patterns (Fig S3D). Subclusters of cluster 9 also show significant regionalization where subclusters 9.1, 9.3, 9.5, and 9.6 show localization to the SGZ (Fig S3E). Overall, cell clusters with similar expression profiles exhibited similar spatial localizations across the brain with a correlation coefficient of 0.67 (Fig S3G), indicating the existence of archetypal regional expression patterns and potential spatial markers in the brain. These results show that the tissue-optimized HCR seqFISH approach can directly identify a variety of transcriptional states and quantify broad spatial patterns of expression. Combinatorial expression patterns define fine clusters.

While certain cell clusters contain strong expression of marker genes, not all clusters are defined based on a few genes. How much power do individual genes or groups of genes have in explaining the observed cell clusters? To understand this, we

examined whether subsets of genes can recapitulate the observed clusters (Fig 3C-D). We found that any set of 25 genes recovers about half of the correlation structure in the cell-to-cell correlation map (Fig 3C, S3I, N=10 bootstrap replicates). The fact that the selection of any 25 genes can explain the gross patterns in the data is likely due to the high correlations amongst the expression patterns of genes, as shown in the gene-to-gene correlation map (Fig S3J). Thus, a small subset of the measured genes can provide sufficient information to infer the gross transcriptional states of the cells. Interestingly, this may be the same reason why low-coverage single cell sequencing methods such as drop-seq and inDrop (Klein et al., 2015; Macosko et al., 2015) can capture the large distinction of cell types, because many highly expressed genes are correlated to other genes that collectively define cell types. At the same time, the finer correlation structure in the data, required to define the cell clusters accurately, can only be captured with accurate quantitation of many genes (Fig 3C-D). Consistent with this, using a "random-forest" machine learning algorithm (Breiman, 2001) to classify cell clusters, we found that 75 genes are needed to classify cells with 50% accuracy, indicating that correct cluster assignment requires more detailed information from many genes (Fig 3C). Supporting this view, the first 10 principal components (PC) explained 59.5% of the variation in the data, while the rest of the variation required the remaining 115 PCs (Fig 3D, S3F). The "random forest" algorithm required 10 PCs to predict the cell cluster assignments with 50% accuracy (Fig 3D), but accuracy steadily increased with more PCs. These observations indicated two levels of information in the data: a coarse level, where large distinctions in cell clusters are observable by a few genes, and a fine level, where subtle distinctions require many more genes.

These results suggest two points experimentally. First, multiplexing at the level of 20 genes by seqFISH can give broad cell cluster identification that is not available with 2-3 gene smFISH experiments. Although single marker genes are useful for inference, we find that they frequently are not sufficient for cell classification. For example, all DG specific granule cells (clusters 7) have Gpc4 and Vps13c as their enriched marker genes (Fig 3B); yet, Gpc4 and Vps13c are also strongly expressed in other hippocampal cells outside of the DG, as seen in both our experiments and the ABA. Thus, smFISH against Gpc4 and Vps13c alone would not be sufficient to uniquely identify the DG granule cells. Furthermore, even the strongly bimodal markers that are known to define cell types (i.e. Mgfe8, Gja1, etc.) are correlated enough to overall expression profiles that cells fall into the appropriate cluster even when these genes are excluded. This point suggests that while marker genes can be

essential in assigning a cell to a known cell type, they are not necessary to identify unique clusters in the dataset provided enough measurements are made.

Second, accurate measurement of combinatorial expression of many genes enabled by seqFISH can allow for more specific cell cluster identification. As a comparison, in single cell RNAseq data, CA1 pyramidal cells are clustered into a single cluster (Zeisel et. al, 2015; Habib et. al 2016) potentially because of the relatively lower detection efficiency. In our seqFISH experiments, measuring hundreds of genes quantitatively, we can resolve several clusters and subclusters with robust regionalization within the CA1 (Fig 3B, S3C-E).

Cells are patterned in the dentate gyrus.

To further visualize the spatial organization of cells, we mapped cluster definitions of cells back into the images. In the DG, we observed a striking lamina layering of cell classes. The two blades of the DG (Fig 4A-B) showed mirror arrangements of cells, with cluster 9 cells, forming the subgranular zone (SGZ), leading into a granule cell layer (GCL) dominated by a single cluster of granule cells (cluster 7) (Fig 3B). In the 125 gene data set, the cells of the GCL were found to be dominated by expression of Gpc4 and Vps13c matching ISH data from the ABA (Fig S8B). Cluster 7 was found to be further subdivided into 6 subclusters (Fig S3D). These subclusters were found to have varying levels of calbindin D-28K (Calb1) expression which is known to increase with granule cell maturation (Yang et al., 2015). On the other hand, the cells of the SGZ were found to be significantly enriched in astrocyte markers such as Mfge8 and Mertk, which has been also been observed previously (Miller et al, 2013) and in the ABA data (Fig S8A). However, these cells do not cluster with typical astrocytes (cluster 12 and 13) because their combinatorial expression patterns are different from astrocytes, consistent with their classification as a completely different population of cells.

In the fork region of the DG, the layer of cluster 9 cells appeared on the interior surface of the fork, followed by a layer of granule cells (cluster 7) (Fig 4C). A different layering pattern is seen at the crest of the DG, where astrocytes, microglia, and some other glial cells line the exterior of the crest ensheathing the GCL (Fig 4D). In both brains of the 125 gene experiments, the same cell clusters and spatial arrangements are observed. Furthermore, because the mRNAs are imaged in 3D in the 10-15um brain slices, we can obtain a 3D view of the expression profiles, shown in the fork regions of the DG (Fig 4F).

Distinct regions of CA1 and CA3 are composed of different combination of cell clusters.

While each region of the DG contains similar compositions of cells, distinct subregions within the CA1 and CA3 contained different combinations of cell classes (Fig 5, S6F). In the CA1, there were 3 distinct regions defined by their individual cellular compositions. In the dorsal region of CA1 (CA1d), neuron cluster 6 (enriched in Nell1, a protein kinase C binding protein) (Table S2) was the major cell type in the pyramidal layer, with astrocyte, microglia and other cells (clusters 10-13) intercalating into the stratum pyramidale (SP) (Fig 5A-C). Transitioning into the CA1 intermediate region (CA1i) (Fig 5D), pyramidal cell cluster 4 displaced cell cluster 6 as the dominant cell, with the co-appearance of cluster 1 and 2 pyramidal cells.

As the middle of the CA1i region was reached, a small amount of cluster 4 pyramidal cells remain, while cluster 1 and 2 pyramidal cells dominate (Fig 5E-F). Cluster 1 and 2 are enriched in Nell1 (EGF like protein), Npy2r (neuropeptide Y receptor), Slc4a8 (sodium bicarbonate transporter) and B3gat2 (glucuronosyltransferase) (Table S2). The CA1i region displayed a characteristic spatial organization where glial cells line the outermost regions, while pyramidal cell cluster 1 and 2 longitudinally partitioned the pyramidal layer. This separation of the inner versus the outer layers of CA1 matches those observed in previously (Dong et al., 2008). Furthermore, interneurons (cluster 3) were found to preferentially line the inner edge of the pyramidal layer in the CA1i region (Fig 5E-F). This patterning of interneurons, particularly subcluster 3.1 cells which were enriched in Slc5a7, a choline transporter, was consistent with the patterning of cholinergic interneurons observed with ChAT-GFP labeling (Yi et al., 2015). Finally, the largest amount of heterogeneity in the CA1 was seen in the ventral CA1 region (CA1v), where cell clusters 3, 5, and 10 began to mix in with clusters 1 and 2 (Fig 5G-I).

Similarly, the CA3 was found to have four transcriptionally distinct regions with different pyramidal cell compositions and abrupt transitions. The ventral most region of CA3 contained a high level of heterogeneity of pyramidal cell clusters (Fig 5J-K), while the intermediate region of CA3 contain a mixture of cell clusters 1 and 2 (Fig 5L-M). As the CA3 progressed towards the hilus of the DG, the cell types transitioned first to primarily cluster 4 neurons (enriched in dcx, doublecortin, and Col5a1, a collagen), and then to almost exclusively cluster 6 neurons in the region most proximal to the DG hilus (Fig 5O-P). It is interesting to note that while cluster

6 cells appear in both the CA1 (subcluster 6.8) and CA3 (subclusters 6.1 and 6.4), sub-clusters of 6 show distant regional localization (Fig S3E), suggesting that the gene expression differences in CA1 and CA3 cells are captured in the seqFISH data. We note that similar patterns of homogeneous dorsal and heterogeneous ventral cell populations are observed when only hippocampal cells are clustered (Figure S5).

The regionalized expression patterns we observed in the hippocampus match closely to those observed in previous literature (Thompson et al Neuron 2008 and Dong et al PNAS 2009). For example, CA1d, CA1i, CA1v boundaries correspond to the boundaries shown in Dong et al Fig 2B. In CA3, the subregions observed in our experiment match the CA3 subregion 4-7 in Thompson et al. (Thompson et al., 2008).

Lastly, we note that the two slices from two different mice in the 125 gene experiment show not only the same subregional structure (Fig 4-6), but also the same clusters of cells (Fig 5 and 6) in the different subregions of the hippocampus (Fig S6F). In both brains, the CA1d consists of relatively homogenous population of cluster 6 cells, which transition to a mixture of 1 and 2 cells in CA1i, and finally to a mixture of 1-6 and 10 cells in the CA1v (Fig S6F). These results together show that the sub-regions of the hippocampus are a robust feature in the organization of CA1 and CA3, consisting of cells classes with distinct expression profiles. The stereotypical nature of the spatial arrangement of these structures suggest further experiments with seqFISH and other functional assays to probe the distinct functions of the different cell clusters in the CA1 and CA3.

249 gene multiplex experiments show the same hippocampal subregions

To further show that the sub-regional structure of the hippocampus is independent of the target genes, we performed a 249 gene seqFISH experiment on a third coronal section. Of these 249 genes, only 22 genes overlapped with the 125 gene experiment set (Table S8). For this set of genes, 214 were selected from a list of transcription factors and signaling pathway components and the remaining 35 were selected from cell identity markers from another single cell RNAseq dataset (Tasic et al, 2016). The 214 genes were barcoded by 5 rounds of hybridization, while the remaining genes were imaged in 7 rounds of non-barcoding serial hybridization. To quantify the efficiency of this experiment, 4 genes in the barcoding set (Smarca4, Sin3a, Npas3, and Neurod4) were re-probed with smHCR. The barcoding efficiency of the 249 gene probe set was found to be 71% with and R value of 0.80 (Fig S6D).

In single cells, we detect on average 2807±1660 (mean±s.d., N=2050 cells) total barcoded barcodes.

The same arrangement in the DG was observed in the 249 gene experiment, despite different genes used, indicating robust identification of the layering in the DG by seqFISH (Fig 7S-T). In particular, the cells in the SGZ are clustered independently from cells in the GCL, similar to the layers observed in the 125 gene experiment. In the SGZ cells, we observed enrichment of Sox11, a key transcription factor in neurogenesis (Miller et al, 2013). Other transcription factors involved in neurogenesis, NFIA and Tbr1 are also enriched in the SGZ cells as seen in our data and the ABA images (Fig S8A, Table S7). The observations of this distinct layer in both the 249 and 125 gene experiment and the combined gene enrichment pattern (increased Sox11, Sox9, NFIA, and Tbr1 in the 249 gene experiment and increased Mertk and Mfge8 in the 125 gene experiment) suggests that many cells in this layer are involved in adult neurogenesis in the SGZ. Supplementary figure 7B shows distinctive marker gene expression in the GCL of the dentate gyrus.

In addition, the same regionalized cellular patterns are observed in CA1d, CA1i, and CA1v, where different subregions utilize different cell classes in characteristic ratios (Fig S6G). As seen with the 125 gene experiment, while the CA1d uses only a few cell classes and is relatively homogeneous, while the CA1v region is made up of many different cell classes resulting in a high level of cellular heterogeneity. Furthermore, the distinction between CA1 and CA3 cell clusters are more clear in the 249 gene experiment suggesting more resolving power of spatial patterns (Fig 7A-K). The 249 gene experiment also suggests that the CA3 may be composed of 3-4 subregions based on cell cluster composition (Fig 7L-R). The cellular heterogeneity of the CA3 is again shown to mirror that of the CA1, where the cellular heterogeneity increases along the dorsal to ventral axis. Cells with distinctive marker gene expression in the hippocampus and dentate gyrus are shown in supplementary figure 7.

3.4 Discussion

Single cell data resolves cellular organizations in the sub-regions of the CA1 and CA3.

Two conflicting views of the cell types in the hippocampus have been proposed based on the analysis of the Allen Brain Atlas data (Thompson 2008) as well as recent RNA-seq data (Cembrowski et al., 2016, Zeisel et al 2015). Analysis of the ABA

in situ data showed that distinct subregions of the hippocampus expressed different molecular markers, indicating that the CA1 and CA3 are "regionalized" into distinct sub-structures (Fanselow and Dong, 2010; Thompson et al., 2008). However, recent bulk RNA-seq experiments on the CA1 found that gene expression patterns changed gradually along the dorsal to ventral axis, contradicting the sharp boundaries observed in the ABA analysis (Cembrowski et al., 2016). Further supporting this "continuous" cell type view of the hippocampus, analysis of the single cell RNA-seq data (Zeisel et al, 2015) identified a single continuous population of cells in the CA1 region. Our data provides a single cell resolution picture of the spatial organization of cells in the hippocampus and reconciles both the RNA-seq and the ABA data. While our data mostly supports a regionalized view of the hippocampus, we observe that a single cell class does not in general define CA1 and CA3 sub-regions. Instead, we observed that different subregions of CA1 and CA3 are composed of distinct combinations of cell clusters (Fig 5-7). For example, CA1d consists primarily of cluster 6 pyramidal cells (Fig 5A-C), in addition to the cluster 1,2, 10, and 12 cells, while CA1v consists of a large set of cell classes including cluster 1-6 and 10 cells, but at different relative abundances (Fig 5-6, Fig S6 F-G). Due to this intermixing of cell classes in each sub-region, a bulk measurement of transcription profiles would find a lack of regionalization, but single cell analysis with spatial resolution would identify these distinct regions based on their unique cell class compositions. Indeed, when we averaged the single cell expression profile within each sub-region of the CA1, we can reproduce the continuous correlation profiles found by bulk RNA-seq between CA1v, CA1i, and CA1d (Fig 8) (Cembrowski et al., 2016). The bulk RNA-seq observation that CA1i lacked specific marker genes can also be explained. This is in fact consistent with our findings that CA1i contained cell classes present in both CA1d and CA1v (Fig 5-7). This organization of cell classes is observed in both the 125 gene experiments as well as in the 249 gene experiment.

It is worth noting that the complexity of cell populations observed in the CA1d versus the CA1v matches the functional differences in CA1. CA1d is responsible for spatial learning and navigation and contains a higher concentration of place cells and send projections to dorsal subiculum and cortical retrosplenial area (Cenquizca and Swanson, 2007; Jung et al., 1994; Risold et al, 1997; O'Keefe and Dostrovsky, 1971). We observed that CA1d is composed of a relatively homogeneous population of cells, predominantly of cluster 6 cells. In contrast, the ventral region is involved in a variety of cognitive tasks, such as stress response, emotional and social behavior (Cenquizca and Swanson, 2007; Jung et al., 1994; Fanselow and Dong, 2010; Kishi

et al., 2006; Muller et al., 1996; Petrovich et al., 2001; Pitkänen et al., 2000; Saunders et al., 1988; Witter and Amaral, 1991; Yi et al., 2015). Correspondingly, we observed a large set of cell classes in the CA1v regions. It is intriguing to hypothesize that the different cell classes identified based on molecular profiles may correspond to neurons with distinct connectivity and functional patterns. This hypothesis can be investigated in future experiments combining anterograde tracing as well as electrophysiological recording followed by seqFISH.

SeqFISH cell classes versus single cell RNA-seq cell types

While the accurate measurement of 100-200 genes can provide distinctions between the large functional classes found by RNA-seq, the clusters found by seqFISH, in general, should not be interpreted as cell types. RNA-seq measurements at the whole transcriptome level defines cell types based on highly variable genes. On the other hand, seqFISH provides highly accurate measurements of fewer genes, but uses the combinatorial expression patterns to group cells into clusters. However, because only 100-200 genes are targeted in the seqFISH experiments, not all of the "cell types" are equally represented in the gene list and seqFISH cannot catalogue "cell types" in the same fashion that single cell RNAseq can. For example, in our 125 gene experiments, we cannot resolve the distinct subpopulation of interneurons because we lacked marker genes such as Vip and Sst. seqFISH and RNA-seq provide two different, yet complementary, levels of resolution into the transcriptional profiles of cells. RNA-seq measures the transcription levels of thousands of genes but at a lower quantitative accuracy, while seqFISH measures only 100's of genes but with much greater quantitative power. The differing nature of the two sets of data informs how the data should be analyzed and interpreted. Thus, seqFISH and single cell RNAseq have complementary roles in elucidating distinct cell subpopulations in tissues. SeqFISH could be applied to find finer distinctions within cell types found by RNA-seq or to look at the spatial patterning of cell types found by RNA-seq.

seqFISH provides a generalized method to multiplex mRNA imaging in tissues

seqFISH with amplification and error correction provides a highly quantitative method to profile hundreds of mRNA species directly in single cells within their native anatomical context. Our method of stripping the probes from the RNA has many advantages. DNAse digestion of probes allows false positives to be rejected as nonspecifically bound probes do not colocalize between different

rounds of hybridization (Fig 2A). In addition, the same region of the transcript can be hybridized in every round, allowing seqFISH to efficiently target mRNAs shorter than 1kb, enabling targeting of most genes. Lastly, seqFISH allows exponential scaling of barcode numbers, thus 4-5 rounds of hybridization can code for hundreds of transcripts with a simple error correction scheme. Theoretically, the entire transcriptome can be coded for with error correction by using 8-9 rounds of hybridization with seqFISH. These advantages of HCR seqFISH allows robust multiplexed RNA detection in tissues, shown here in the mouse brain. Ultimately, the multiplexing capability of seqFISH is limited by the amount of optical space within a cell, and not by the coding capacity of the method (supplementary text). We showed previously that super-resolution microscopy can significantly increase the optical space available in the cell for transcription profile imaging, but super-resolution microscopy experiments proved difficult to image in samples thicker than 1μm, and were experimentally cumbersome and time consuming to image (Lubeck and Cai, 2012). A recent development in expansion microscopy as well as correlation methods (Coskun et al., 2016) however offers promise for multiplexing to levels of high transcript density (Chen et al., 2015a; Treweek et al., 2015, Chen et al., 2016). In addition, by labeling subcellular components (i.e. dendrites and axons) with antibodies, the local transcriptome in compartments of the cell can be measured.

We observed that, because expression patterns amongst genes are highly correlated, the distinction between large classes of cells can be determined from 10-20 genes, while a finer classification of cell clusters depends on the quantitative measurement of the combinatorial expression patterns of many genes (Fig 3C-D). This correlation amongst genes can be used to "stitch" our seqFISH data with single cell RNAseq data, similar to the approach explored with single cell RNAseq and ISH in Satija et al (Satija et al., 2015). By correlating seqFISH data to single cell RNA-seq expression data, cells types identified based on RNA-seq can be "mapped" back into our seqFISH data.

As shown here, seqFISH with hundreds of genes in tissues can become a general and widely used tool to answer a wide range of fundamental questions in biology and medicine. For neuroscience, by combining the insights into the spatial organization of transcription provided by seqFISH with connectomics and electrophysiological measurements, we can obtain a comprehensive understanding of the molecular basis of the neuroanatomy of the brain.

3.5 Experimental Procedure

Probe Design.

Genes were selected from the Allen Brain Atlas database. We identified genes that are heterogeneously expressed in coronal sections containing the hippocampus at Bregma coordinates -2.68 mm anterior. We selected 100 genes that had high variances across these distinct regions and that also had low-medium expression levels. Probe sequences were designed using software developed in house. Full details are described in Supplemental Experimental Procedures.

Probe Generation.

All oligoarray pools were purchased as 92k synthesis from Customarray Inc. Probes were amplified from array-synthesized oligo pool as previously described (Chen et al., 2015b). Full details are described in Supplemental Experimental Procedures.

Brain extraction and sample mounting.

C57BL/6 with Ai6 Cre-reporter (uncrossed) (Jackson Labs, SN: 007906) female mice aged 50-80 days were anesthetized with isoflurane according to institute protocols (protocol #1701-14) (Madisen et al., 2012). Mice were perfused with 4% PFA and the brain was dissected out and placed in a 4% PFA buffer for 2 hours at room temperature. The brain was then immersed in 4C 30% RNAse-free Sucrose\1x PBS until the brain sank. Once sunk, the brain was embedded in OCT and sectioned. Full details are described in Supplemental Experimental Procedures.

Sample permeabilization, hybridization, and Imaging.

Sections were permeabilized in 4C 70% EtOH for 12-18 hours. Brains were further permeabilized by the addition of rnase-free 8% SDS. A hybridization chamber was adhered around the brain section. RNA integrity test probes were hybridized overnight at 37 in hybridization buffer (Table S3). Samples were washed in 30% wash buffer (WB) for 30 minutes. Probes were amplified. Following amplification, samples were washed in the same 30% WB for at least 10 minutes to remove excess hairpins. Samples were stained with DAPI and submerged in pyranose oxidase antibleaching buffer (Lubeck et al., 2014). If the RNA was deemed to be intact, DAPI data was collected in this hybridization. Samples were digested with DNAse I for 4 hours at room temperature on the scope. Following DNAse I the sample was washed several times with 30% WB and the probes were hybridized overnight

(Table S4). Samples were again washed and amplified. Repeating this cycle with the appropriate probes for each hybridization developed barcode digits. Fluorescent Nissl stain was collected at the end of the experiment along with images of multi-spectral beads to aid chromatic aberration corrections. Full details are described in Supplemental Experimental Procedures.

Image Processing.

The images were first corrected for to remove the uneven illumination profiles in each channel and to remove the effects of chromatic aberration. The background intensity in the images was then subtracted. A 150-pixel border region around the image was ignored in all analysis to avoid errors from edge effects of illumination. Full details are described in Supplemental Experimental Procedures.

Image Registration.

The processed images were then registered by first taking a maximum intensity projection along the z direction in each channel. All of the maximum projections of the channels of a single hybridization were then collapsed resulting in 4 composite images containing all the points in a particular round of hybridization. Each of these composite images of hybridization 1-3 was then cross-correlated individually with the composite image of hybridization 4 and the position of the maxima of the cross-correlation was used as the translation factor to align hybridizations 1-3 to hybridization 4.

Cell Segmentation.

For cells in the cortex, the cells were segmented manually using the DAPI images taken in the first round of hybridization and the fluorescent nissl stain taken at the end of the experiment. Furthermore, the density of the point cloud surrounding a cell was taken into account when forming cell boundaries, especially in cells that did not stain with the nissl stain. For the hippocampus, the cells were segmented by first manually selecting the centroid in 3D of each DAPI signal of every cell. Transcripts were first assigned based on nearest centroids. These point clouds were then used to refine the centroid estimate and create a 3D voronoi tessellation with a 10% boundary-shrinking factor to eliminate ambiguous mRNA assignments from neighboring cells. Regional segmentation was performed manually using the ImageJ ROI tool.

Barcode calling.

The potential mRNA signals were then found by LOG filtering the registered images and finding points of local maxima above a specified threshold value. Once all potential points in all channels of all hybridizations were obtained, dots were matched to potential barcode partners in all other channels of all other hybridizations using a 1-pixel search radius to find symmetric nearest neighbors. This procedure was repeated using each hybridization as a seed for barcode finding and only barcodes that were called similarly in at least 3 out of 4 rounds were used in the analysis. The number of each barcode was then counted in each of the assigned cell volumes and transcript numbers were assigned based on the number of on-target barcodes present in the cell volume. All image processing and image analysis code can be obtained upon request. Full details are described in Supplemental Experimental Procedures.

Clustering.

To cluster the dataset with two brain measured with 125 genes, we first Z-score normalized each of the slices based on gene expression (Table S5 and Table S6). Once the single cell gene expression data is converted into z-scores, we compute a matrix of cell-to-cell correlations using Pearson correlation coefficients for all of the cells in the two brains. Then hierarchical clustering with Ward linkage is performed on the cell-to-cell correlation data using cells taken from the center of the field of view. To analyze the robustness of individual clusters, a random forest model was trained using varying subsets of the data and used to predict the cluster assignment of the remaining cells (Breiman, 2001). For Figure 4-6, the entire field of cells was classified using the clustered cells as the training set. A bootstrap analysis by dropping different sets of cells was performed in increments (Fig S5). To determine the effect of dropping out genes on the accuracy of the clustering analysis, we used a random forest decision tree to learn the cluster definition based on the 125 gene data. Then we ask the decision tree to re-compute the cluster assignment on cell-to-cell correlation matrices with fewer and fewer genes (Fig 3C-D, green line). Bootstrap resampling was also performed with this analysis (Fig 3C-D, blue lines). The cell-to-cell correlation in Fig S3EI was calculated with increasing number of principal components dropped (have their eigenvalues set to zero). The cluster assignment accuracy is again computed through the random forest decision tree. The 249 gene experiment was clustered independently with Z-score normalized data.

3.6 Figures

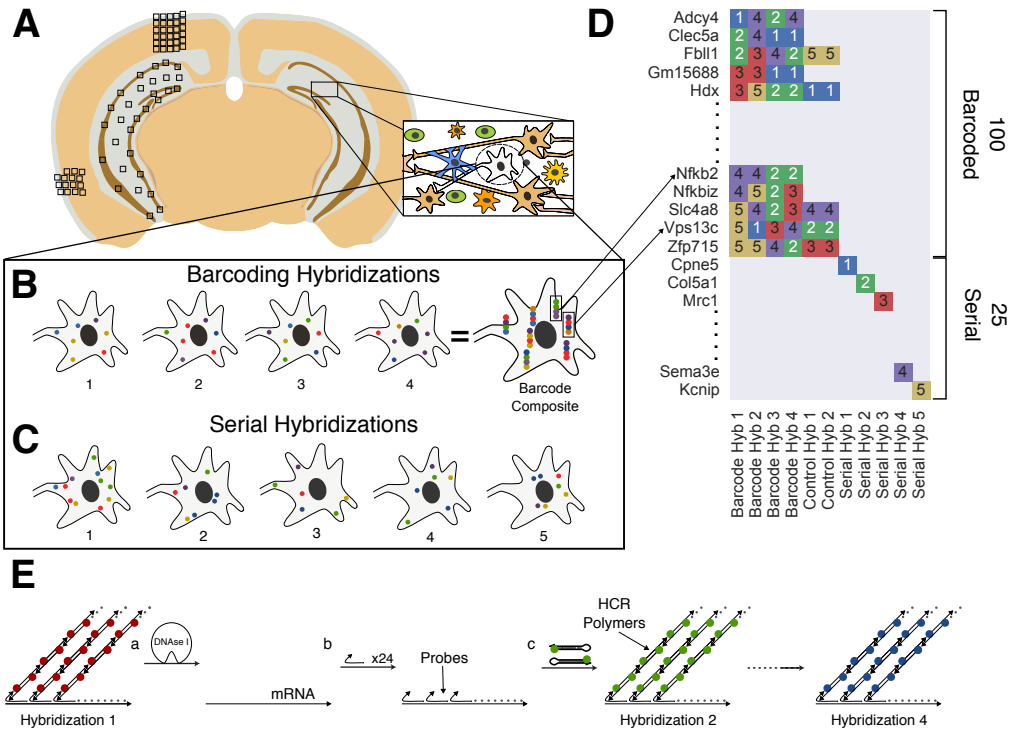

A
B
Barcoding Hybridizations
1
2
3
4
Barcode Composite
C
Serial Hybridizations
1
2
3
4
5
D
Adcy4
Clec5a
Fbll1
Gm15688
Hdx
Nfkb2
Nfkbiz
Slc4a8
Vps13c
Zfp715
Cpne5
Col5a1
Mrc1
Sema3e
Kcnip
100 Barcoded
25 Serial
Barcode Hyb 1
Barcode Hyb 2
Barcode Hyb 3
Barcode Hyb 4
Control Hyb 1
Control Hyb 2
Serial Hyb 1
Serial Hyb 2
Serial Hyb 3
Serial Hyb 4
Serial Hyb 5
E
a
DNAse I
mRNA
b
x24
Probes
c
HCR Polymers
Hybridization 1
Hybridization 2
Hybridization 4

Figure 0 *(previous page)*: Overview of the Sequential barcode FISH (seqFISH) in brain slices. **A.** A coronal section from a mouse brain was mounted on a slide and imaged in all boxed areas. Each image was taken at 60x magnification. **B.** Example of barcoding hybridizations from one cell in field from A. The same points are re-probed through a sequence of 4 hybridizations (numbered). The sequence of colors at a given location provides a barcode readout for that mRNA ("barcode composite"). These barcodes are identified through referencing a lookup table abbreviated in D and quantified to obtain single cell expression. In principle, the maximum number of transcripts that can be identified with this approach scales to F^N, where F is the number of fluorophores and N is the number of hybridizations. Error correction adds another round of hybridization. **C.** Serial smHCR is an alternative detection method where 5 genes are quantified in each hybridization and repeated N times. Serial hybridization scales as F*N. **D.** Schematic for multiplexing 125 genes in single cells. 100 genes are multiplexed in 4 hybridizations by seqFISH barcoding. This barcode scheme is tolerant to loss of any round of hybridization in the experiment. 25 genes are serially hybridized 5 genes at a time by 5 rounds of hybridization. Each number represents a color channel in single molecule HCR. As a control, 5 genes are measured both by double rounds of smHCR as well as barcoding in the same cell. **E.** SmHCR amplifies signal from individual mRNAs. After imaging, DNAse strips the smHCR probes from the mRNA, enabling rehybridization on the same mRNA (step a). The "color" of an mRNA can be modulated by hybridizing probes that trigger HCR polymers labeled with different dyes (step b). mRNA are amplified following hybridization by adding the complementary hairpin pair (step c). The DNAse smHCR cycle is repeated on the same mRNAs to construct a predefined barcode over time.

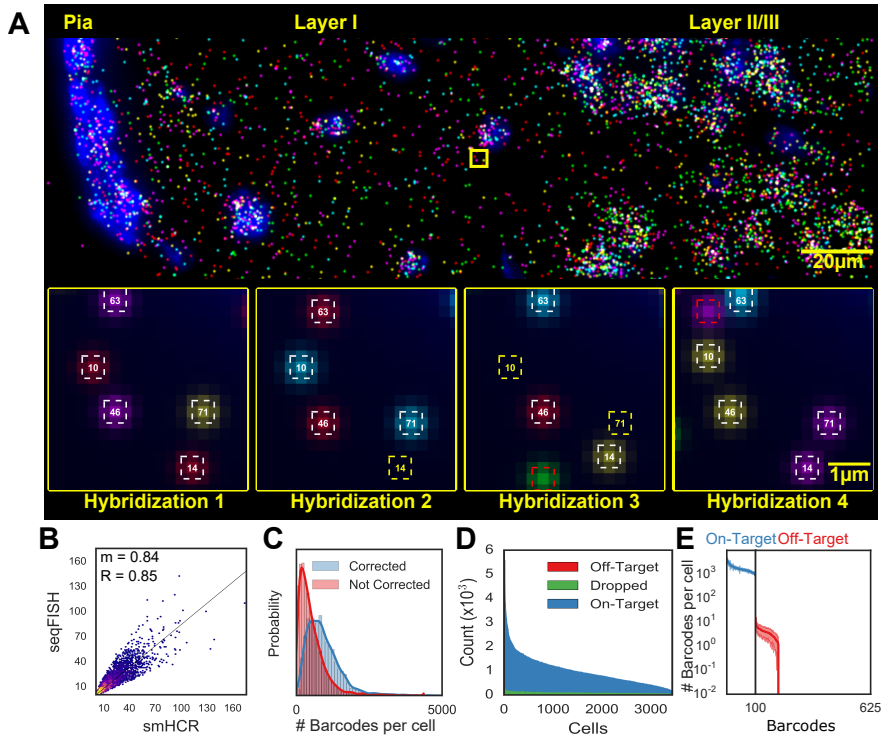


Figure 1: seqFISH generates accurate in situ quantification of mRNA levels. **A.** Image of seqFISH barcoding 100 genes in the outer layer of the mouse cortex. RNA dots in the image are z projected over 15μm. Individual mRNA points are shown across 4 hybridizations in the inset images. White squares correspond to identified barcodes, yellow squares correspond to missing transcripts in a particular hybridization, red squares correspond to spurious false positives and are not counted in any barcode measurements. Numbers in the squares correspond to barcode indices. **B.** seqFISH correlates with smHCR counts. After barcoding, 5 target mRNAs were measured twice by smHCR in the same cells, providing absolute counts of the transcripts. The two techniques correlate with an R=0.85 and a slope (m) of 0.84 (n=3851 measurements). The 2D histogram intensity shows the distribution of points around the regression line. A high density of points is seen along the regression line. The density falls off steeply around the regression line. **C.** Error correction results in a median gain of 373 (25%) counts per cell (n=3497). Red and blue curves correspond to the total barcode counts per cell before and after error correction. **D.** Dropped and off-target barcodes represent a small source of error in seqFISH. 100 on-target barcodes and 525 off-target barcodes are measured per cell. Dropped barcodes are due to at least two overlapping dots appearing within the same region. **E.** Off-target barcodes are rarely observed and contribute minimally to the expression profile in single cells. Each of the 100 on-target barcodes (blue) and 525 off-target barcodes (red) are quantified per cell. The mean is shown with shaded regions corresponding to 1 SD (N=41 imaged regions).

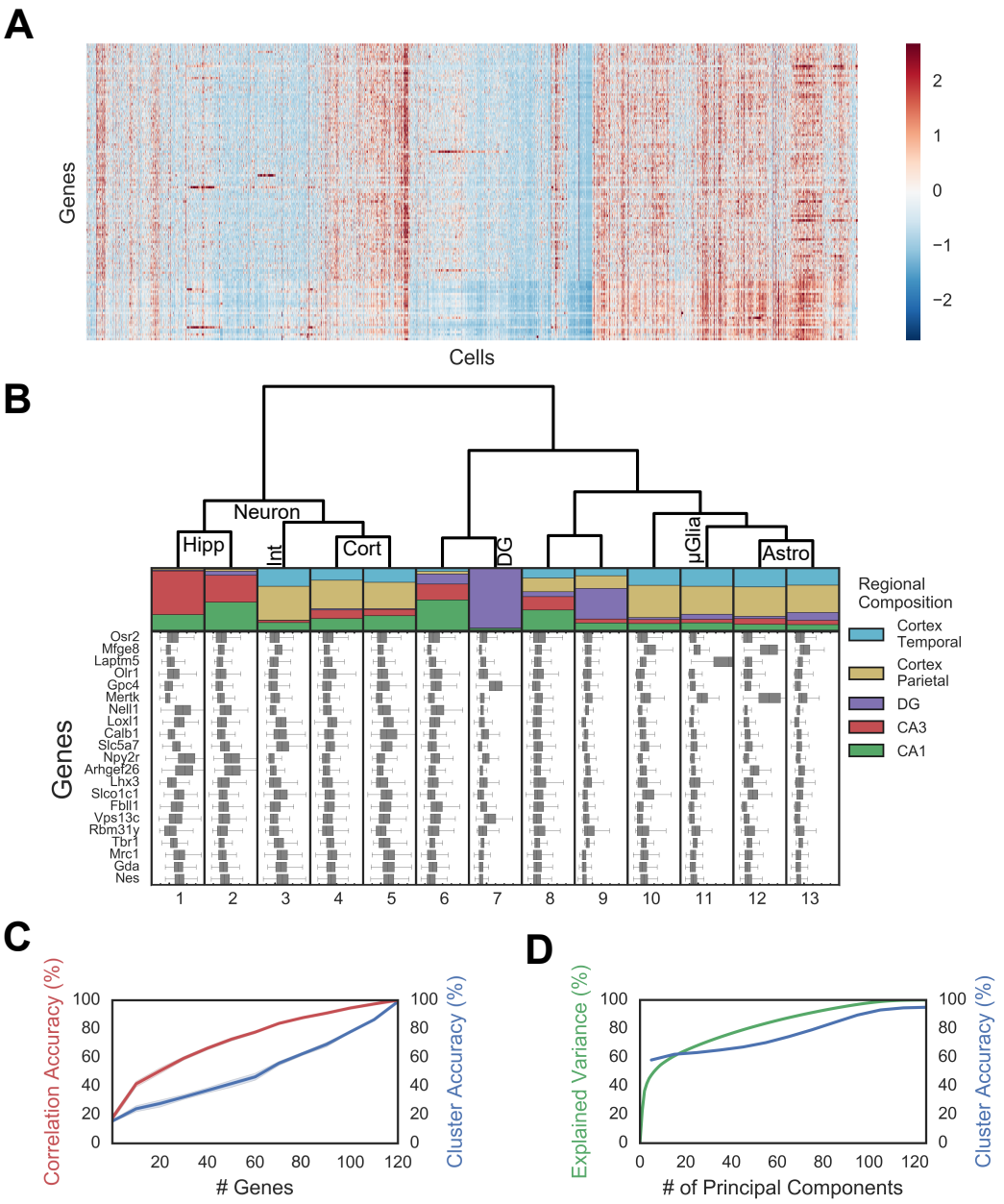

A
Genes
Cells
2
1
0
−1
−2
B
Neuron
Hipp
Int
Cort
DG
μGlia
Astro
Regional Composition
Cortex Temporal
Cortex Parietal
DG
CA3
CA1
Genes
Osr2
Mfge8
Laptm5
Olr1
Gpc4
Mertk
Nell1
Loxl1
Calb1
Slc5a7
Npy2r
Arhgef26
Lhx3
Slco1c1
Fbll1
Vps13c
Rbm31y
Tbr1
Mrc1
Gda
Nes
1 2 3 4 5 6 7 8 9 10 11 12 13
C
Correlation Accuracy (%)
Cluster Accuracy (%)
Genes
D
Explained Variance (%)
Cluster Accuracy (%)
# of Principal Components

Figure 3 *(previous page)*: Distinct clusters of cells exhibit different regional localization in the brain. **A.** Gene expression of 14,908 cells presented as a Z-score normalized heatmap. **B.** Regional compositions of 13 cell clusters are visualized as stacked bar plots with the area corresponding the to number of cells in each region. Hippocampal regions are: CA3, CA1, Dentate Gyrus (DG). Cortical regions: parietal and temporal. Box plot of the Z scores of 21 representative genes are plotted for each cell class. The major tick marks correspond to Z score 0 while every minor tick is a z score interval of 1. Cell type assignments are shown on the dendrogram. Abbreviations: Hippocampus pyramidal (Hipp), cortex (Cort), Dentate Gyrus (DG), Interneurons (Int), Astrocyes(Astro), Microglia (μGlia). **C.** Any random subset of 25 genes can recapitulate approximately 50% of the information in the correlation amongst cells (red), but a larger number of genes are required to accurately assign cells to cluster using a random forest algorithm (blue) (n=10 bootstrap replicates; shading is 95% CI), indicating that fine structures in the data require quantitative measurements of combinatorial expression of many genes. **D.** Similar to C, while the first ten PCs explain the coarse structure, a larger number of principal components (PCs) are required to describe the full data. Expected variation (green) and accuracy in predicting cell identity using a random forest model (blue).

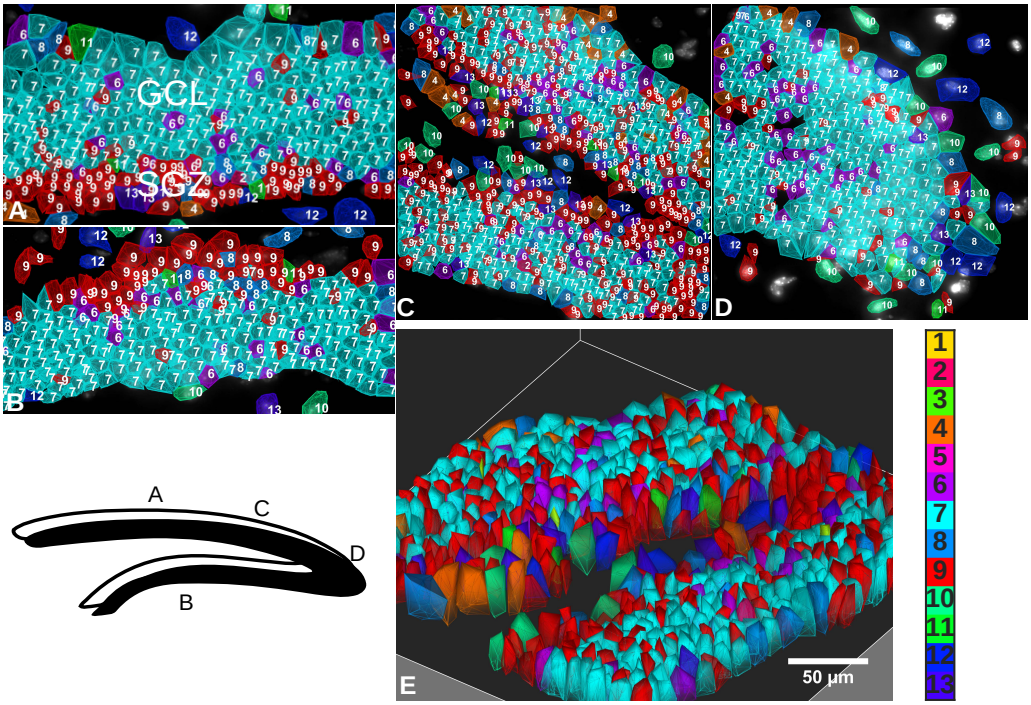


Figure 4: Spatial layering of cell classes in the Dentate Gyrus (DG). **A-B.** Suprapyramidal and infrapyramidal blades of DG. Cells of the subgranular zone and granule cells are arranged in lamina layers in mirror symmetric patterns on the upper and lower blades. **C.** The SGZ stays on the inner layer of the DG fork. **D.** Cells are patterned in the crest. Numbered color key corresponds to cluster numbers in Fig3b. **E.** Letters in the cartoon of DG correspond to images. **F.** 3D image of the fork region shown in C.

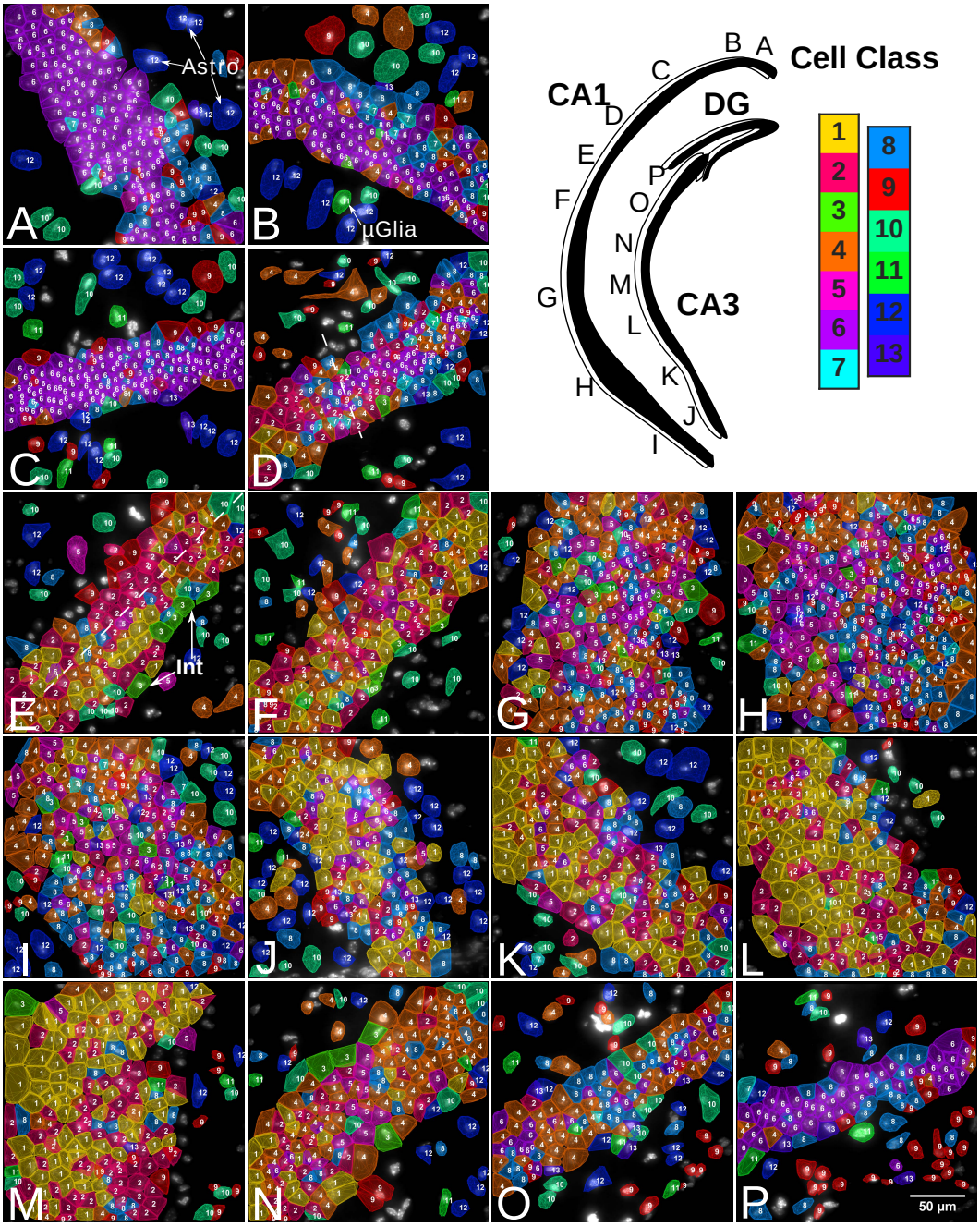

Astro
μGlia
Int
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
50 μm
CA1
DG
CA3
Cell Class
1
2
3
4
5
6
7
8
9
10
11
12
13

Figure 5 *(previous page)*: Subregions of the hippocampus are composed of distinct compositions of cell classes based on the first 125 gene experiment. Upper right panel. Cartoon of hippocampus with imaged regions labeled. Color key corresponds to the classes in Fig3b. **A-D.** These images are regions from the CA1d. Astrocytes (Astro) are marked in image A and a microglia cell (μGlia) is marked in image B. Moving along the hippocampus from CA1 dorsal to ventral, cell classes transition from a homogenous dorsal population (C to D) to a mixed population in the CA1 intermediate **(E-F)** to regions of even larger cellular diversity in the CA1 ventral region **(G-I).** The dotted line in D marks the transition point of the CA1d to the CA1i. E shows two laterally segregated cell classes (marked by a dotted line) in the CA1i along with cholinergic interneurons (Int) on the interior surface of the CA1i. The ventral **(J-K)** and intermediate CA3 **(L-M)** have similar cell classes compositions to the CA1v and CA1i. The two last regions **(O-P)** of the dorsal CA3 shows distinct cell classes compositions that are relatively homogeneous within a field but are different than other fields of CA3. The cell class composition of field P is similar to that of the CA1d, but these cluster 6 cells are grouped into a distinct subcluster.

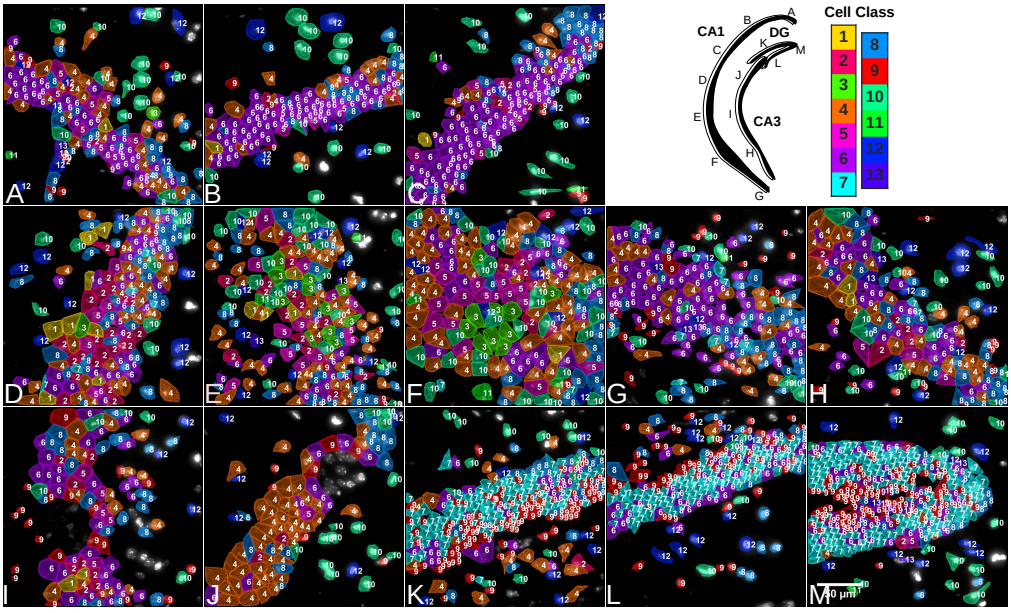


Figure 6: Mapping of cell types to a second brain slice with 125 genes. Upper right panel. Cartoon of hippocampus with imaged regions labeled. Color key corresponds to the classes in Fig3b. **A-D.** Similar to the cell class compositions shown for the hippocampus in Fig 5, CA1d in this second coronal section from a second mouse is composed of mostly cluster 6 cells. **(E)** CA1i region and **(F-G)** the CA1 ventral regions are again composed of similar cell classes to that shown in figure 5 with increasing diversity of cell class compositions from the CA1d to the CA1i to finally the CA1v. **(H-J)** CA3 regions. **(K-M)** DG regions showing the same cell classes and layer pattern of the GCL and SGZ shown in Figure 4.

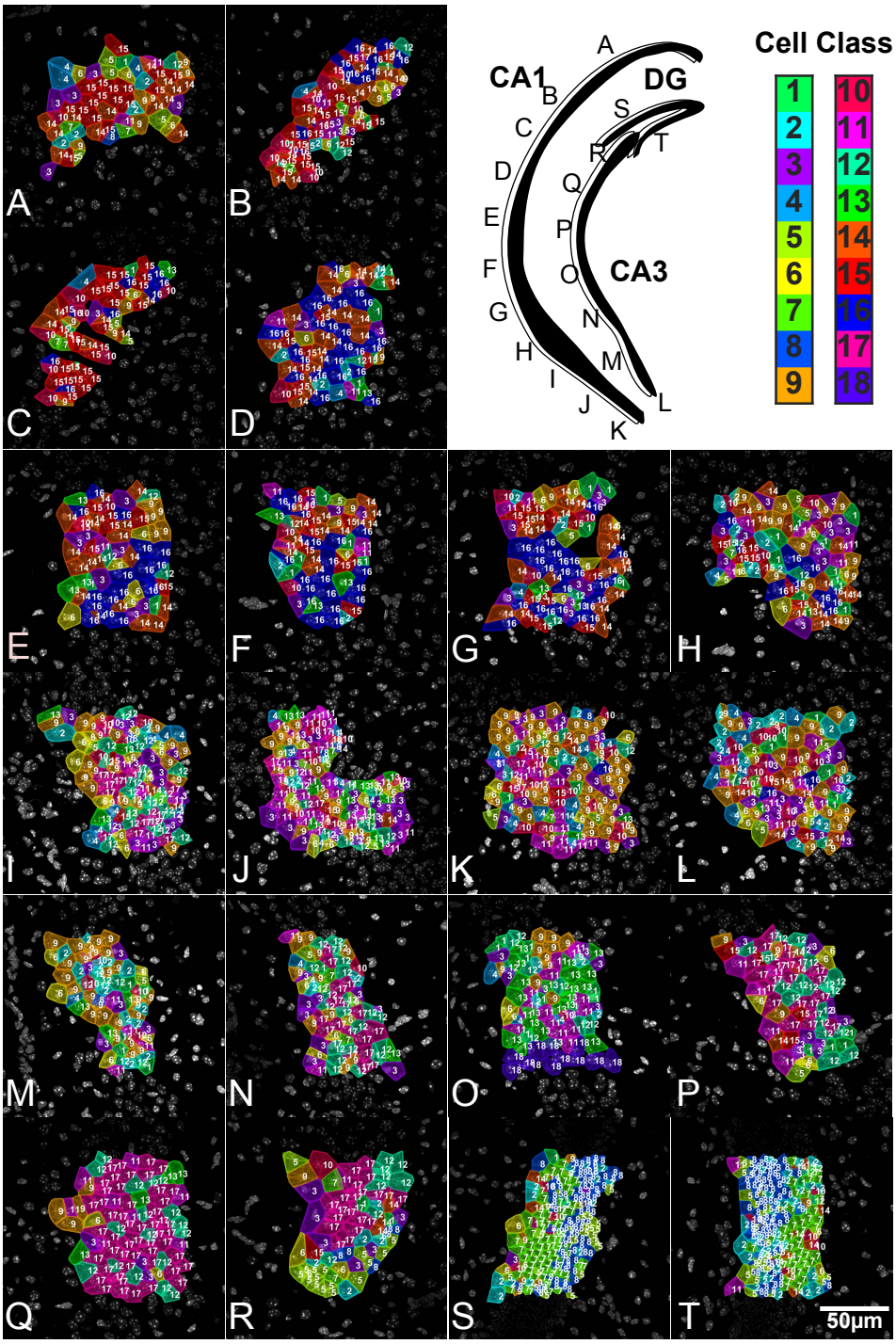

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
CA1
DG
CA3
Cell Class
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
50µm

Figure 7 *(previous page)*: Mapping of cell types to a third brain slice with 249 genes. Upper right panel. Cartoon of hippocampus with imaged regions labeled. Color key corresponds to the classes in Fig S6C. **A-C.** Similar to the slice shown in Fig 5 and 6, CA1d is relatively homogenous in cell cluster composition. **D-G.** Images from the CA1i region show that the cell class composition is different from that of the CA1d. **H-K**. Again, similar to Fig 5 and 6, images from the CA1 ventral regions shows a much more complicated cellular composition and a high degree of cellular heterogeneity. **L-R.** Images from the CA3 region show that the cellular compositions also creates 3-4 subregions within the CA3. The cellular heterogeneity of the CA3 subregions mirrors that of the CA1, where the ventral region of the CA3 is very heterogenous while the dorsal region of the CA3 is relatively homogenous. **S-T.** The DG regions show the distinct SGZ versus GCL layering pattern seen in the previous two brains.

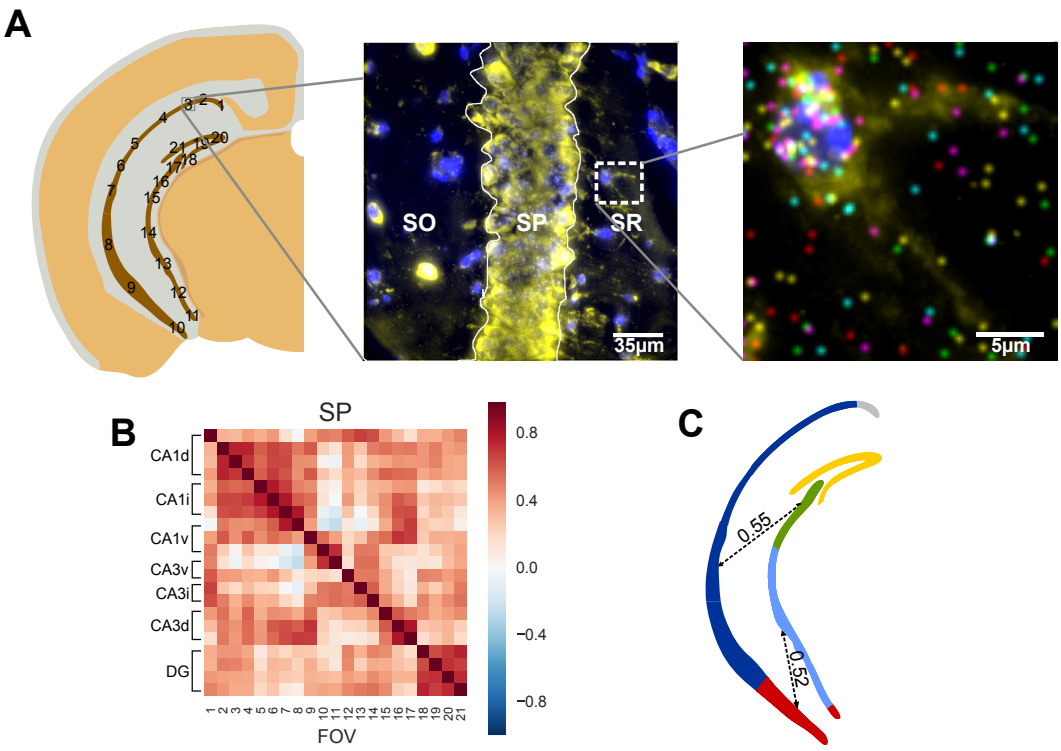


Figure 8: Correlations of the transcription profile across the pyramidal layer. **A.** mRNA counts in the cell bodies in the Stratum Pyramidale (SP) are grouped within each field of view. A single cell in the Stratum Radiatum (SR) is shown to illustrate individual mRNA localization. Stratum Oriens (SO) is labeled for orientation. **B.** mRNAs in different subregions of pyramidal layer show both long-distance spatial correlations as well as local correlations between neighboring fields. Both CA1 and Dentate Gyrus (DG) show high regional correlations. Correlation is calculated based on the 125 gene experiment. **C.** Illustration of regional and long distance correlation patterns observed in B. Correlated regions are colored and long distance correlations are shown as dotted lines with their median correlation coefficient written over the dotted line.

3.7 Supplemental Data and Figures

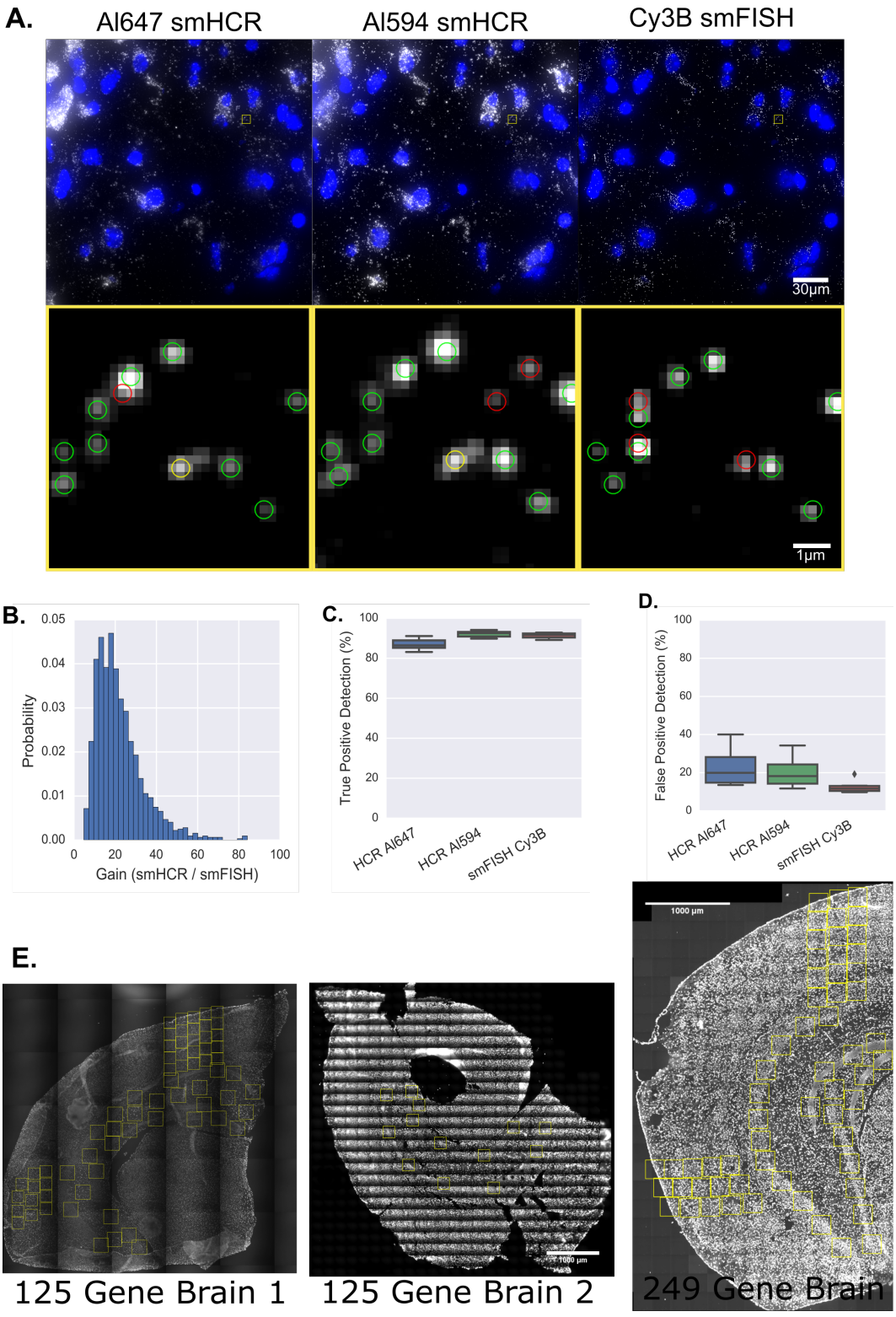
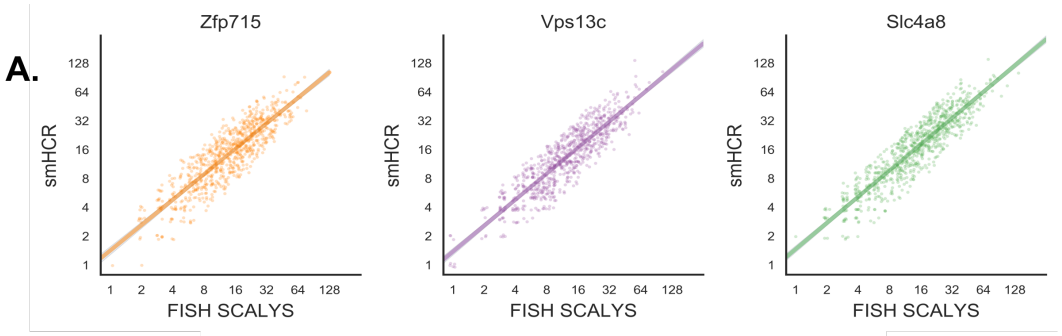
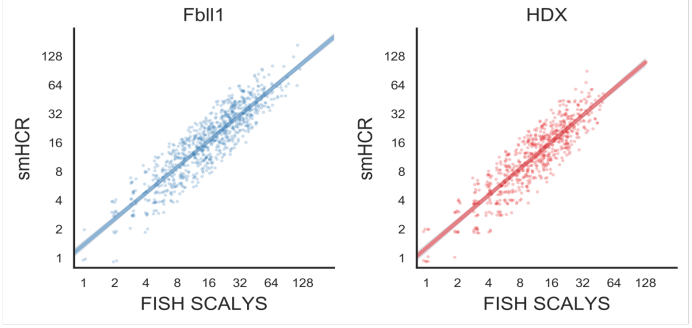
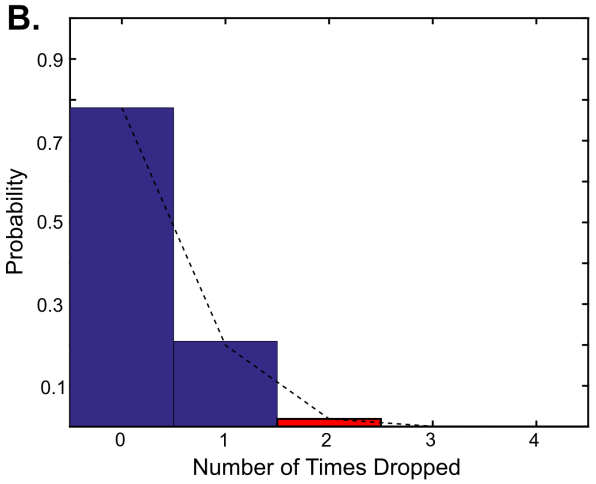
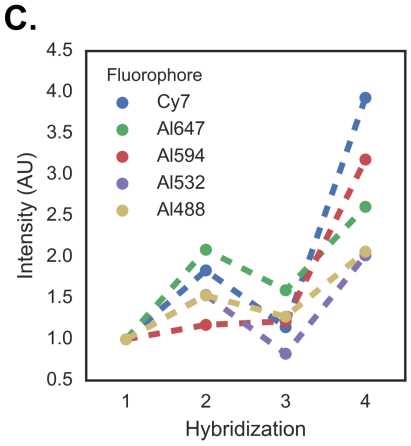


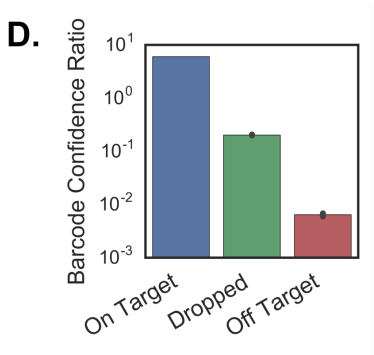
Figure S1 *(previous page)*: SmHCR performance metrics as compared to smFISH, Related to Figure 1. **A.** Raw data of Pgk1 transcripts imaged in a brain slice. The transcript was targeted with 2 hcr probes sets and 1 smFISH probe set, each consisted of 24 oligonucleotide probes. The probe sets were hybridized together and were imaged in 3 different channels. Green circles are transcripts detected in all channels, yellow circles signify transcripts detected in 2 out of 3 channels, and red circles represent signal found in only 1 channel (false positives due to nonspecific binding). These images show that smHCR and smFISH have similar sensitivity, specificity, and spot size. **B.** Gain of smHCR vs smFISH. The mean gain of smHCR is 22.1 ± 11.55 vs smFISH (n=1338). **C.** True positive detection rate of smHCR and smFISH per channel. The percent of true positives (transcripts detected with at least 2 out of 3 probe sets) detected with each probe set (n=1338). **D.** False positive rate of smHCR and smFISH. Percent of total dots in a channel not detected in any other channel for 3 color Pgk1 (n=1338). **E.** All the regions imaged in the coronal section are boxed. Each box represents a field of 216 µm x 216 µm. The brain section used for figure 4 and 5 is shown on the left. The middle section is used for figure 6 and the right section is used for figure 7.


A.
Zfp715
Vps13c
Slc4a8
smHCR
FISH SCALYS
1
2
4
8
16
32
64
128


Fbll1
HDX
smHCR
FISH SCALYS
1
2
4
8
16
32
64
128


B.
Probability
0.9
0.7
0.5
0.3
0.1
0
1
2
3
4
Number of Times Dropped


C.
Fluorophore
Cy7
Al647
Al594
Al532
Al488
Intensity (AU)
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
1
2
3
4
Hybridization


D.
Barcode Confidence Ratio
10^1
10^0
10^-1
10^-2
10^-3
On Target
Dropped
Off Target

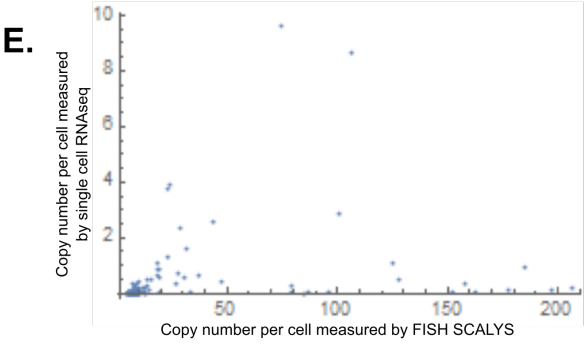

E.
Copy number per cell measured by single cell RNAseq
10
8
6
4
2
50
100
150
200
Copy number per cell measured by FISH SCALYS

Figure S2 *(previous page)*: Quantitation of seqFISH, Related to Figure 2. **A.** All control genes show high correlations between seqFISH and smHCR. **B.** Number of dropped hybridizations from the barcode. Blue bars represent measured probability and the red bars represent inferred values from binomial distribution fitting of measured probability. The ratio of the full barcodes (4 hybridizations) vs 3 hybridization barcodes indicate that transcripts that are mis-hybridized in 2 rounds are rare. Transcripts missed in 2 or more hybridizations (red bars) could not be recovered from the error-correction algorithm and would be dropped from our quantifications (N=2,115,477 total barcodes). **C.** Intensity of barcode hybridizations overtime. All dots belonging to barcodes are quantified in each hybridization and their mean intensity is plotted over time normalized to the first hybridization. 99% CI ratio of mean is plotted as a bar over points, but is not visible due to its small size (n=60143 to 111284 points per channel). **D.** Barcoding confidence ratio. Barcode classes in D are compared to a null model of barcode observations where random chance observation should give a ratio of 1. Off target barcodes are observed 0.005 times less than expected, suggesting that seqFISH has high accuracy in correctly counting barcoded transcripts (n=3493 cells). Dark bars on top of bar plots correspond to 99.999% confidence interval determined by bootstrap resampling. **E.** Comparison of average copy numbers per gene as measured by Zeisel et al. and seqFISH. Single cell RNA-seq underestimates copy numbers compared to seqFISH.

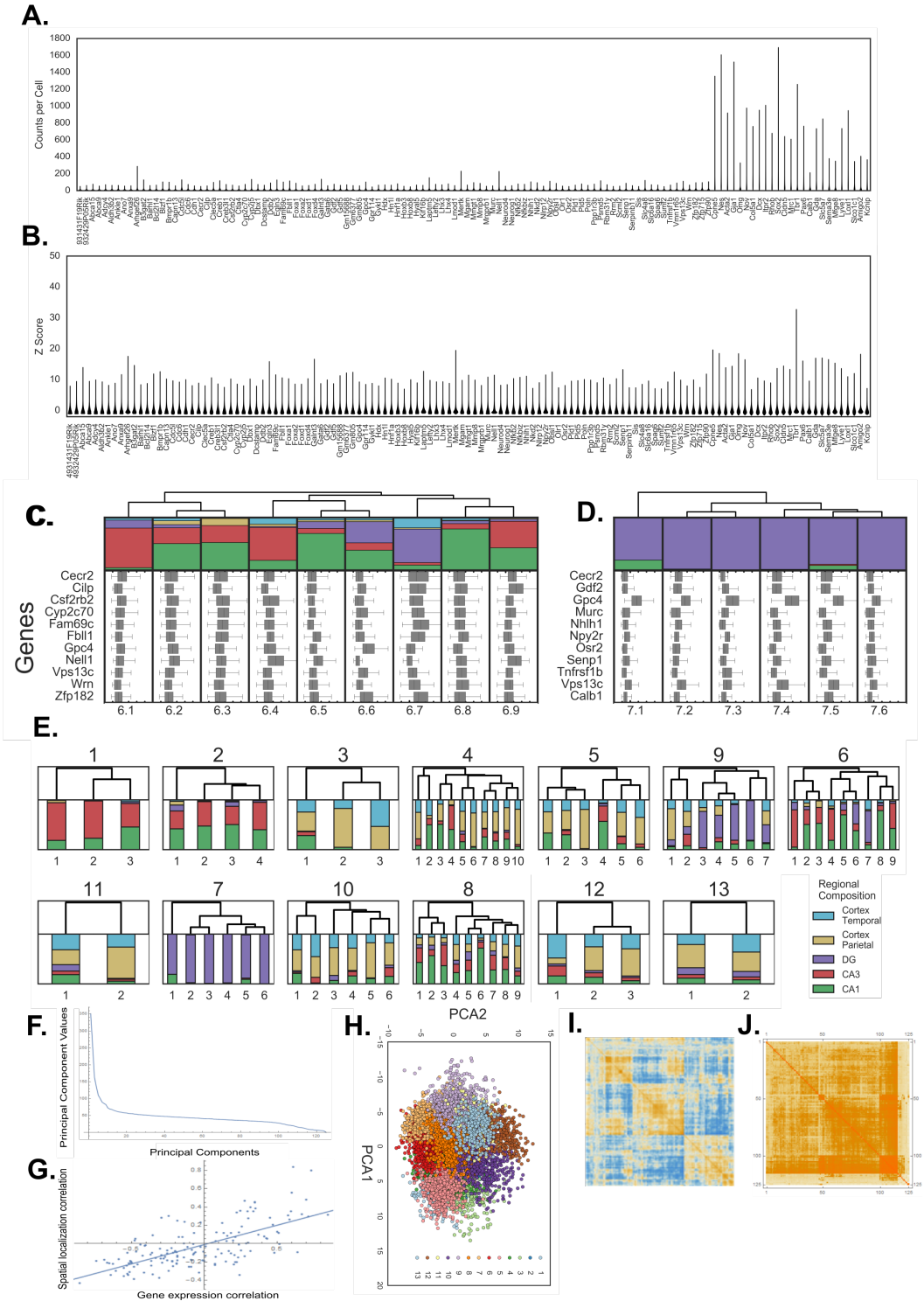

A.
Counts per Cell
B.
Z Score
C.
Genes
Cecr2
Cilp
Csf2rb2
Cyp2c70
Fam69c
Fbll1
Gpc4
Nell1
Vps13c
Wrn
Zfp182
6.1
6.2
6.3
6.4
6.5
6.6
6.7
6.8
6.9
D.
Cecr2
Gdf2
Gpc4
Murc
Nhlh1
Npy2r
Osr2
Senp1
Tnfrsf1b
Vps13c
Calb1
7.1
7.2
7.3
7.4
7.5
7.6
E.
1
2
3
4
5
9
6
11
7
10
8
12
13
Regional Composition
Cortex Temporal
Cortex Parietal
DG
CA3
CA1
F.
Principal Component Values
Principal Components
G.
Spatial localization correlation
Gene expression correlation
H.
PCA2
PCA1
I.
J.

Figure S3 *(previous page)*: Gene expression patterns and clustering of the 125-gene dataset, Related to Figure 3. **A.** Overview of 125 gene expression. Plots show the distribution of each transcript in all 14,908 imaged cells. Note the last 25 genes have higher expression and were imaged with serial hybridization. **B.** Violin plots of Z-score distribution for 125 genes. **C.** Subclusters of cluster 6 cells and their regional localization and gene expression profile displayed under the dendrogram. Subcluster 6.1 is enriched in the CA3, while 6.7 is enriched in the DG. **D.** Subclusters of cluster 7 cells are shown. Almost all cells are localized in the GCL but have different combinatorial expression profiles. Note Calb1 expression, which marks out granule cell maturation, differs amongst subclusters. **E.** Sub-cluster hierarchy of each of the 13 clusters identified in Figure 3B. **F.** PCA eigenvalue analysis of the cell-to-cell correlation matrix. First 125 PC and their eigenvalues are shown. As observed in Fig 3, the first 10 PCs explain 59.5% of the variation in the data, while the remaining 115 PCs are needed to explain remaining data. Reflecting this, the eigenvalues of the first 10 components are high, while the remaining eigenvalues are uniform. **G.** Correlation between gene expression and spatial localization. Each dot represents a pair of cell classes and their correlations in gene expression expression space (x) and spatial localization patterns (y) (N=153 pairwise correlations between classes, R=0.67). Classes that are similar in expression have similar localization patterns. **H.** PCA decomposition separates cells into coherent clusters corresponding to cell classes. Cells are colored according to the clusters displayed in the dendrogram. **I.** Cell to cell correlation map for all 14,908 cells images in the 125 gene experiment. **J.** Gene to gene correlation map for all 125 genes measured in the 125 gene experiment.

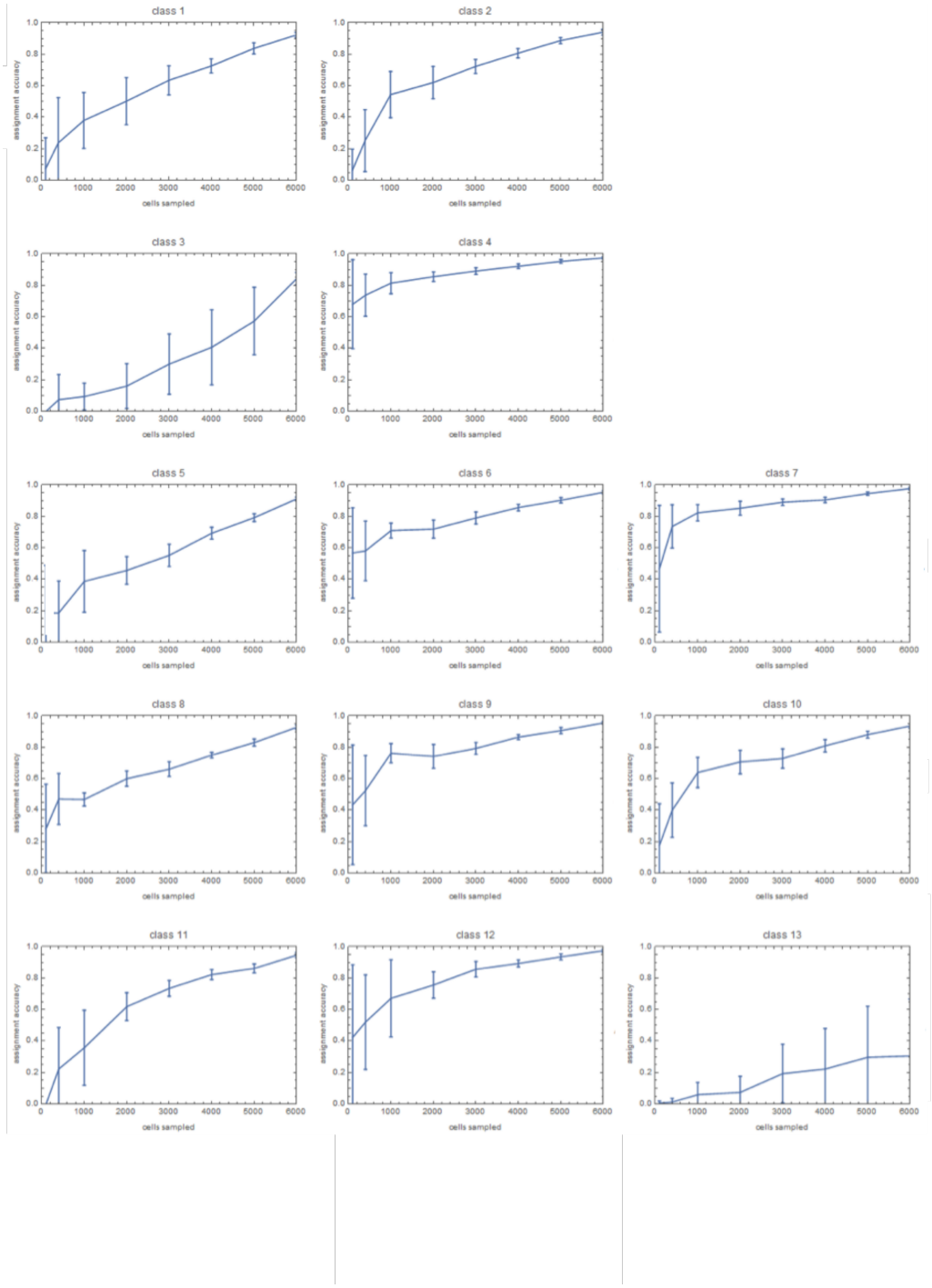
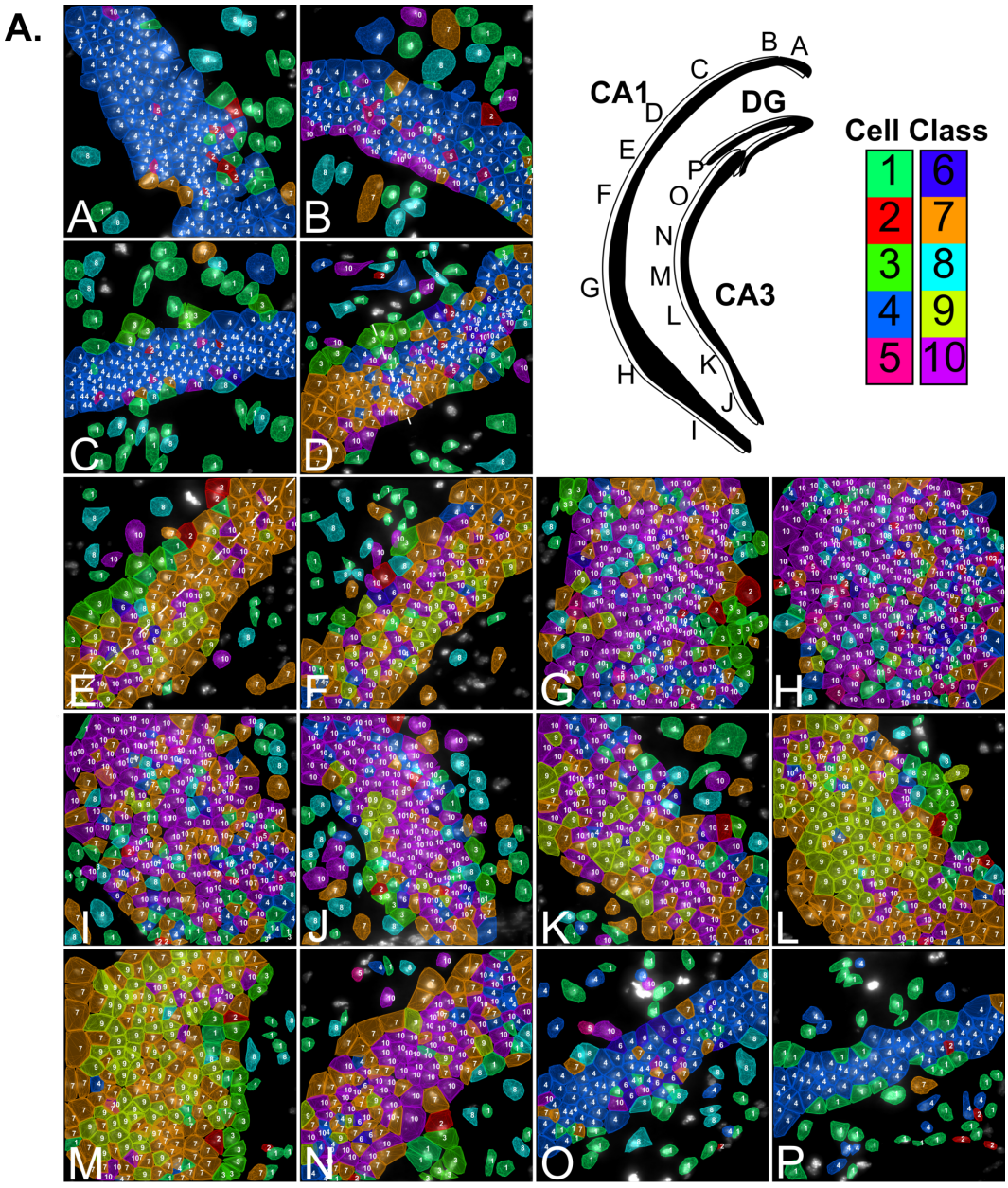


Figure S4: Robustness of cell classes to downsampling of cells, Related to Figure 3. To measure how well cluster assignments perform with a limited number of cells, a random forest model was trained on the cell-to-cell correlation matrix of subsets of 14,908 cells. The robustness of the clusters was calculated by applying this model to classify the remaining cells and determining the percent accuracy of correct assignment to the clusters presented in 3b. While some classes can be assigned accurately even with a small number of cells as the initial training set, several classes require large number of cells to accurately assign (n=10 bootstrap replicates, S.E.)


A.
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
CA1
DG
CA3
Cell Class
1
2
3
4
5
6
7
8
9
10

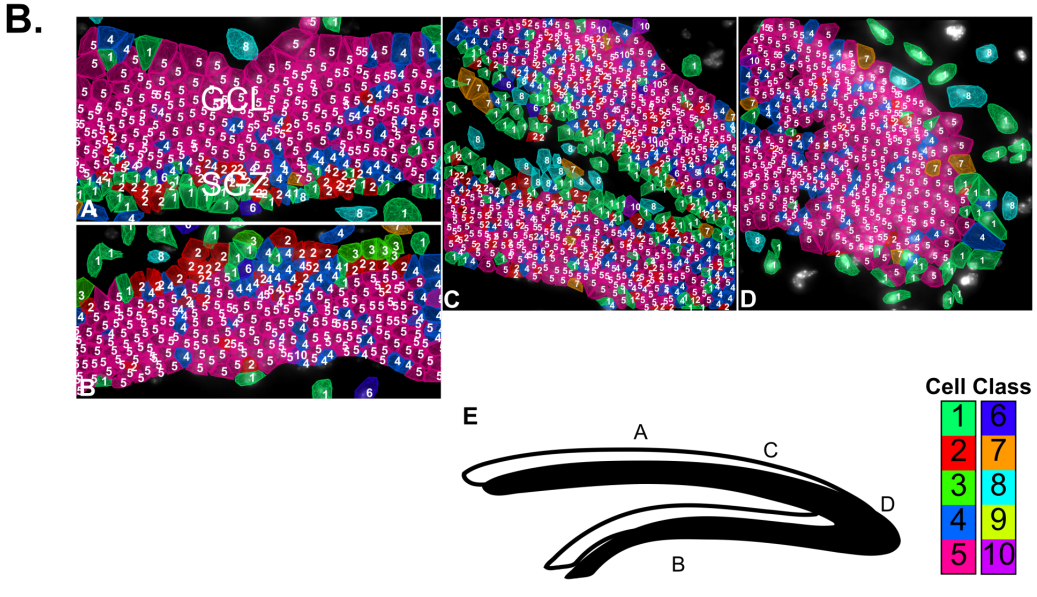

B.
GCL
SGZ
A
B
C
D
E
Cell Class
1
2
3
4
5
6
7
8
9
10

Figure S5 *(previous page)*: The same pattern of hippocampal subregions are observed when only hippocampal cells are clustered, related to Figures 3-5. **A.** In Figure 5 and 6, both cortex and hippocampal cells were used in the clustering. When only cells from the hippocampus are used for clustering, the same patterns are observed with homogenous cell populations in CA1d and CA3d. The intermediate and ventral subregions contain heterogeneous cell clusters. **B.** The laminar patterning in the dentate gyrus is also observed similar to Figure 4.

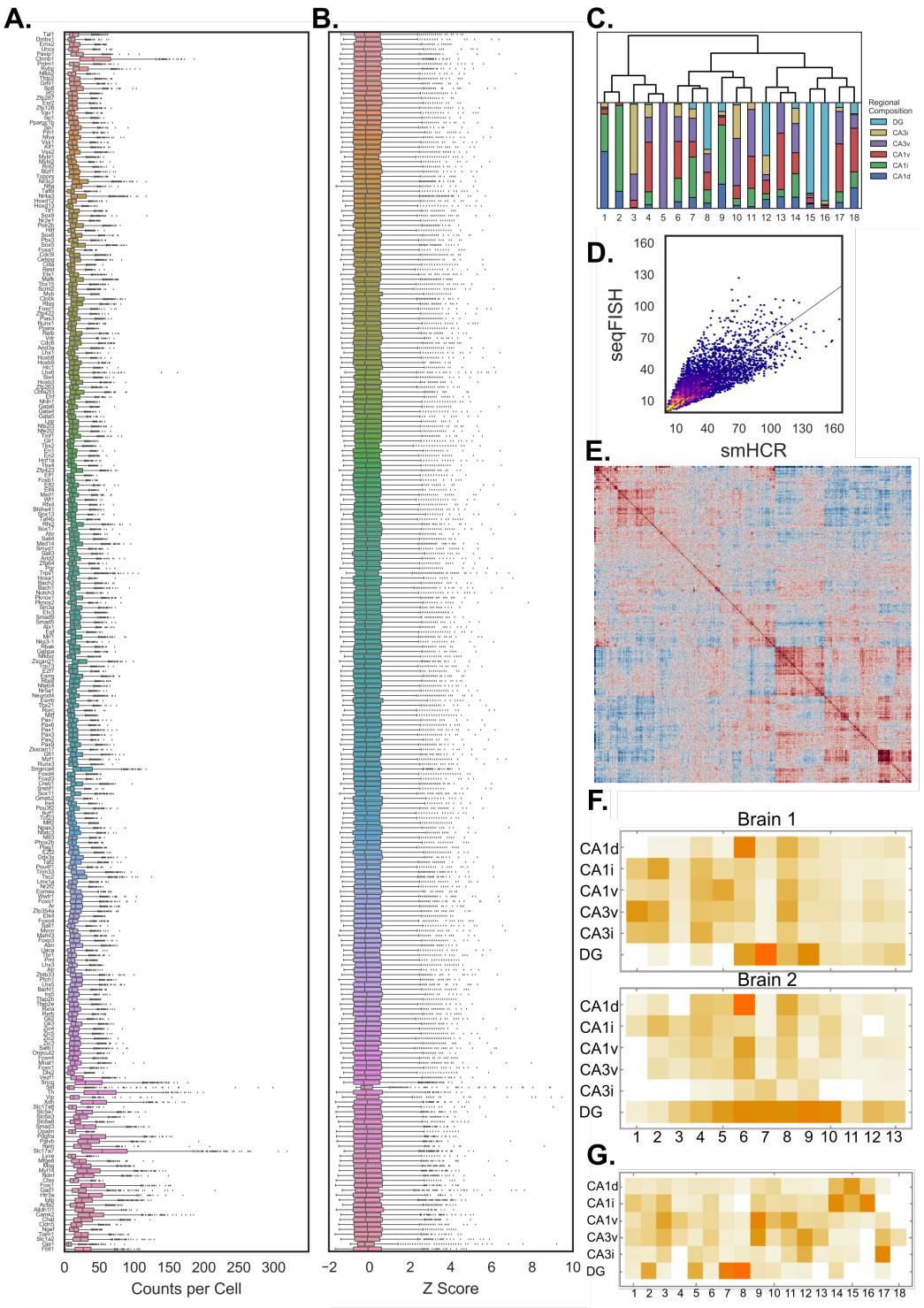
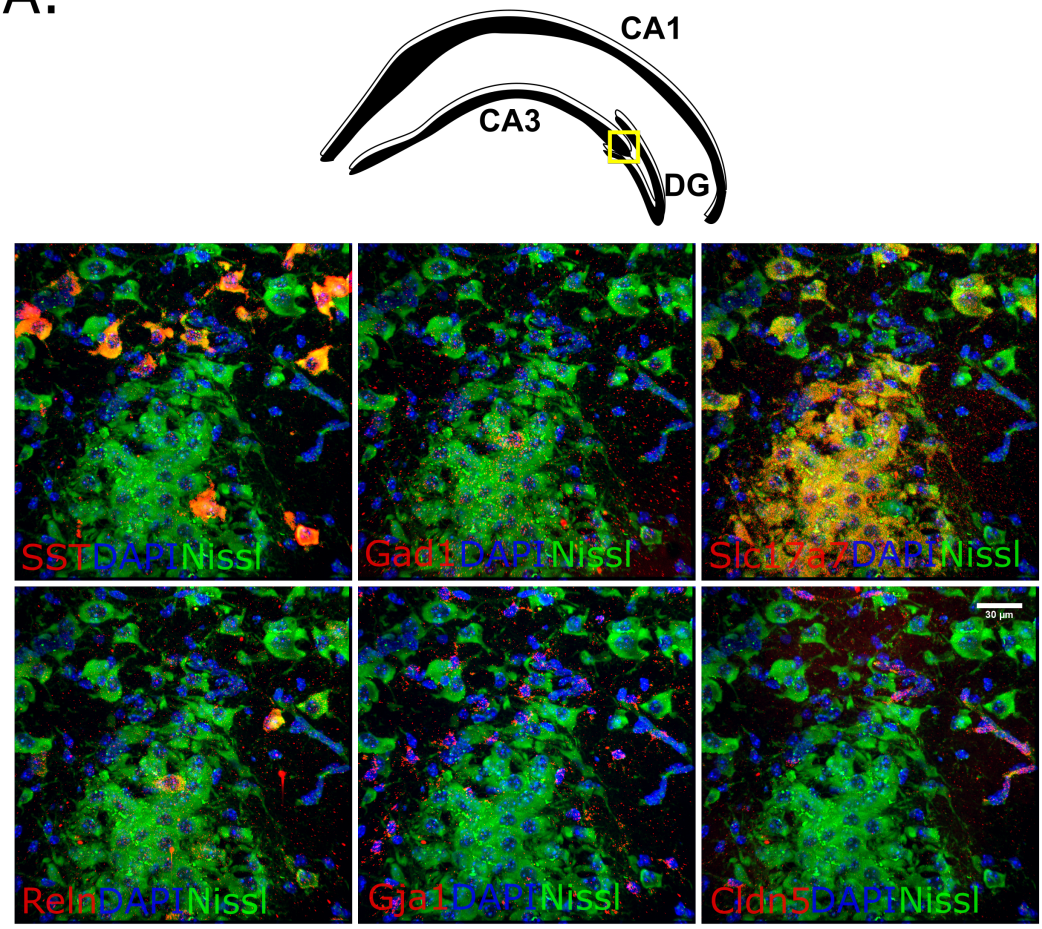

A.
B.
C.
D.
E.
F.
G.
0 50 100 150 200 250 300
Counts per Cell
−2 0 2 4 6 8 10
Z Score
Regional Composition
DG
CA3i
CA3v
CA1v
CA1i
CA1d
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
seqFISH
smHCR
10 40 70 100 130 160
Brain 1
Brain 2
CA1d
CA1i
CA1v
CA3v
CA3i
DG
1 2 3 4 5 6 7 8 9 10 11 12 13

Figure S6 *(previous page)*: Gene expression patterns and clustering of the 249-gene dataset, Related to Figure 7. **A.** Overview of 249-gene expression. Plots show the distribution of each transcript in all 2050 imaged cells in the hippocampus. Note the last 35 genes have higher expression and were imaged with serial hybridization. **B.** Violin plots of Z-score distribution for 249 genes. **C.** Dendrogram with regional localization of the 18 cell clusters for the 249-gene experiment. **D.** Correlation of seqFISH counts to smHCR counts for the 249-gene experiment. The 2D density histogram shows a high density of points around the regression line that fall off towards the edges of the distribution. **E.** Cell-to-cell correlation for all 2050 cells in the 249-gene dataset. **F.** Heat map of the percentage of each cell class in each region of the hippocampus for both the 125-gene experiments. These heat maps show that in both 125-gene experiments the same cell classes are used in roughly the same proportions. **G.** Heat map of the percentage of each cell class in each region of the hippocampus for the 249-gene experiment. The same patterns are seen as the 125 gene experiment (i.e. different regions use different cell classes in varying amounts).

A.


CA1
CA3
DG
SSTDAPINissl
Gad1DAPINissl
Slc17a7DAPINissl
ReInDAPINissl
Gja1DAPINissl
Cldn5DAPINissl
30 µm

B.

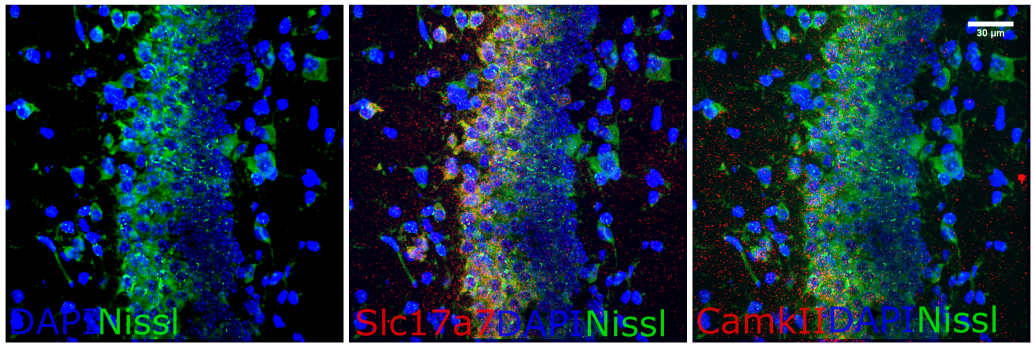

DAPINissl
Slc17a7DAPINissl
CamkIIDAPINissl
30 µm

Figure S7 *(previous page)*: Marker genes robustly identify cell types, Related to Figure 7. **A.** The top panel outlines the region of the hippocampus being shown in a yellow box. The images show the raw gene expression patterns seen using smHCR in our data at the dorsal most tip of the CA3 for a representative set of cell identity markers used in the 249 gene experiment. The transcript expression profile is shown in red, Nissl staining is shown in green, and DAPI staining is shown in blue. Each image shown is the full field of view and a maximum intensity projection over 15 um. **B.** Set of images showing the distinction between the GCL and SGZ. The GCL shows a high level of Nissl staining and expression of neuronal genes such as slc17a7 and camkII. The SGZ shows an absence of Nissl staining and terminal neuron marker genes. The transcript expression profile is shown in red, Nissl staining is shown in green, and DAPI staining is shown in blue. Each image shown is the full field of view (216 um x 216 um) and a maximum intensity projection over 15 µm.

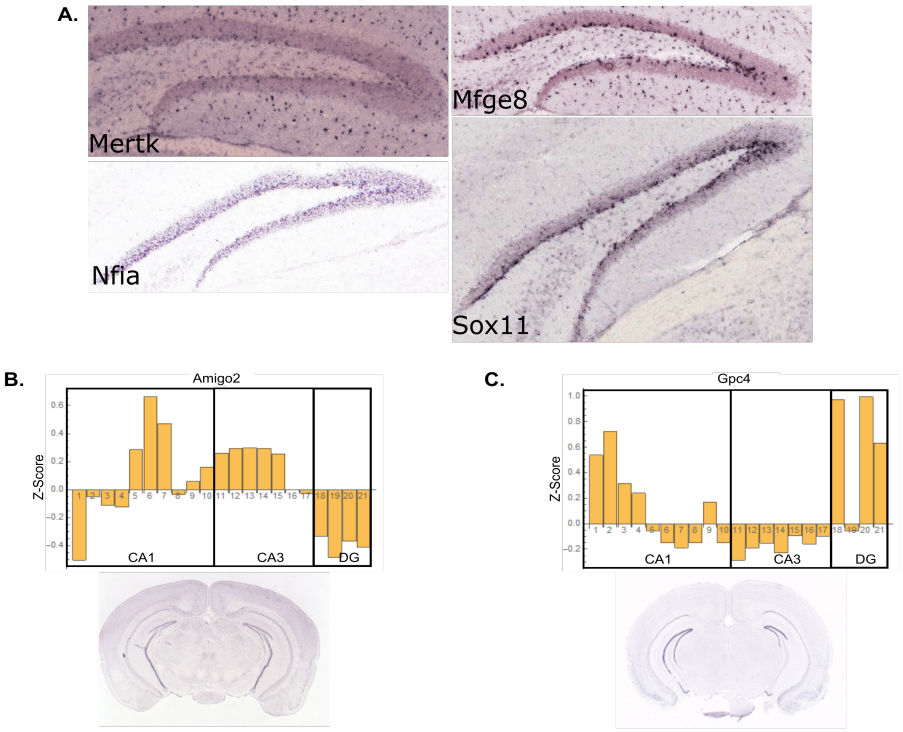


Figure S8 *(previous page)*: Comparison of SeqFISH expression data to Allen Brain Atlas expression data, Related to Figure 8. **A.** ISH data from the Allen Brain Atlas for genes seen to be enriched in the SGZ in the 125 and 249 gene seqFISH experiments. In the 125 gene experiment, mertk and mfge8 were found to be enriched in the SGZ. In the 249 gene experiment, nfia and sox11 were seen to be enriched in the SGZ. ABA ISH data shows similar patterns to those observed with seqFISH for the SGZ. B-C. Comparison of averaged z-score values per cell from seqFISH to ABA data across hippocampus. **B.** Amigo2 Z-score profile found across the different fields of the hippocampus using seqFISH is shown on top and the ABA ISH image for Amigo2 is shown on the bottom. **C.** Gpc4 Z-score profile found across the different fields of the hippocampus using seqFISH is shown on top and ABA ISH image for Gpc4 is shown on the bottom.

3.8 Supplemental Tables

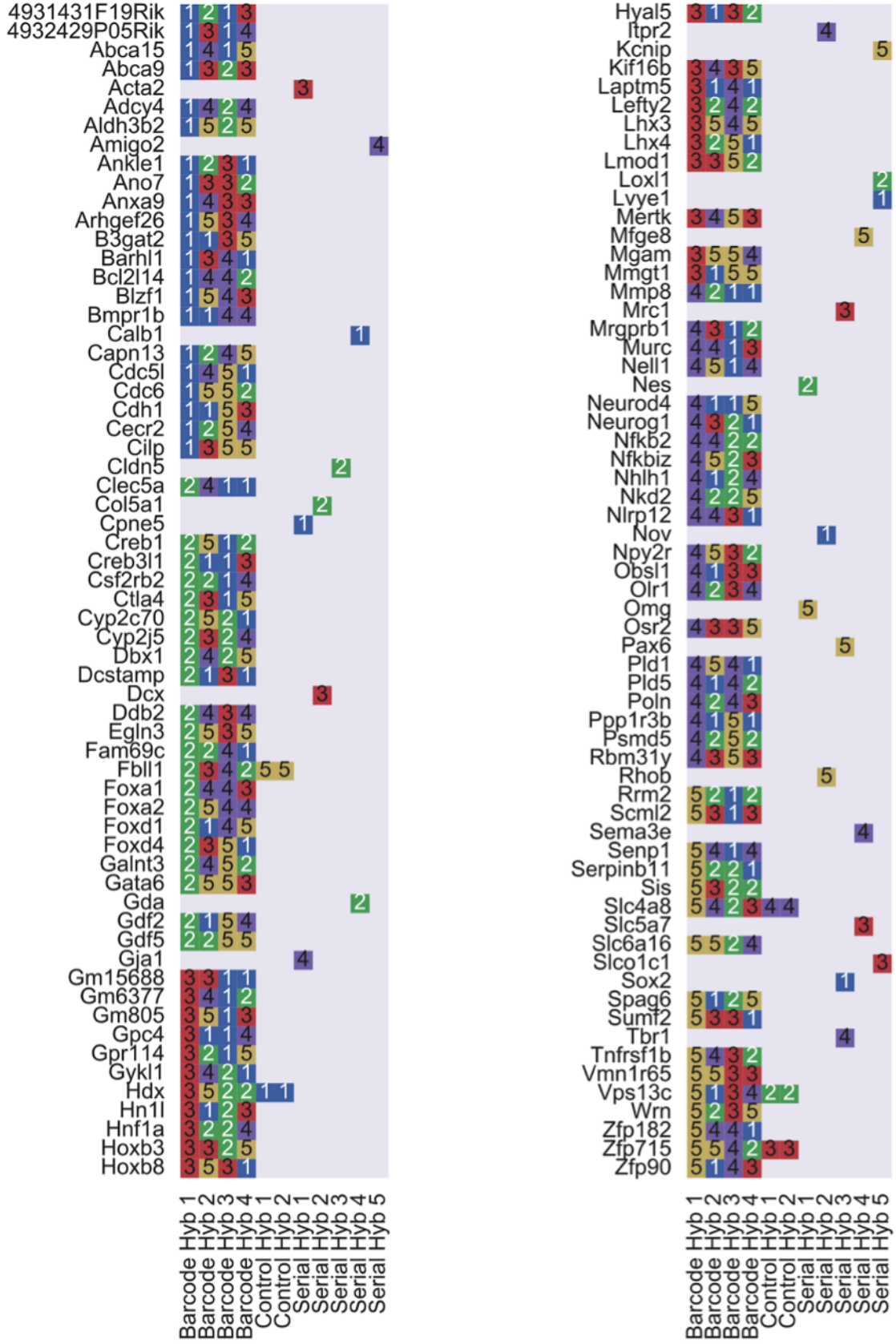


Table S1, Related to Figure 1.[1]

Barcode assignments in the 125-gene seqFISH and serial experiment. 125 genes are profiled, 100 of which are barcoded and 25 are identified by serial smHCR hybridizations. Five control genes (Hdx, Vps13c, Zfp715, Fbll1, Slc4a8) were quantified by both techniques. The smHCR round of hybridization of control genes were performed twice to colocalize signal to obtain an absolute count.

Table S2, Related to Figure 3-6. (Provided as a separate Excel file)

Cluster group data for both 125-gene experiments. The "Major Cluster" column A defines the large cluster number. The "sub-cluster index" column B gives the subcluster number of the cluster within the major cluster. The number of cells in each subcluster and the location of those cells are tabulated in the columns C-I. Column J lists the top 4 enriched genes in the subcluster.

Table S3, related to Figure 1 and Figure S1. (Provided as a separate Excel file)

Sequences for RNA integrity test. 48 probes targeting the PGK1 transcript was used. 24 probes were amplified with initiator B1 and the remaining 24 probes were amplified with initiator B3.

Table S4, Related to Figures 1-6. (Provided as a separate Excel file)

Probe list for 125-gene barcoding experiment. Sheet 1 to sheet 4 gives the full oligoarray synthesized sequence for hybridizations 1 to 4 respectively. Sheet 5 and 6 give the sequences of the 25 high copy number genes that were targeted. For sheets 1-4, Column A gives the gene name. Columns B and J give the forward and reverse amplification primer, respectively. Columns C and I give the restriction site sequences. Column D and H give the restriction site spacer sequences. The final probe sequence is can be made by concatenating columns E-G. For sheets 5-6, column A is the gene name and concatenating Column B-D gives the final probe. The second sheet gives the readout probes. Column A is the gene name and concatenating Column B-D gives the final readout probe for HCR.

Table S5, related to Figure 4, 5, and 6. (Provided as a separate Excel file)

Raw expression data for 125 genes in brain 1 and brain 2 cells. For sheet 1, each row represents a single cell and each column represents the mRNA count within a

[1]All supplementary tables can be downloaded as Microsoft Excel files from the following link: http://www.cell.com/neuron/fulltext/S0896-6273(16)30702-4

cell for a specific gene in brain 1. For sheets 2-4, each column represents a single cell and each row represents the mRNA count within a cell for a specific gene.

Table S6, Related to Figure 7. (Provided as a separate Excel file)

Raw expression data for 249 genes in Brain 3 cells. Each column represents a single cell and each row represents the mRNA count within a cell for a specific gene.

Table S7, Related to Figure 7. (Provided as a separate Excel file)

Cluster group data for 249-gene experiment. The "Major Cluster" column A defines the cluster number. The number of cells in each sub-cluster and the location of those cells are tabulated in the columns C-I. Column J lists the top 4 enriched genes in the cluster.

Table S8, Related to Figure S7. (Provided as a separate Excel file)

Barcode assignments in the 249-gene seqFISH and serial experiment. 249 genes are profiled, 214 of which are barcoded and 35 are identified by serial smHCR hybridizations. Four control genes (Smarca4, Sin3a, Npas3, and Neurod4) were quantified by both techniques.

3.9 References

Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., et al. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Natl. Acad. Sci. U. S. A. 109, 21301–21306.

Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., and Hess, H.F. (2006). Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. Science 313, 1642–1645.

Breiman, L. (2001). Random Forests. Mach. Learn. 45, 5–32.

Cajigas, I.J., Tushev, G., Will, T.J., Dieck, S. tom, Fuerst, N., and Schuman, E.M. (2012). The Local Transcriptome in the Synaptic Neuropil Revealed by Deep Sequencing and High-Resolution Imaging. Neuron 74, 453–466.

Cembrowski, M.S., Bachman, J.L., Wang, L., Sugino, K., Shields, B.C., and Spruston, N. (2016). Spatial Gene-Expression Gradients Underlie Prominent Heterogeneity of CA1 Pyramidal Neurons. Neuron 89, 351–368.

Cenquizca, L.A., and Swanson, L.W. (2007). Spatial organization of direct hippocampal field CA1 axonal projections to the rest of the cerebral cortex. Brain Res. Rev. 56, 1–26.

Chen, F., Tillberg, P.W., and Boyden, E.S. (2015a). Expansion microscopy. Science 347, 543–548.

Chen, F., Wassie, A.T., Cote, A.J., Sinha, A., Alon, S., Asano, S., Daugharthy, E.R., Chang, J.-B., Marblestone, A., Church, G.M., Raj, A., Boyden, E.S., 2016. Nanoscale imaging of RNA with expansion microscopy. Nat Meth advance online publication. doi:10.1038/nmeth.3899

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

Choi, H.M.T., Beck, V.A., and Pierce, N.A. (2014). Next-Generation in Situ Hybridization Chain Reaction: Higher Gain, Lower Cost, Greater Durability. ACS Nano 8, 4284–4294.

Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Gephart, M.G.H., Barres, B.A., and Quake, S.R. (2015). A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. 112, 7285–7290.

Dong, H.-W., Swanson, L.W., Chen, L., Fanselow, M.S., and Toga, A.W. (2009). Genomic–anatomic evidence for distinct functional domains in hippocampal field CA1. Proc. Natl. Acad. Sci. 106, 11794–11799.

Fan, Y., Braut, S.A., Lin, Q., Singer, R.H., and Skoultchi, A.I. (2001). Determination of transgenic loci by expression FISH. Genomics 71, 66–69.

Fanselow, M.S., and Dong, H.-W. (2010). Are the dorsal and ventral hippocampus functionally distinct structures? Neuron 65, 7–19.

Femino, A.M., Fay, F.S., Fogarty, K., and Singer, R.H. (1998). Visualization of Single RNA Transcripts in Situ. Science 280, 585–590.

Habib, N., Li, Y., Heidenreich, M., Swiech, L., Trombetta, J.J., Zhang, F., Regev, A., 2016. Div-Seq: A single nucleus RNA-Seq method reveals dynamics of rare adult newborn neurons in the CNS. bioRxiv 045989. doi:10.1101/045989

Jung, M.W., Wiener, S.I., and McNaughton, B.L. (1994). Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. J. Neurosci. 14, 7347–7356.

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

Kishi, T., Tsumori, T., Yokota, S., and Yasui, Y. (2006). Topographical projection from the hippocampal formation to the amygdala: A combined anterograde and retrograde tracing study in the rat. J. Comp. Neurol. 496, 349–368.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 161, 1187–1201.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., et al. (2014). Highly Multiplexed Subcellular RNA Sequencing in Situ. Science 343, 1360–1363.

Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168–176.

Lubeck, E., and Cai, L. (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nat. Methods 9, 743–748.

Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. Nat. Methods 11, 360–361.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202–1214.

Madisen, L., Zwingman, T.A., Sunkin, S.M., Oh, S.W., Zariwala, H.A., Gu, H., Ng, L.L., Palmiter, R.D., Hawrylycz, M.J., Jones, A.R., et al. (2010). A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. Nat. Neurosci. 13, 133–140.

Madisen, L., Mao, T., Koch, H., Zhuo, J., Berenyi, A., Fujisawa, S., Hsu, Y.-W.A., Iii, A.J.G., Gu, X., Zanella, S., et al. (2012). A toolbox of Cre-dependent optogenetic transgenic mice for light-induced activation and silencing. Nat. Neurosci. 15, 793–802.

Miller, JA. Jason Nathanson, Daniel Franjic, Sungbo Shim, Rachel A. Dalley, Sheila Shapouri, Kimberly A. Smith, Susan M. Sunkin, Amy Bernard, Jeffrey L. Bennett, Chang-Kyu Lee, Michael J. Hawrylycz, Allan R. Jones, David G. Amaral, Nenad Sestan, Fred H. Gage, Ed S. Lein (2013). Conserved molecular signatures of neurogenesis in the hippocampal subgranular zone of rodents and primates. Development. 140(22): 4633–4644. doi: 10.1242/dev.097212

Muller, R., Stead, M., and Pach, J. (1996). The hippocampus as a cognitive graph. J. Gen. Physiol. 107, 663–694.

O'Keefe, J., and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. Brain Res. 34, 171–175.

Petrovich, G.D., Canteras, N.S., and Swanson, L.W. (2001). Combinatorial amygdalar inputs to hippocampal domains and hypothalamic behavior systems. Brain Res. Brain Res. Rev. 38, 247–289.

Pitkänen, A., Pikkarainen, M., Nurminen, N., and Ylinen, A. (2000). Reciprocal Connections between the Amygdala and the Hippocampal Formation, Perirhinal Cortex, and Postrhinal Cortex in Rat: A Review. Ann. N. Y. Acad. Sci. 911, 369–391.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biol 4, e309.

Risold, P.Y., and Swanson, L.W. (1996). Structural evidence for functional domains in the rat hippocampus. Science 272, 1484–1486.

Rust, M.J., Bates, M., and Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Meth 3, 793–796.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. Nat Biotech 33, 495–502. doi:10.1038/nbt.3192

Saunders, R.C., Rosene, D.L., and Van Hoesen, G.W. (1988). Comparison of the efferents of the amygdala and the hippocampal formation in the rhesus monkey: II. Reciprocal and non-reciprocal connections. J. Comp. Neurol. 271, 185–207.

Shah, S., Lubeck, E., Schwarzkopf, M., He, T., Greenbaum, A., Sohn, C. ho, Lignell, A., Choi, H.M.T., Gradinaru, V., Pierce, N.A., Cai, L., 2016. Single-molecule RNA detection at depth via hybridization chain reaction and tissue hydrogel embedding and clearing. Development dev.138560. doi:10.1242/dev.138560

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P.I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., Frisén, J., 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353, 78–82. doi:10.1126/science.aaf2403

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. advance online publication.

Thompson, C.L., Pathak, S.D., Jeromin, A., Ng, L.L., MacPherson, C.R., Mortrud, M.T., Cusick, A., Riley, Z.L., Sunkin, S.M., Bernard, A., et al. (2008). Genomic Anatomy of the Hippocampus. Neuron 60, 1010–1021.

Treweek, J.B., Chan, K.Y., Flytzanis, N.C., Yang, B., Deverman, B.E., Greenbaum, A., Lignell, A., Xiao, C., Cai, L., Ladinsky, M.S., et al. (2015). Whole-body tissue stabilization and selective extractions via tissue-hydrogel hybrids for high-resolution intact circuit mapping and phenotyping. Nat. Protoc. 10, 1860–1896.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 85.

Witter, M.P. (1993). Organization of the entorhinal—hippocampal system: A review of current anatomical data. Hippocampus 3, 33–44.

Witter, M.P., and Amaral, D.G. (1991). Entorhinal cortex of the monkey: V. Projections to the dentate gyrus, hippocampus, and subicular complex. J. Comp. Neurol. 307, 437–459.

Yang, B., Treweek, J.B., Kulkarni, R.P., Deverman, B.E., Chen, C.-K., Lubeck, E., Shah, S., Cai, L., and Gradinaru, V. (2014). Single-Cell Phenotyping within Transparent Intact Tissue through Whole-Body Clearing. Cell.

Yang SM, Alvarez DD, Schinder AF. (2015). Reliable Genetic Labeling of Adult-Born Dentate Granule Cells Using Ascl1 CreERT2 and Glast CreERT2 Murine Lines. J Neurosci. 35(46):15379-90.

Yi, F., Catudio-Garrett, E., Gábriel, R., Wilhelm, M., Erdelyi, F., Szabo, G., Deisseroth, K., and Lawrence, J. (2015). Hippocampal “cholinergic interneurons” visualized with the choline acetyltransferase promoter: anatomical distribution, intrinsic membrane properties, neurochemical characteristics, and capacity for cholinergic modulation. Front. Synaptic Neurosci. 7.

Zeisel, A., Manchado, A.B.M., Codeluppi, S., Lönnerberg, P., Manno, G.L., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science aaa1934.

3.10 Supplementary Experimental Procedures

Probe Design.

Genes were selected from the Allen Brain Atlas database. We identified genes that are heterogeneously expressed in coronal sections containing the hippocampus at Bregma coordinates -2.68 mm anterior. Using the ABA region definitions, we break down the voxels representing the ABA data in those brain sections into 160 distinct regions and average the expression values within each region. We selected 100 genes that had high variances across these distinct regions and that also had low-medium expression levels. These genes included transcription factors and signaling pathways components as well as ion channels and other functional genes. Lastly, we chose 25 genes from single cell RNA-seq data that were enriched in certain cell types. Briefly, the design criteria used were 1) constant regions of all spliced isoforms were identified, 2) Masked regions of UCSC genome were removed from possible probe design, 3) 35mer sequences were tiled 4nt apart, 4) sets of non-overlapping probes with tightest GC range around 55% were found, 5) probes were blasted for off-target hits. Any probe with an expected total off-target copy number of more than 5000 was dropped. Once all possible probes for every target gene was acquired, the probe set oligo-pool was optimized using the following criteria: 1) Expected # of off-target hits for entire probe pool was calculated, 2) probes were sequentially dropped from genes until any off-target gene was hit by no more than 6 probes from entire pool, 3) HCR adapters were added to designed probes and 10nt in either direction of the adapter junction was blasted and screened for off-target hits, 4) probe pools were searched for regions of 18mer complementary, 5) the probe sets for a given transcript was refined down to 24 probes by dropping probes in order of the expected number of off-target hits, 6) Cutting sites and hybridization specific primers were added to probes.

Probe Generation.

All oligoarray pools were purchased as 92k synthesis from Customarray Inc. Probes were amplified from array-synthesized oligo pool as previously described (36), with the following modifications: (i) a 35nt RNA-targeting sequence for in situ hybridization, (ii) a 35nt HCR initiator sequence designed to initiate one color of 5 possible HCR polymers, (iii) two hybridization specific flanking primer sequences to allow PCR amplification of the probe set and (iv) EcoRI (5'-GAATTC—3') and KpnI (5'-GGTACC-3') sites for cutting out flanking primers to reduce probe size. Ethanol precipitation was used to purify the final digested probes.

Brain extraction and sample mounting.

C57BL/6 with Ai6 Cre-reporter (uncrossed) (Jackson Labs, SN: 007906) female mice aged 50-80 days were anesthetized with isoflurane according to institute protocols (protocol #1701-14) (38). No randomization of mice was used and blinding was not necessary as the study was exploratory. Mice were perfused for 8 minutes with perfusion buffer (10U/ml heparin, 0.5% NaNO2 (w/v) in 0.1M PBS at 4C). Mice were then perfused with fresh 4% PFA\0.1M PBS buffer at 4C for 8 minutes. The mouse brain was dissected out of the skull and immediately placed in a 4% PFA buffer for 2 hours at room temperature under gentle mixing. The brain was then immersed in 4C 30% RNAse-free Sucrose (Amresco 0335-2.5KG)\1x PBS until the brain sank. After the brain sank, the brain was frozen in an dry ice\isopropanol bath in OCT media and stored at -80C. Fifteen micron sections were cut using a cryotome and immediately placed on an aminosilane modified coverslip.

Sample permeabilization, hybridization, and Imaging.

Brain sections mounted to coverslips were permeabilized in 4C 70% EtOH for 12-18 hours. Brains were further permeabilized by the addition of rnase-free 8% SDS (Ambion AM9822) for 10 minutes. Samples were rinsed to remove SDS, desiccated and a hybridization chamber (Grace Bio-Labs 621505) was adhered around the brain section. Samples were hybridized overnight at 37C with Split Color PGK1 Probes (Table S3) in Hybridization Buffer (2X SSC (Invitrogen 15557-036), 10% Formaldehyde (v/v) (Ambion AM9344), 10% Dextran Sulfate (Sigma D8906), 2mM Vanadyl Ribonucleoside Complex (VRC; NEB S1402S) in Ultrapure water (Invitrogen 10977-015)). Samples were washed in 30% Wash Buffer (WBT: 2X SSC, 30% Formaldehyde (v/v)] 10% Dextran Sulfate, 0.1% Triton-X 100 (Sigma X-100), 2mM VRC in Ultrapure water) for 30 minutes. While washing aliquoted HCR hairpins (Molecular Instruments Inc) were heated to 95C for 1.5 minutes and allowed to cool to RT for 30 minutes. HCR hairpins were diluted to a concentration of 120nM per hairpin in amplification buffer (2X SSC, 10% Dextran Sulfate) and added to washed tissue for 45 minutes. Following amplification, samples were washed in the same 30% WBT for at least 10 minutes to remove excess hairpins. Samples were stained with DAPI and submerged in pyranose oxidase antibleaching buffer (12). Sample port covers were closed with a glass coverslip or a transparent polycarbonate sheet to exclude oxygen.

Samples were imaged using a standard epifluorescence microscope (Nikon Ti Eclipse with custom built laser assembly) for the 125-gene experiment. Exposures

times were 200 ms for cy7 and alexa 488 channels and 100 ms for alexa 647, alexa 594, and cy3b channels. For the 249-gene experiment, a Yokogawa CSU-W1 spinning disk confocal unit attached to an Olympus IX-81 base was used for imaging. The exposure times were 500 ms for each channel. At this stage, intact and accessible mRNA should always appear in two channels. If the RNA was deemed to be intact, DAPI data was collected in this hybridization. Samples were digested with DNAse I (Roche 04716728001) for 4 hours at room temperature on the scope. Following DNAse I the sample was washed several times with 30% WBT and hybridized overnight with 70% Formamide HB and the experiment probes at 1 nM concentration per probe sequence at room temperature (Table S4 and S5). Samples were again washed and amplified as before. Barcode digits were developed by repeating this cycle with the appropriate probes for each hybridization. Fluorescent Nissl stain (ThermoFisher N-21480) was collected at the end of the experiment along with images of multispectral beads to aid chromatic aberation corrections.

Image Processing.

To remove the effects of chromatic aberration, the multispectral beads were first used to create geometric transforms to align all fluorescence channels. Next, the background illumination profile of every fluorescence channel was mapped using a morphological image opening with a large structuring element. These illumination profile maps were used to flatten the illumination in post-processing resulting in relatively uniform background intensity and preservation of the intensity profile of fluorescent points. The background signal was then subtracted using the imagej rolling ball background subtraction algorithm with a radius of 3 pixels. Finally, the calculated geometric transforms were applied to each channel respectively. The 150 pixel border region around the image was ignored in all analysis to avoid errors from edge effects of illumination.

Image Registration.

The processed images were then registered by first taking a maximum intensity projection along the z direction in each channel. All of the maximum projections of the channels of a single hybridization were then collapsed resulting in 4 composite images containing all the points in a particular round of hybridization. Each of these composite images of hybridization 1-3 were then cross-correlated individually with the composite image of hybridization 4 and the position of the maxima of the cross-correlation was used as the translation factor to align hybridizations 1-3 to

hybridization 4.

Cell Segmentation.

For cells in the cortex, the cells were segmented manually using the DAPI images taken in the first round of hybridization and the fluorescent nissl stain taken at the end of the experiment. Furthermore, the density of the point cloud surrounding a cell was taken into account when forming cell boundaries, especially in cells that did not stain with the nissl stain. For the hippocampus, the cells were segmented by first manually selecting the centroid in 3D of each DAPI signal of every cell. Transcripts were first assigned based on nearest centroids. These point clouds were then used to refine the centroid estimate and create a 3D voronoi tessellation with a 10% boundary-shrinking factor to eliminate ambiguous mRNA assignments from neighboring cells.

Barcode calling.

The potential mRNA signals were then found by LOG filtering the registered images and finding points of local maxima above a specified threshold value8. Once all potential points in all channels of all hybridizations were obtained, dots were matched to potential barcode partners in all other channels of all other hybridizations using a 1 pixel search radius to find symmetric nearest neighbors. Point combinations that constructed only a single barcode were immediately matched to the on-target barcode set. For points that matched to construct multiple barcodes, first the point sets were filtered by calculating the residual spatial distance of each potential barcode point set and only the point sets giving the minimum residuals were used to match to a barcode. If multiple barcodes were still possible, the point was matched to its closest on-target barcode with a hamming distance of 1. If multiple on target barcodes were still possible, then the point was dropped from the analysis as an ambiguous barcode. This procedure was repeated using each hybridization as a seed for barcode finding and only barcodes that were called similarly in at least 3 out of 4 rounds were used in the analysis. The number of each barcode was then counted in each of the assigned cell volumes and transcript numbers were assigned based on the number of on-target barcodes present in the cell volume. All image processing and image analysis code can be obtained upon request.

Clustering.

To cluster the dataset with 14,908 cells and 125 genes profiled, we first z-score normalized the data based on gene expression (Table S6). Once the single cell gene expression data is converted into z-scores, we compute a matrix of cell-to-cell correlations using Pearson correlation coefficients. Then hierarchical clustering with Ward linkage is performed on the cell-to-cell correlation data with cells in the center field of view. The cluster definitions are then propagated to the remaining cells using a random forest machine learning algorithm. To analyze the robustness of individual clusters, a random forest model was trained using varying subsets of the data and used to predict the cluster assignment of the remaining cells (22). A bootstrap analysis by dropping different sets of cells was performed in increments (Fig S5). To determine the effect of dropping out genes on the accuracy of the clustering analysis, we used a random forest decision tree to learn the cluster definition based on the 125 gene data. Then we ask the decision tree to re-compute the cluster assignment on cell-to-cell correlation matrices with fewer and fewer genes (Fig 3F, green line). Bootstrap resampling was also performed with this analysis (Fig 3F, bluelines). The PCA and tSNE analysis were performed using the same cell-to-cell z-scored Pearson correlation matrix. The cell-to-cell correlation in Fig 3E was calculated with increasing number of principal components dropped (have their eigenvalues set to zero). The cluster assignment accuracy is again computed through the random forest decision tree.

3.11 Supplementary Text

Error correction barcode design

Designing an error correction code to correct for k number of errors in a message of n length is analogous to packing as many spheres of radius k in a n dimensional cube. There are examples of "perfect codes" such as Golay and Hamming codes that can be as efficient as possible in this packing design. These perfect codes are important in digital communication because the word lengths are long, up to billions of letters for gigabytes of data, and many forms of errors can occur, including deletion and insertions. However, in the seqFISH experiments, as the code lengths are short, a perfect code correction system is not necessary, especially as the "correct" codes are already defined. One of the major source of error is deletions due to loss of a hybridization. Thus, it is possible to design simple correction schemes that are not completely efficient (i.e. obtain the tightest packing density for the n-spheres) but can achieve good error correction with just a few

extra rounds of hybridization. To design a barcode scheme that can tolerate loss of a single round of hybridization is akin to a problem where any n-dimensional hypercube is collapsed by 1 dimension to a n-1 dimensional hypercube without having any two points on the n-dimensional hypercube mapping to the same point. In order for this to be true, no two barcodes can be connected by a 1D line running parallel to any of the axes. There are many solutions to generate this 1 round loss tolerant code. A barcode generator (i, (i+j+k) mod 5,j,k) is used to generate the barcodes used in our experiment. This design can correct for loss of 1 hybridization for an arbitrarily long barcode sequence with minimal extra effort. For example, 7 rounds of hybridization with 5 colors can cover 57= 78,125 transcripts, more than the transcriptome, with 8 hybridizations the entire transcriptome can be coded with error correction using the barcoding system proposed. Another consideration in designing error-tolerant barcodes is that the mechanism of re-hybridzation should guide the robustness of error correction. In the merFISH implementation of seqFISH, null signal, or "0", along with "1" which is cy5 fluoroscence, is used to form a binary barcode. However, it is difficult to determine whether no signal is due to mis-hybridization or actual null signal. In our seqFISH implementation using positive signals as readouts during each round of hybridization reduces the need for error correction because false positive signal is unlikely to re-occur in the same position during another hybridization due to DNAse stripping between hybridizations. Thus implementation of seqFISH with 5 colors and 1 extra round of hybridization to error correct is both efficient and accurate, and allows imaging of a large tissue sections since imaging time is ultimately limiting in multiplexing experiments.

Optical Space for Barcodes in Cells

The theoretical upper limit for the number of barcodes that can be identified accurately within a cells primarily depends on the volume of the cell. As mRNA spots are diffraction limited, if a microscope is configured to have sub-diffraction limited pixel size, the ability to identify smFISH signal without any super-resolution would require no two mRNA signals to be immediately adjacent to each other in x, y or z dimension. We will call these minimum required voxels "coding voxels." The absolute upper limit of the number of transcripts that can be coded unambiguously without any super-resolution methods is solely a function of the number of coding voxels present in a cell. Assuming a diffraction limit of λ um and a resolution of z um in the z direction, there exists $V/((3\lambda)^2 z)$ coding voxels per cell, where V is the volume of the cell in microns. In our seqFISH method, we use 5 or more channels

to hold mRNA spots which would increase the total number of coding voxels by a multiplicative factor equal to the number of channels used for barcoding. Therefore,

$$\#B = (FV)/((3\lambda)^2 z) \tag{3.1}$$

where #B is the maximum number of unambiguous barcodes a cell can hold, and F is the number of channels used. As mammalian cells range from about 500 – 4000 microns in volume, these cells can accommodate roughly between 6100 – 49,000 barcodes assuming 5 fluorescence channels are being used, the diffraction limit is 0.3 um, and the z resolution is 0.5 μm. In principle, this calculation would provide the total number of perfectly discernible spots a cell can accommodate. In our actual experimental data, we have some amount of dropped barcodes due to ambiguity in barcode assignment due to spot overlaps. This is one of the main factors that reduces the efficiency of seqFISH as compared to single transcript detection (i.e. smFISH or smHCR). Expansion microscopy could further increase the number of coding voxels in a cell by the expansion factor leading to fewer drops and imaging of denser transcripts.

seqFISH clustering analysis vs single cell RNAseq analysis

As we do not sample the entire transcriptome with seqFISH, the dendrogram structure and cell clusters will be influenced by the genes chosen. This may be best illustrated by an analogy: at an international conference, you are surveying the attendees. You ask 100 questions, 80 of them about their profession and 20 on their country of origin. You might find that many people are biologists, and others are engineers. Similarly, there are many people from the US and others from elsewhere. If you cluster the data, you will find that because 80 out of the 100 questions are about their profession, the biologist vs engineer split will be the first split along the dendrogram, and the citizenship will split next. This is because two biologists from different countries will look more similar to each other (according the 100 questions asked), than an engineer. If the questions were 80 about citizenship, and 20 about profession, then the citizenship would be first split in the clustering. Now if you throw in more questions about gender in the survey, then you can get clusters that have both men and women engineers in the US. But if there is only 1 question about gender and the remaining 99 questions are about other things, then it is expected that gender would not factor importantly into the unsupervised clustering of the data.

The clustering dendrogram is determined by the distance between different clusters, and this distance is affected by how many marker genes are in the target

list. If we had 2 genes in our target list that differ in their expression between neuron vs non-neurons, and 10 genes that differ between DG vs CA1 neurons all with similar expression distributions, then the dendrogram is going to split the DG and CA1 neurons first because the distance between those clusters are larger than the neurons vs non-neurons distances (10 vs 2). Thus the dendrogram for the 100-200 genes experiment should not match completely the dendrogram for the RNAseq data, because the composition of marker genes in the 100 or 200 gene list is not the same proportions as the transcriptome.

In our analysis, all genes are weighted equally regardless of expression levels or "canonical importance." We believe this is the most direct and unbiased way to perform the analysis on our data. Because our data is 100-200 genes, but with higher accuracy per gene than single cell RNAseq for those 100-200 genes, it is going to be fundamentally different than single cell RNAseq data and our analysis method was selected to best match the nature of our data. Most single cell RNAseq analysis methods either select subsets of genes for clustering at each level or iteratively select genes with the highest variations to define cell types. Our analysis uses the full set of genes probed to detect combinatorial expression differences amongst cells. We have tried to implemented several analysis methods similar to ones used in single cell RNAseq, but obtained poor separation between cells clusters as compared to the analysis method herein presented. This is because we lose significant combinatorial information by ignoring genes when defining clusters. The accurate measurement of 100 genes can provide a great deal of explanatory power because the combinatorial expression pattern contains more information than just individual genes.

On the other hand, the main limitation of the 100-200 gene experiment is that it does not measure all of the genes. So the data will not cluster the same as RNAseq data, and does not identify all the big "cell types" in the top branches of the dendrogram. However, we can detect fine differences between cells. single cell RNAseq can be used for cataloging all the major cell types, while seqFISH can be used to focus on a specific region or "cell type" and investigate the spatial mRNA expression patterns between cells or fine differences within a "cell type".