Plus Ultra: Genome-wide Spatial Transcriptomics with RNA seqFISH+

Thesis by

Chee-Huat Linus Eng

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2021
Defended May 25, 2021

Chee-Huat Linus Eng
ORCID: 0000-0002-2521-9696

ACKNOWLEDGEMENTS

I love to play role-playing video games. In the beginning of the journey, you are weak. But through hard work, persistence, curiosity, and many other values (you name it!), you increase your level, skills, and abilities, and after multiple attempts, struggles, and not-giving-up, you beat the final boss! Of course, it would not be fun at all if it were not for the people you met during this entire journey and helped you through the many obstacles. It is the same in the PhD journey except not with magic, but with science.

First, I would like to thank my PhD advisor, Long Cai, for being a great mentor during my graduate school studies. You have been very motivating, encouraging, and patient during all the discussions we had about failed experiments throughout the years. You supported me when I was in doubt about which direction to take and provided helpful guidance like a lantern in the deep dark (you know why I said this!). Thank you for tolerating my jokes all this time- I am glad you are a professor with a great sense of humor! Aside from the scientific growth I gained from you, your kindness towards people is a precious value that I truly admire and appreciate, and greatly influences me to be a better human being in my life!

Next, I would like to acknowledge my thesis committee, which includes Dr. Rustem Ismagilov, Dr. Mitchell Guttman, and Dr. Matt Thomson, for their guidance and support these past years. Their insightful suggestions have been very helpful to my project development.

During my time at Caltech, I became greatly indebted to the past and current Cai lab members for their help and input with my projects. First, I would like to thank Yodai Takei, my “senpai” and good friend, who joined the lab a few months earlier than me, for being very accommodating during my first arrival to the Cai lab. Thank you for driving me around LA to eat tasty Japanese ramen despite that time you were a newbie in driving and made me worried so much at the highway with your driving skills (of course now your driving skills leveled up!). I enjoy discussing a lot of scientific ideas with you while walking to grab coffee on Lake Avenue even though it was extremely hot that time. Your scientific accomplishments have been very inspiring, and I look forward to more great things in your future! Oh yeah, I hope you get married soon! Next is Noushin Koulena, my “alliance” in the Cai lab who is very reliable and dedicated in helping me with my experiments for the longest time. Without you, I think I would have spent weeks to generate those probes! Aside from science, I especially enjoy the time chatting with you because of your sense of

humor, which makes the conversation lively and fun! I would also like to thank Jina Yun, the "snack provider" in the lab. Thank you for feeding me with those delicious snacks before COVID time. In addition, thank you for your hard work in conjugating those readout probes which require a lots of efforts and precision. There are more current and past members who I would like to extend my gratitude to such as Julian Thomassie, Sheel Shah, Mike Lawson, Nico Pierson, Yandong Zhang, Elsy Buitrago-Delgado, Lincoln Ombelets, and Simone Schindler for their help computationally or experimentally in my projects. I would not have finished the project smoothly without great help from these people.

I met a lot of precious friends during the journey at Caltech. I wish to thank Hyeong Chan, Wzy, Marcus, Richard, and Karena for being such good friends and for hanging out with me despite my attendance score not being high! I would also like to express my gratitude to the Malaysian friends at Caltech, including Voon, Ying Shi, Cynthia, Shu Fay, and Marcus for the gathering nights, cooking Malaysian food, and playing board games. I also enjoyed the time participating in the international students' events with you all! Remember the transferring of curry with my car, the worry that it could just splash out during my drive? I would like to thank my previous housemate and good friend Kai Chen, who has always been very inspiring to me because of his passion and dedication to science. I would also like to thank my current housemate and colleague, Carsten Tischbirek, for being caring and generous in sharing with me lots of the cakes and ice cream that he bought. By the way, he is my friend too! (requested by him to mention him as a friend).

Finally, I would love to thank my family back in Malaysia for their unconditional love and support towards my dream. Before I forget, I should thank myself too, for being fearless and persistent, striving for what people deemed impossible at the time.

ABSTRACT

Visualizing single cells and their organization in intact tissue is crucial to understanding their governing biological function. Even though single cell RNA sequencing has provided many insights into the heterogeneity and gene expression profiles across many tissue types, the dissociation process which loses the spatial information is hindering our deeper understanding of how these transcriptional distinct cell types are organized and interacting in their native tissue environment.

The thesis begins by giving a background on how single cell RNA sequencing has transformed biology and the emergence of spatial technology such as sequential fluorescence in situ hybridization (seqFISH). While spatial methods are useful for mapping the cell types identified from single cell RNA sequencing, the need for turning spatial technology such as seqFISH, which has high detection efficiency of the transcriptome with spatial information, into an *in situ* discovery tool is discussed as the scientific community's goal heads towards building spatial atlases for every human tissues and organs such as the brain.

While seqFISH has high detection efficiency, it is still limited in the number of genes capable of profiling at once. The major obstacle is the optical crowding problems when more RNA species are targeted and imaged using a fluorescence microscope. In Chapter 2, we first investigated, if the RNA molecules are instead captured on a coverslip and profiled with sequential barcoding strategy, the FISH-based method will reliably characterize the transcriptome when molecular crowding is not an issue.

Finally, in Chapter 3, we demonstrate the barcoding strategy to break through the molecular crowding limit of multiplexed FISH. From being able to profile hundreds to a thousand genes by various multiplexed FISH methods at that time in the field, we succeeded in profiling 10,000 genes by RNA seqFISH+, an evolved version of seqFISH, in various intact tissue sections, turning seqFISH+ into a spatial discovery technology with its genome-wide coverage and high detection efficiency. The work described in this part of the thesis is highlighted in Nature Method's Method of The Year 2020- Spatially-resolved Transcriptomic article.

PUBLISHED CONTENT AND CONTRIBUTIONS

Eng, Chee-Huat Linus, Sheel Shah, Julian Thomassie, and Long Cai. 2017. “Profiling the Transcriptome with RNA SPOTs.” *Nature Methods* 14 (12): 1153–55. https://doi.org/10.1038/nmeth.4500.

C.-H.L.E and L.C. conceived and designed experiments. C.-H.L.E. performed all the experiments. C.-H.L.E performed image analysis with the guidance of S.S. C.-H.L.E. and L.C. performed data analysis. C.-H.L.E participated in writing the manuscript.

Eng, Chee-Huat Linus, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, et al. 2019. “Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH+.” *Nature* 568 (7751): 235–39. https://doi.org/10.1038/s41586-019-1049-y.

C.-H.L.E. and L.C. conceived of the idea and designed experiments. C.-H.L.E. performed all the experiments. C.-H.L.E., M participated in data analysis. C.-H.L.E. participated in cell segmentation and generated the primary probes. C.-H.L.E involved in the validation the readout probes. C.-H.L.E. participated in writing the manuscript.

TABLE OF CONTENTS

Acknowledgements……...iii
Abstract ……iv
Published Content and Contributions…….....v
Table of Contents…… vi
Chapter I: Introduction 1
1.1 Transformative biology by single cell RNA sequencing 1
1.2 Emergence of spatial transcriptomics technologies 3
1.3 The need for spatially transcriptomics methods as discovery tools..... 5
1.4 Towards spatial cell atlas 5
1.5 References 6
Chapter II: Profiling the transcriptome by RNA SPOTs..... 18
2.1 Abstract..... 18
2.2 Introduction 18
2.3 Results 19
2.4 Discussion..... 22
2.5 Supplementary Data and Figures 23
2.6 Methods 37
2.7 References 44
Chapter III: Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+..... 47
3.1 Abstract..... 47
3.2 Introduction 47
3.3 Results..... 48
3.4 Discussion..... 55
3.5 Supplementary Data and Figures..... 56
3.6 Methods 73
3.7 References 87

Chapter 1

Introduction

1.1 Transformative biology by single cell RNA sequencing

In the past decade, bulk RNA sequencing has been used to study gene expression in various tissue samples at population level[1]. Despite the great resources provided by bulk measurements, the lack of single cell resolution is limiting our understanding on biological problems. It is the advent of single-cell RNA sequencing (scRNA-seq) by multiple groups which enabled the studies of the transcriptome to dissect single cell heterogeneity and discover unexpected biological insights relative to classical transcriptomic profiling methods [2–7]. ScRNA-seq was first applied to a four-cell stage blastomere by Tang et. al. in which the sequencing library preparations were performed manually in individual tubes, hindering the single cells that could be studied at once. Subsequent years, multiple scRNA-seq technologies were developed to assay many cells at once including single cell tagged reverse transcription sequencing (STRT-seq)[8], cell expression by linear amplification and sequencing (CEL-seq) [9] , MARS-seq [10], SMART-seq [11] , and many others. The invention of droplet based sequencing such as Drop-seq and In-Drop further outperformed the commonly used plate-based sequencing platform Fluidigm C1 due to their capability to assay tens of thousands of cells as well as lower cost [12] [13].

The unbiased genome-wide profiling and high number of single cells assayed in scRNAseq have allowed researchers to make biological discoveries in all fields such as in neuroscience and developmental biology. For example, Tasic et al had used scRNAseq to characterize the cortical cells in the primary visual cortex into distinct transcriptional types including 23 GABAergic, 19 glutamatergic, and 7 non-neuronal cells. The author further confirmed that the transcriptomic states of these clusters could be associated with their electrophysiological and axon projection properties [14]. Later years, the author applied scRNAseq to dissect the similarity and differences of the 133 transcriptomic cell types between anterior lateral motor cortex and primary visual cortex [15]. In developmental biology, multiple studies have used scRNA-seq to understand the developmental trajectories of cells and investigate the transcriptional states of the cells and its descendants during development and regeneration. For example, the development of sci-RNA_seq (Single cell Combinatorial Indexing RNA sequencing) to profile almost 50,000 cells from the nematode *Caenorhabditis elegans* at the L2 stage revealed some rare neuronal cell types which only exists as few as one or two cells at that stage. Together with the defined 27 cell types, these data serve as a powerful resource to the nematode community[16]. On the other hand, Wagner et. al. used InDrop scRNAseq to sequence more than 90,000 cells following the developmental stages of zebrafish embryos. Their results uncover the progression of cell-state landscape across axis patterning, formation of germ cell layers, and organogenesis. The authors further developed TracerSeq which barcodes cell lineages during development. They found that the history of cell lineage does not always follow or reflect the cell-state graph topology based on the profiled

transcriptome[17]. Similar studies by Farrell et. al. which also applied scRNAseq to >38,000 cells during early zebrafish embryogenesis revealed some interesting findings. Other than characterizing these cells into distinct transcriptional clusters which represent the developmental stages of the embryo, they identified modules of coexpressed genes by these cells across the developmental time. In addition, they revealed that at some developmental branches, multilineage priming exists in some of the cells based on co-expression of multiple genes characteristic of downstream cell fates [18]. Many more scRNA-seq studies applied on other model organisms such as planarian *Schmidtea mediterranea* [19] , the western claw-toed *Xenopus tropicalis* [20] , and the house mouse *Mus musculus* [21–24] all served as valuable resources to the wide community. It is worthwhile to mention that in the studies conducted by Cao et. al., the number of cells being assayed at once in scRNAseq is particularly impressive with more than 2 millions single cells from 61 embryos which span across embryonic developmental stages of 9.5 to 13.5 are sequenced in a single experiment [25] . Their results demonstrate the global view of mouse organogenesis based on the transcriptome profiled, as well as the dynamics of the gene expression within cell types and trajectories across this critical developmental process. Finally, scRNAseq is also highly applied on human samples to comparatively study the transcriptome between healthy and diseased tissues[26–30]. In particular, scRNAseq has also been applied to study the coronavirus disease (COVID-19) pandemic to identify the immune response as well as the impact on various tissue organs of post-infection which helps scientists to understand the disease better[31–35].

Indeed, scRNAseq has transformed how biological problems are studied nowadays, transitioning from small numbers of cells to millions of cells with the unbiased discovery power of genome-wide transcriptome profiling. However, it is also known that scRNAseq also suffers from multiple limitations. All scRNAseq libraries preparation requires the dissociation of tissue into single cell suspension. This created a few problems. The most precious spatial information of each single cell within its intact tissue is lost. Moreover, it is known that transcriptome-wide changes can be induced by the dissociation process. Lastly, different tissues have different dissociation efficiency and that certain cell populations could be lost in detection by scRNA-seq. This particular weakness has motivated scientists to instead, isolate the nuclei of single cells for transcripts instead of isolating the whole cytoplasmic RNA[36,37]. Moreover, inefficient reverse transcription, amplification bias of library preparation and high dropout rate of scRNAseq caused the detection of lowly expressed transcripts challenging, rendering a low detection efficiency of 1-20% of the transcripts per single cell depending on the platform used. The detection efficiency further drops when shallow depth of sequencing is performed in exchange for a higher cell number profiled[38]. In fact, all these problems can be potentially solved by the emergence of spatially resolved transcriptomics profiling methods in recent years.

1.2 The emergence of spatial transcriptomic technologies

It is crucial to understand single cells within their spatial organization in intact tissues as neighboring cell-cell interactions can govern the cell fate decision. For instance, classical studies showed how the "organizer" cells organize the dorsal ectoderm into a neural tube and the mesoderm into anterior-posterior axis through series of induction[39,40]. Techniques such as *in situ* hybridization (ISH) with colorimetric readout applied on mouse or human brain slices, one gene at a time, by the Allen Brain Institute to create a reference map atlas has been useful to the neuroscience community over the past years[41,42]. However, ISH does not provide single cell resolution and the measurement is not quantitative. Current gold standard measurement in absolute transcripts quantitation method is still the single molecule fluorescence in situ hybridization (smFISH) technology developed by Raj et. al. which uses multiple fluorophore-labeled short oligonucleotides designed to bind to the same targeted transcript, yielding diffraction-limited spots when imaged under a fluorescence microscope. By counting these diffraction-limited spots which each dot represent a single RNA transcript, it allows the quantification of gene expression with near 100% detection efficiency[43].

The limitation of smFISH technology is the number of fluorescence channels one can use in microscopy, generally 4-5 fluorescence channels are available, which limits the scalability of the number of genes one can profile at once. In 2012, Lubeck and Cai scaled up the measurement of smFISH to 32 genes through super-resolution imaging and combinatorial spatial and spectral barcoding. However, super-resolution is difficult to perform on a thicker sample, as well as the gene throughput is limited by the number of fluorophores available to perform super-resolution imaging[44]. This motivates the development of sequential FISH (seqFISH) which fundamentally works by generating temporal barcodes on each transcript by sequential rounds of FISH hybridization. They took advantage of the fact that since transcripts are fixed in cells, and the corresponding fluorescent dots should remain in place for multiple rounds of hybridization and by aligning these spots, one can identify the unique fluorophore barcode designed for each gene. The advantage of seqFISH is that the number of barcodes scales as F^N , which F represents the number of fluorophores and N represents the number of sequential hybridization rounds performed. They further demonstrated that by introducing a redundant round of hybridization, one can decode the barcode assigned to each gene more robustly[45]. It is since this exciting technology development, that subsequent years, multiple spatial methods are developed. In particular, multiplex error robust (MERFISH) expanded the error correction scheme in the original seqFISH demonstration by using a Hamming distance of 4 based barcodes in the RNA detection in cell cultures[46]. Despite the highly quantitative power of smFISH, it is known to suffer from low signal to noise ratio. It has then been shown by combining tissue clearing technologies such as CLARITY and PACT (Passive CLARITY)[47,48] with FISH can improve the signal-to-noise ratio in tissue sections. Even so, smFISH signals could be further amplified by branched DNA or hybridization chain reaction (HCR)[49]. For example, Shah et. al. performed up to 249 genes HCR-seqFISH measurement in the tissue sections which robustly characterizes the dentate gyrus spatial organization into distinct transcriptional clusters without any tissue clearing due to the benefits of ~20-fold signal amplification[50]. A hybrid method which combines FISH and in situ sequencing, STARmap (spatially-resolved transcript amplicon readout mapping)

was developed to detect up to 1000 genes in the mouse primary visual cortex tissue section. STARmap begins by performing FISH using a pair of DNA probes which can be ligated when hybridized in close proximity. Once ligated, enzymatic amplification by phi29 DNA polymerase generates DNA nanoballs around the transcripts, the sample is then embedded in an acrylamide hydrogel to anchor the amplicons, followed by tissue digestion to improve the tissue transparency and signal-to-noise ratio. Finally the tissue-hydrogel hybrid can then be sequenced out *in situ* for the barcodes assigned for each gene[51]. Around the same time as multiplexed smFISH technologies are evolving, in situ sequencing of RNA in single cells are also actively developing. To give an instance, fluorescent in situ RNA sequencing (FISSEQ) directly reverse transcribes the mRNA in intact cells, followed by amplicons generations through rolling circle amplification, and finally sequence-by-ligation to identify the RNA sequences. Despite being an attractive concept and method, FISSEQ suffers from very low detection efficiency (< 0.01% reported) , likely due to the inefficient reverse transcription step as well as difficulties to ligate the complementary DNA (cDNA) *in situ*[52]. Targeted in situ sequencing which involves reverse transcription step, followed by FISH, improves the detection efficiency but it is still lower than direct FISH to RNA as conventional smFISH does[53].

Imaging-based spatially resolved multiplexed FISH measurements are highly quantitative with high detection efficiency but have been limited to the hundreds of genes up to 1,000 genes. On the other hand, slide-based spatial sequencing technologies such as Spatial Transcriptomics (ST) technology[54] captures spatial information by using spatially barcoded and oligo(dT) probes printed as microarray spots on the surface of glass slides. Then, cryosectioned tissue slices are placed on top and digested away enzymatically to release the mRNA, allowing the mRNA molecules to be captured by the surface probes. Current optimized version in 10x Genomics in Visium platform contains 5000 barcoded spots which are 55um in diameter yielding an average resolution of 1-10 cells per spot. In order to improve the spatial resolution, Slide-seq and HDST were developed. Differ from ST technology, both Slide-seq and HDST packed a monolayer of beads on a rubber-coated glass coverslip with the former using 10um beads and the latter using 2um beads. Since these beads are assembled with random barcodes, the spatial identity of the beads required either sequencing by SOLiD chemistry for Slide-seq or sequential hybridization for HDST to decode, followed by matching the spatial barcodes to the sequenced amplicons[55,56]. Despite the genome-wide profiling and high throughput measurement, these slide-based spatial sequencing technologies have a few shortcomings. First of all, despite some of the methods approaching single-cell resolution, it is challenging to define a single cell boundary in the sample and hence impossible to study subcellular localization of transcripts. Second, the lateral diffusion of transcripts during dissociation of the tissue likely will cause the intermixing of transcripts leaked from one cell to another. Lastly, the capture and detection efficiency of transcripts are much lower than multiplexed FISH. For example, the optimized version of Slide-seq V2 reported <50% of any conventional droplet based sequencing of detection efficiency which has about 1-5% depending on sequencing depth [57].

1.3 The need for spatial transcriptomics methods as discovery tools

Each of the spatial technologies has its strength and weakness. For example, slide-based spatial sequencing enables large numbers of cells to be profiled genome-wide, at the cost of low detection efficiency and not at single-cell resolution. On the other hand, targeted approaches such as seqFISH have very high detection efficiency, however this high detection efficiency is hindered by the molecular crowding as the number of genes profiled increases. Till now, all spatial methods have been used for spatial mapping of transcriptional clusters identified from scRNAseq, demonstrating the complementary of both technologies[50,51,58–60]. However, given cells function and interact with their neighbors, being able to study these neighboring cell interactions in addition to their spatial organization is crucial in dissecting the complex biological problems. If highly multiplexed FISH technology can only profile up to 1,000 genes with high efficiency, it is difficult to serve as an *in situ* discovery tool with such gene coverage. The major bottleneck of scaling up multiplexed FISH is because of the molecular crowding as the number of genes targeted increases, rendering decoding impossible. In fact, all spatial technologies can be benefited from expansion microscopy which the hydrogel-embedded sample is homogenized through enzyme digestion, followed by low osmolarity solution to physically expand the sample, thus pulling apart the cross linked RNA molecules, rendering super-resolution[61,62]. However, expanded samples require more technical attention to handle such as sample staging, prevention of molecules movement during imaging, as well as dramatically increases the imaging time as the sample volume increases. Hence, there is a need to scale up the number of genes detected by seqFISH to tens of thousands of genes efficiently with reasonable imaging time.

1.4 Towards spatial cell atlas

Over the past few years, the single cell spatial transcriptomics field has evolved so quickly as seen by more and more studies applying spatial technologies to study complex biological problems. In 2020, *Nature Method* has recognized spatially resolved transcriptomics as the method of the year, in which the thesis work described here is highlighted[63]. With huge projects like the Human Cell Atlas[64] which requires loads of collaborative effort to define all human cell types, spatial information becomes obviously indispensable. A comprehensive reference atlas without spatial information is not an atlas, and with the improvement of these spatial technologies, one should expect more and more spatial atlas covering tens of millions of cells with genome-wide measurement emerging in the near future.

Just like how scRNAseq changes biology, I believe spatial technology will transform biology for the next decades.

1.5 References

1. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
2. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 3010–3014 (1992).
3. Brady, G., Barbara, M. & Iscove, N. N. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. *Methods Mol. Cell. Biol.* **2**, 17–25 (1990).
4. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
5. Tang, F. *et al.* Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468–478 (2010).
6. Tang, F. *et al.* Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS One* **6**, e21208 (2011).
7. Brouilette, S. *et al.* A simple and novel method for RNA-seq library preparation of single cell cDNA analysis by hyperactive Tn5 transposase. *Dev. Dyn.* **241**, 1584–1590 (2012).
8. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
9. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
10. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).

11. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

12. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

13. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

14. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).

15. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).

16. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

17. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

18. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, (2018).

19. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science* **360**, (2018).

20. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).

21. Consortium, T. T. M. *et al.* A Single Cell Transcriptomic Atlas Characterizes Aging Tissues in the Mouse. doi:10.1101/661728.

22. Consortium, T. T. M. *et al.* Single-cell transcriptomics of 20 mouse organs creates a

Tabula Muris. *Nature* vol. 562 367–372 (2018).

23. He, P. *et al.* The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760–767 (2020).
24. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
25. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
26. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.* **10**, 4706 (2019).
27. He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol.* **21**, 294 (2020).
28. Kamies, R. & Martinez-Jimenez, C. P. Advances of single-cell genomics and epigenomics in human disease: where are we now? *Mamm. Genome* **31**, 170–180 (2020).
29. Nomura, S. Single-cell genomics to understand disease pathogenesis. *J. Hum. Genet.* **66**, 75–84 (2021).
30. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, (2017).
31. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
32. Zhang, J.-Y. *et al.* Single-cell landscape of immunological responses in patients with COVID-19. *Nat. Immunol.* **21**, 1107–1118 (2020).
33. Singh, M., Bansal, V. & Feschotte, C. A Single-Cell RNA Expression Map of Human

Coronavirus Entry Factors. *Cell Rep.* **32**, 108175 (2020).

34. Ren, X. *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895–1913.e19 (2021).
35. Melms, J. C. *et al.* A molecular single-cell lung atlas of lethal COVID-19. *Nature* (2021) doi:10.1038/s41586-021-03569-1.
36. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 130 (2020).
37. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
38. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
39. Zimmerman, L. B., De Jesús-Escobar, J. M. & Harland, R. M. The Spemann organizer signal noggin binds and inactivates bone morphogenetic protein 4. *Cell* **86**, 599–606 (1996).
40. Spemann, H. & Mangold, H. Induction of embryonic primordia by implantation of organizers from a different species. 1923. *Int. J. Dev. Biol.* **45**, 13–38 (2003).
41. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
42. Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
43. 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).

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

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

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

47. Yang, B. *et al.* Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell* **158**, 945–958 (2014).

48. Chung, K. & Deisseroth, K. CLARITY for mapping the nervous system. *Nat. Methods* **10**, 508–513 (2013).

49. 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).

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

51. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).

52. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).

53. Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).

54. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).

55. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).

56. Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat. Methods* **16**, 987–990 (2019).

57. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).

58. Lignell, A., Kerosuo, L., Streichan, S. J., Cai, L. & Bronner, M. E. Identification of a neural crest stem cell niche by Spatial Genomic Analysis. *Nat. Commun.* **8**, 1830 (2017).

59. Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, (2018).

60. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).

61. Chen, F., Tillberg, P. W. & Boyden, E. S. Optical imaging. Expansion microscopy. *Science* **347**, 543–548 (2015).

62. Chen, F. *et al.* Nanoscale imaging of RNA with expansion microscopy. *Nat. Methods* **13**, 679–684 (2016).

63. Marx, V. Method of the Year: spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).

64. Aviv, R. *et al.* The Human Cell Atlas. *eLife; Cambridge* **6**, (2017).

Chapter 2

PROFILING THE TRANSCRIPTOME BY RNA SPOTS

Eng, Chee-Huat Linus, Sheel Shah, Julian Thomassie, and Long Cai. 2017. "Profiling the Transcriptome with RNA SPOTs." *Nature Methods* 14 (12): 1153–55. https://doi.org/10.1038/nmeth.4500.

2.1 Abstract

Single molecule FISH (smFISH) has been the gold standard in quantifying individual transcripts abundances. Here, we demonstrate the scaling up of smFISH to the transcriptome level by profiling of 10,212 different mRNAs from mouse fibroblast and embryonic stem cells. This method, called RNA SPOTs (Sequential Probing of Targets), provides an accurate and low-cost alternative to sequencing in profiling transcriptomes.

2.2 Introduction

RNA sequencing (RNAseq)[1,2] has been a powerful method to quantify RNAs in a diverse range of biological samples. While RNAseq has replaced microarrays as the de-rigueur method for genomics studies because of higher sensitivities and dynamic range, reverse transcription and other steps needed to convert RNA to cDNA to sequencing libraries can introduce biases in the quantitation of mRNAs. Moreover, sequencing the RNAs at nucleotide level is not necessary for counting the abundances of transcripts. Single molecule fluorescence in situ hybridization (smFISH)[3,4], which directly hybridizes DNA oligonucleotide probes to transcripts in cells, is highly sensitive and accurate in quantitating mRNA abundances.

Here, we demonstrate transcriptome level profiling of mRNAs with single molecule sensitivity and high accuracy using a method based on sequential FISH (seqFISH)[5]. We had shown that seqFISH can be applied to image hundreds of transcripts in cells and tissues[6], image dynamics of chromosomes[7] and allow lineage tracking with single cell resolution[8]. However, the major limitation of seqFISH is that optical diffraction limit prevents many mRNAs from being resolved simultaneously in single cells. In principle, super-resolution microscopy[9] and expansion microscopy[10] can resolve the optical density issue *in situ*. However, many applications quantify mRNAs that have been extracted

from cells and tissues. In these cases, capturing transcripts onto an oligonucleotide dT surface and adjusting the dilution factors can easily remove the optical crowding problems and allow the transcriptome to be decoded by seqFISH.

2.3 Results

To distinguish this *in vitro* application from the *in situ* seqFISH experiments, we call this approach RNA SPOTs (Sequential Probing Of Targets). Extracted mRNAs were first captured on a Locked Nucleic Acid(LNA) poly(dT) functionalized coverslip (Fig 1a) and then hybridized with a pool of 323,156 primary probes targeting the coding regions of 10,212 mRNAs with 28 to 32 probes each gene (**Figure 1a-b**,**Supplementary Table 1 and Online Methods)**. To barcode the 10,212 genes with sequential hybridization, we used a 12 "pseudo-color" based scheme such that 4 rounds of barcoding are sufficient to cover the transcriptome (12^4=20,736) (**Supplementary Table 2)**, with an additional round of error correction to compensate for one drop in any round of barcoding[6] (**Fig 1c-d**). The pseudo-colors design shortens the number of barcoding rounds, which reduces the errors in reading out barcodes.

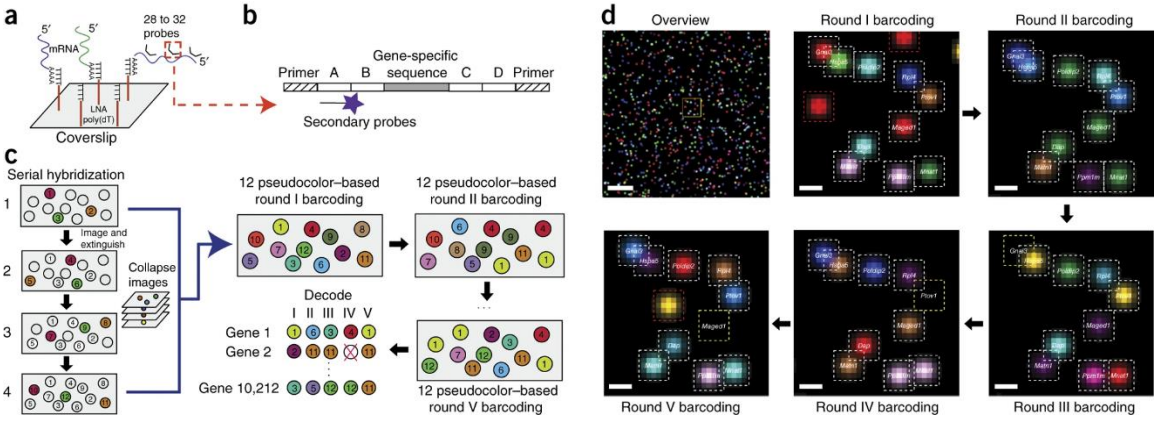


Figure 1. RNA SPOTs profiles 10,212 mRNAs *in vitro*. (**a**) mRNA is captured on a locked nucleic acid (LNA) poly(dT)-functionalized coverslip, and gene-specific primary probes (323,156 total) are then hybridized against the 10,212 targeted mRNAs. Each gene is targeted by 28–32 primary probes. (**b**) Each 149-nt primary probe includes a 25-nt gene-specific sequence complementary to the mRNA, four 20-nt barcodes (A,B,C, and D)—each encoding one of 12 'pseudocolors') which are read out by fluorescent secondary readout probes, single T-nucleotide spacers between readout and gene-specific regions, and two 20-nt PCR primer binding sites. Note that probes for each gene are divided into subsets in which sites A, B, C, and D may correspond to round I, II, III and IV or V, I, II, III, etc. to ensure the gene is measured in every round. (**c**) Each of five barcoding rounds is based on

Figure 1. (continued from above) 12 pseudocolors, which are read out by four serial hybridizations. In each serial hybridization, three readout probes conjugated to Alexa 647, Alexa 594, or Cy3b are hybridized to the primary probes, imaged, and extinguished. Images from four serial hybridizations are then collapsed into a single composite 12-pseudocolor image representing one round of barcoding. Sets of four serial hybridizations are repeated for five barcoding rounds (I to V) for a total of 20 hybridizations. This corresponds to $12^4 = 20{,}736$ codes, with an extra round of barcoding to correct for mishybridizations. (**d**) Digitized composite images based on actual experiments to decode 10,212 distinct mRNA. White dashed squares represent correctly identified barcodes; red dashed squares represent false positives; yellow dashed squares represent barcodes identified despite mishybridization in one round of hybridization. Scale bars: overview, 10 μm; rounds I to V barcoding, 1 μm.

To implement the pseudo-color scheme, we designed the primary probes to contain a 25-nt RNA binding sequence, as well as 4 overhang sites[11] that can be bound by dye-labeled readout oligos (**Figure 1b**). Each site has 12 possible sequences corresponding to the 12 pseudo-colors. To readout the 12 pseudo-colors, three of the readout oligos were hybridized at a time, imaged in the Cy3b, Alexa 594, and Alexa 647 fluorescence channels, and repeated 4 times to iterate through all 12 readout sequences, with disulfide cleavage[12,13] in between the hybridizations to remove the fluorophores (**Supplementary Fig 1 and 2**).

With 5 rounds of barcoding using the 12 pseudo-color readouts scheme, a total of 60 readout oligos were used to decode the 10,212 genes targeted (**Supplementary Fig 1-4 and Supplementary Table 3**). Each set of primary probes that target a specific gene contains 5 unique readout sequences that are spread out over the overhang sites (Fig 1b). A total of 20 rounds of hybridization, or 5 barcoding round each containing 4 serial hybridization **(Supplementary Fig 1)** were performed. A common sequence is present in all primary probes and targeted by an oligo labeled with Alexa 488 to serve as an alignment marker through all 20 rounds of hybridization (**Supplementary Fig 2b**). Each four rounds of serial hybridization were collapsed onto a single image with 12 pseudo colors (Fig 1c). The barcodes were determined from aligning five barcoding rounds of the pseudo-color images. The switching and rehybridization time is fast, with the overall speed limited by imaging speed. Typically, 100-200 fields of view containing more than 10^6 mRNAs can be imaged with 20 rounds of serial hybridization in a 14-hour period through an automated fluidics system. We use Spots per Millions (SPM) to normalize spots counts for individual genes between experiments (**Supplementary Table 4 and 5**).

The false positive rates of detection is low, with 0.72 ± 1.9 SPM per barcode, as determined by the remaining 238,620 off-target barcodes.

To determine the accuracy of the transcriptome level measurements, we compare the decoded RNA SPOTs data with RNAseq data in mouse fibroblasts (NIH/3T3) and mouse embryonic stem cells (mESCs), and found that they correlated with R=0.86 and R=0.9 respectively (**Fig 2a,b** and **Supplementary Fig 5** and **Supplementary Table 4)**. Between two replicates of RNA SPOTs in fibroblasts, the results agree with R=0.94, indicating that RNA SPOTs is a highly robust and reproducible measurement method (**Fig 2c, Supplementary Fig 5- 7)**. Finally, RNA SPOTs correlated with the gold standard smFISH quantitation with a correlation of R=0.86 in mESCs (24 genes)[14] and R= 0.88 in fibroblasts (7 genes) (**Fig 2d** and **Supplementary Fig 8**).

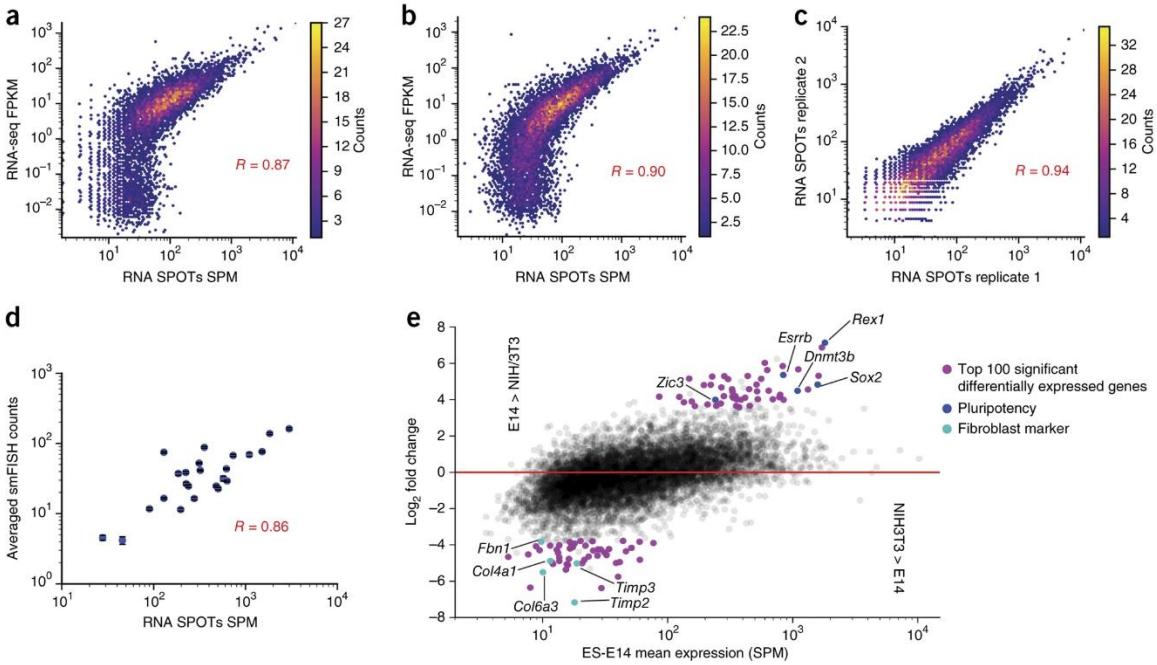


Figure 2. RNA SPOTs is highly accurate and efficient. (**a**) Transcriptomic profiling of mouse NIH/3T3 cells by RNA SPOTs correlates strongly with measurement from RNA-seq. SPM (spots per million) normalizes the number of each decoded mRNA spots (*n* = 581,772) by the total number of spots. FPKM, fragments per kilobase per million reads. (**b**) RNA SPOTs profiling of mouse ES-E14 cell line strongly agrees with RNA-seq measurement. (*n* = 1,688,747 spots). (**c**) Comparison of two RNA SPOTs replicates profiling NIH/3T3 cells illustrates that the method is highly reproducible (n_1 = 581,772 spots; n_2= 453,679 spots). (**d**) Comparison of averaged smFISH copy numbers of 24 genes in ES-E14 cells with RNA SPOTs SPM verifies the high-accuracy measurement of SPOTs.

Figure 2. (continued from above) Error bars represent s.e.m. across different measurements in single cells. (**e**) Differential gene expression between NIH/3T3 and ES-E14 cells. *P* values smaller than 0.05 as determined from two-tailed student *t*-test and $\log_2$ fold change greater and less than ±2, respectively, are used as a threshold for significance. Magenta dots represent top 50 upregulated and top 50 downregulated genes between the two cell lines. Blue dots represent the well-known genes involved in pluripotency. Cyan dots represent the genes involved in maintenance of extracellular matrix.

Comparing genes that were differentially expressed in fibroblasts versus mESCs, we observed the same trend as those detected by RNAseq. For example, pluripotency factors such as *Rex1* (also known as *Zfp42*), *Esrrb* and *Sox2* are highly expressed in mESCs but not expressed in fibroblasts as determined by RNA SPOTs. Similarly, genes involved in extracellular matrix maintenance, such as *Timp2*, *Timp3* and Collagen related genes such as *Col4a1*, *Col6a3* are up-regulated in fibroblast cells compared to mESCs (**Fig 2e** and **Supplementary Table 6**).

2.4 Discussion

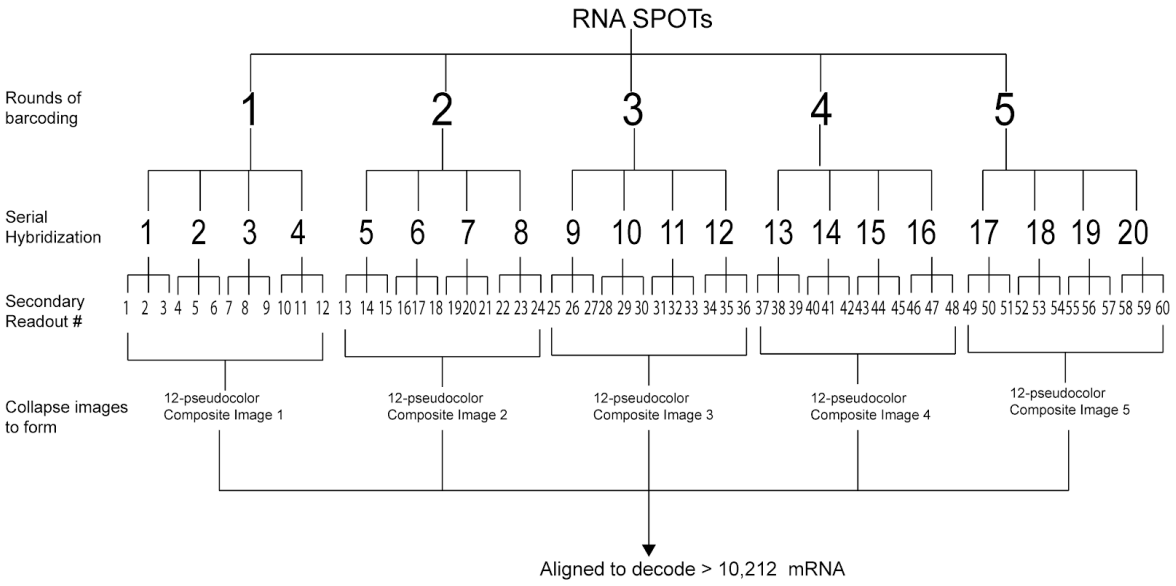
Another advantage of RNA SPOTs compared to RNAseq is that specific sets of genes can be profiled selectively. In this fashion, ribosomal RNA and highly expressed housekeeping genes can be avoided simply by eliminating those probes from the gene set. As each dot detected in our assay corresponds to a single mRNA, RNA SPOTs is more efficient in term of imaging compared to RNAseq, where many sequencing reads are needed to determine the abundance of a transcript. The current barcoding space is sufficient for the entire transcriptome, and noncoding RNAs and other RNAs without polyA tails can be captured in hydrogels (**Supplementary Fig 9**) rather than with dT oligos.

SPOTs is a significant improvement over existing Nanostrings technology[15] because of the genome level coverage and the higher specificity due to the larger number of probes used per gene. By incorporating amplification methods such as HCR [6,16], SPOTs signal can potentially allow faster imaging with air objectives and higher throughput comparable to RNAseq.

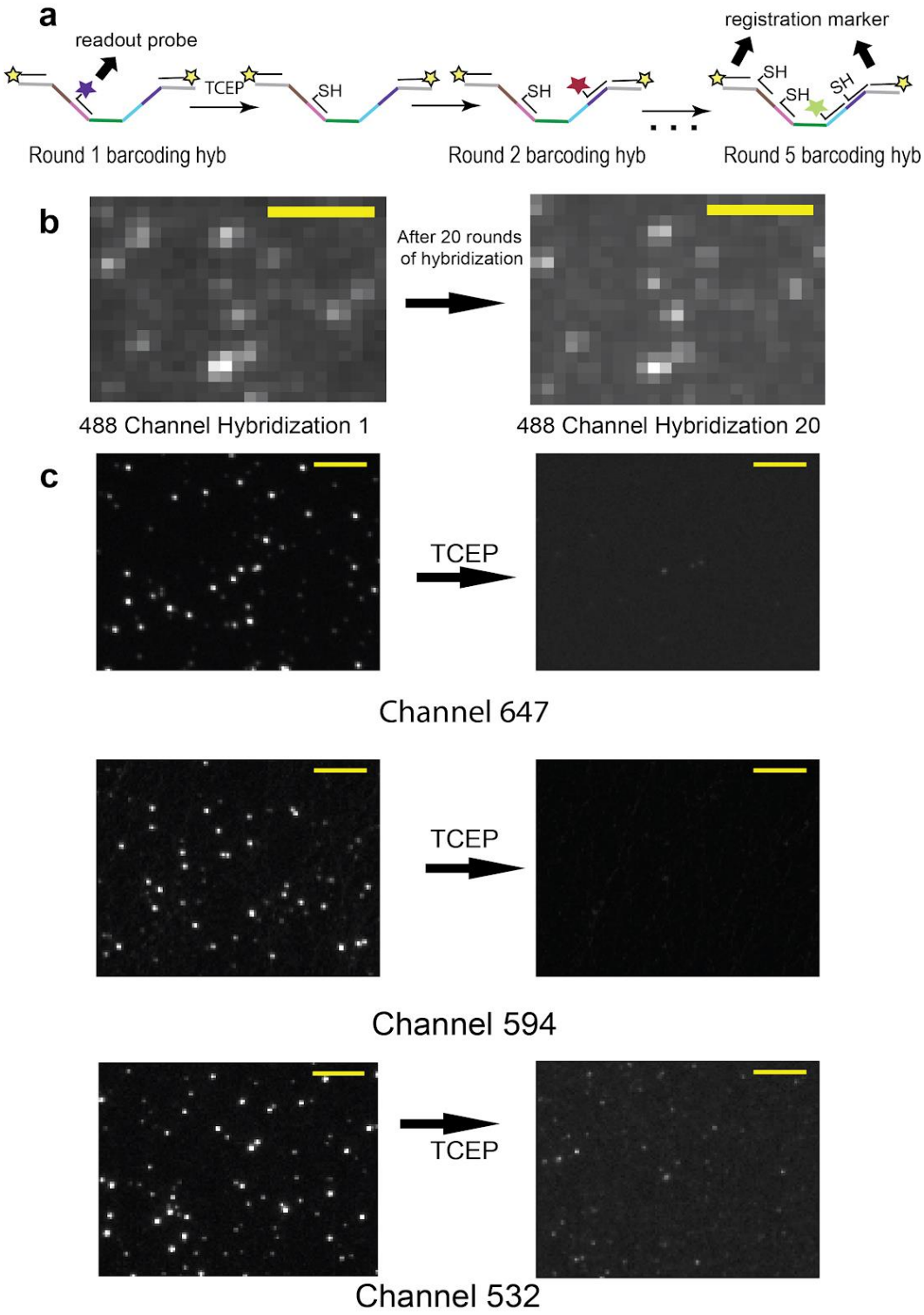
RNA SPOTs can be scaled down to single cell in combination with microfluidics tools to trap and lyse cells[17] or with split-pool molecular indexing methods[18]. While SPOTs cannot be used to discover new RNA sequences, identification of new cell types only requires quantifying the combinatorial expression patterns of genes. Thus, there is no need to re-sequence the mRNAs at the nucleotide level just to count their abundances. With targeted RNA SPOTs, we can choose to probe only for the 2000

transcription factors[19] or 1000 landmark informative genes[20] in single cells, instead of profiling the transcriptome, to capture the essential information in cells and to increase the number of cells sampled. As cost of sequencing is a major limiting factor in many genomics experiments, SPOTs enable an accurate and low-cost alternative to sequencing with many further applications beyond RNA to DNA and proteins.

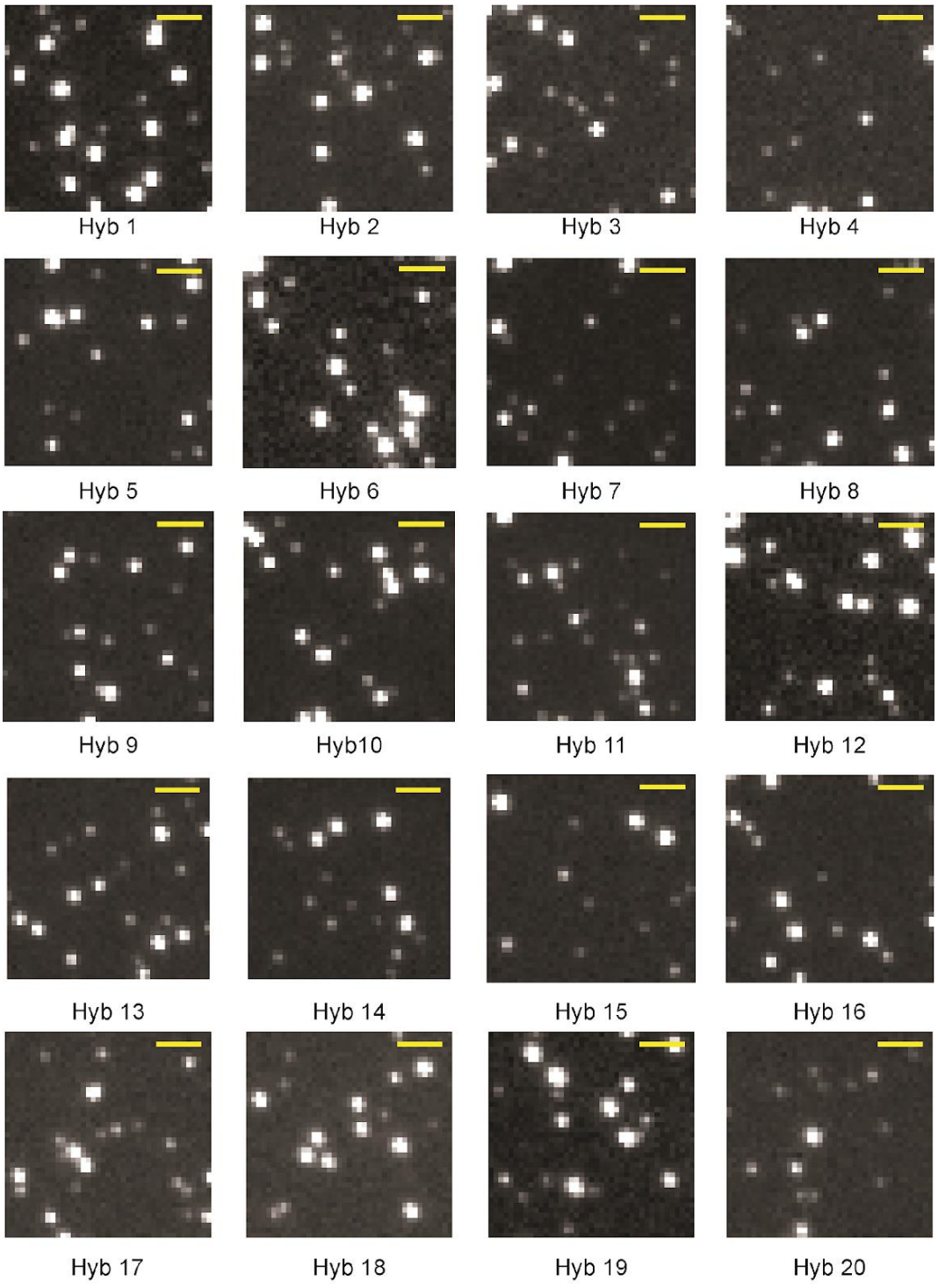
2.4 Supplementary Data and Figures



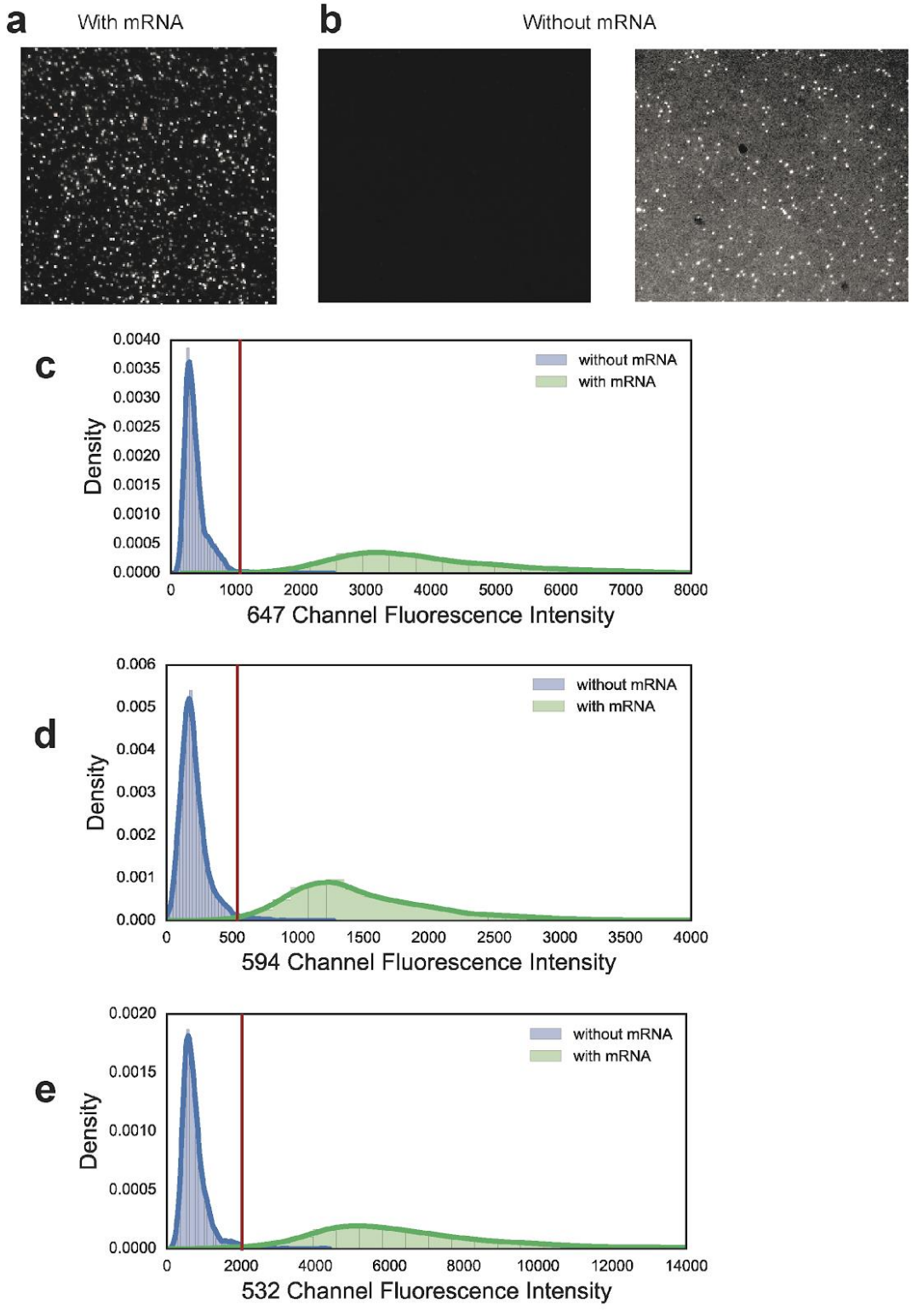
Supplementary Figure 1. RNA SPOTs hybridization and barcoding scheme. To implement transcriptome RNA SPOTs, 5 rounds of barcoding are needed to generate >20,000 different unique error-tolerant barcodes using a 12-base coding scheme to code for the transcriptome. A round of barcoding involves 4 serial hybridizations, each of which uses three unique secondary readout probes fluorescently labeled to Alexa 647, Alexa 594, and Cy3b dyes. The images from each 4 rounds of serial hybridizations are collapsed to form each 12-pseudocolor composite image which is aligned to decode for the barcoded RNA species.


a
readout probe
TCEP
SH
SH
SH
SH
SH
registration marker
Round 1 barcoding hyb
Round 2 barcoding hyb
Round 5 barcoding hyb
b
After 20 rounds of hybridization
488 Channel Hybridization 1
488 Channel Hybridization 20
c
TCEP
Channel 647
TCEP
Channel 594
TCEP
Channel 532

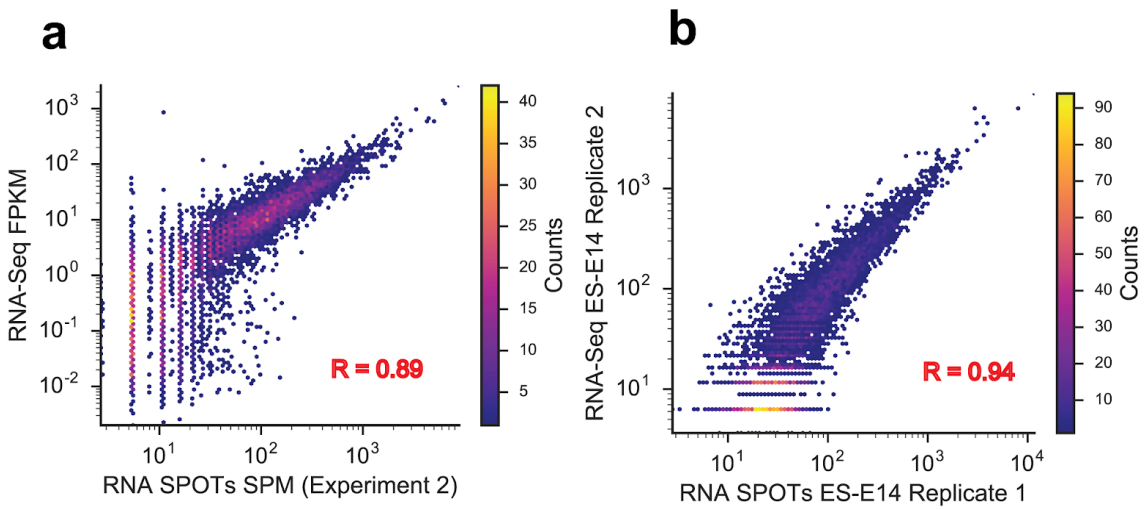
Supplementary Figure 2*(previous page).* Fluorescent switching through cleavage of disulfide conjugate dye on readout probes is highly efficient (a) 20 rounds of hybridization are accomplished by extinguishing fluorescent signals through reduction of disulfide conjugated dye to readout probes using TCEP, followed by re-hybridization of the next unique secondary readout probes. (b) Both priming regions (grey in the probe schematic) used in synthesizing gene specific primary probes are also used as a registration marker through the hybridization of Alexa 488 conjugated readout probes. The majority of the fluorescent spots stay even after 20 rounds of hybridizations. The amide bond between the Alexa 488 dye (shown in yellow) and primer readout probes used as a registration marker is not affected by TCEP. (Scale bars: 2μm.) (c) The fluorescent signals in each channel after treatment of 50mM of TCEP for 5 minutes at room temperature are reduced to minimal to none. (Scale bars: 5μm.)



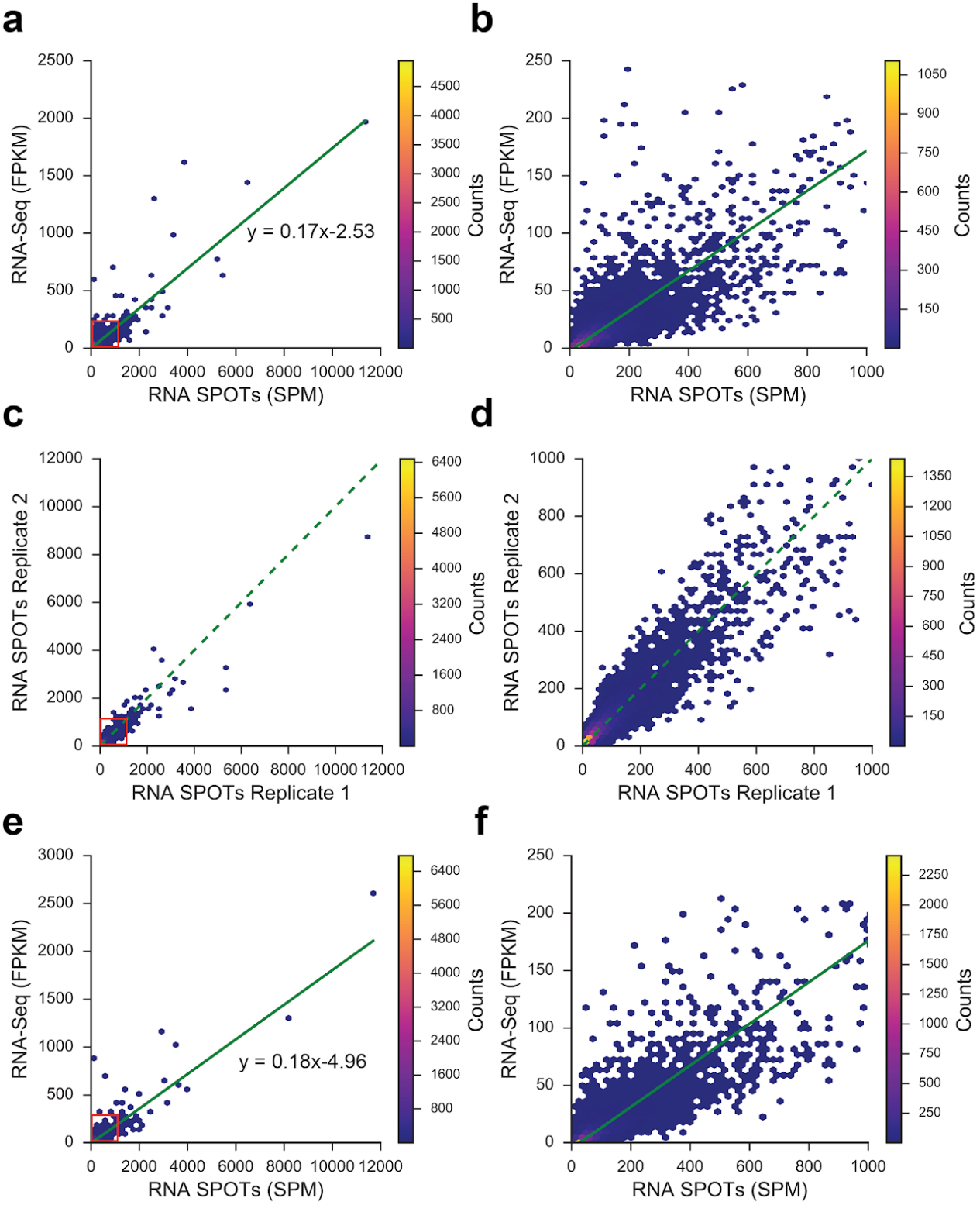
Supplementary Figure 3. Raw images of 20 rounds of fluorescent switching in channel 647. Bright dots are the real targets while dim dots are due to nonspecific binding. The switching between each round of hybridization is complete, with minimal retention of fluorescent signals from the previous round. (Scale bars: 2μm.)


a
With mRNA
b
Without mRNA
c
without mRNA
with mRNA
Density
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
4000
5000
6000
7000
8000
647 Channel Fluorescence Intensity
d
without mRNA
with mRNA
Density
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0
500
1000
1500
2000
2500
3000
3500
4000
594 Channel Fluorescence Intensity
e
without mRNA
with mRNA
Density
0.0020
0.0015
0.0010
0.0005
0.0000
0
2000
4000
6000
8000
10000
12000
14000
532 Channel Fluorescence Intensity

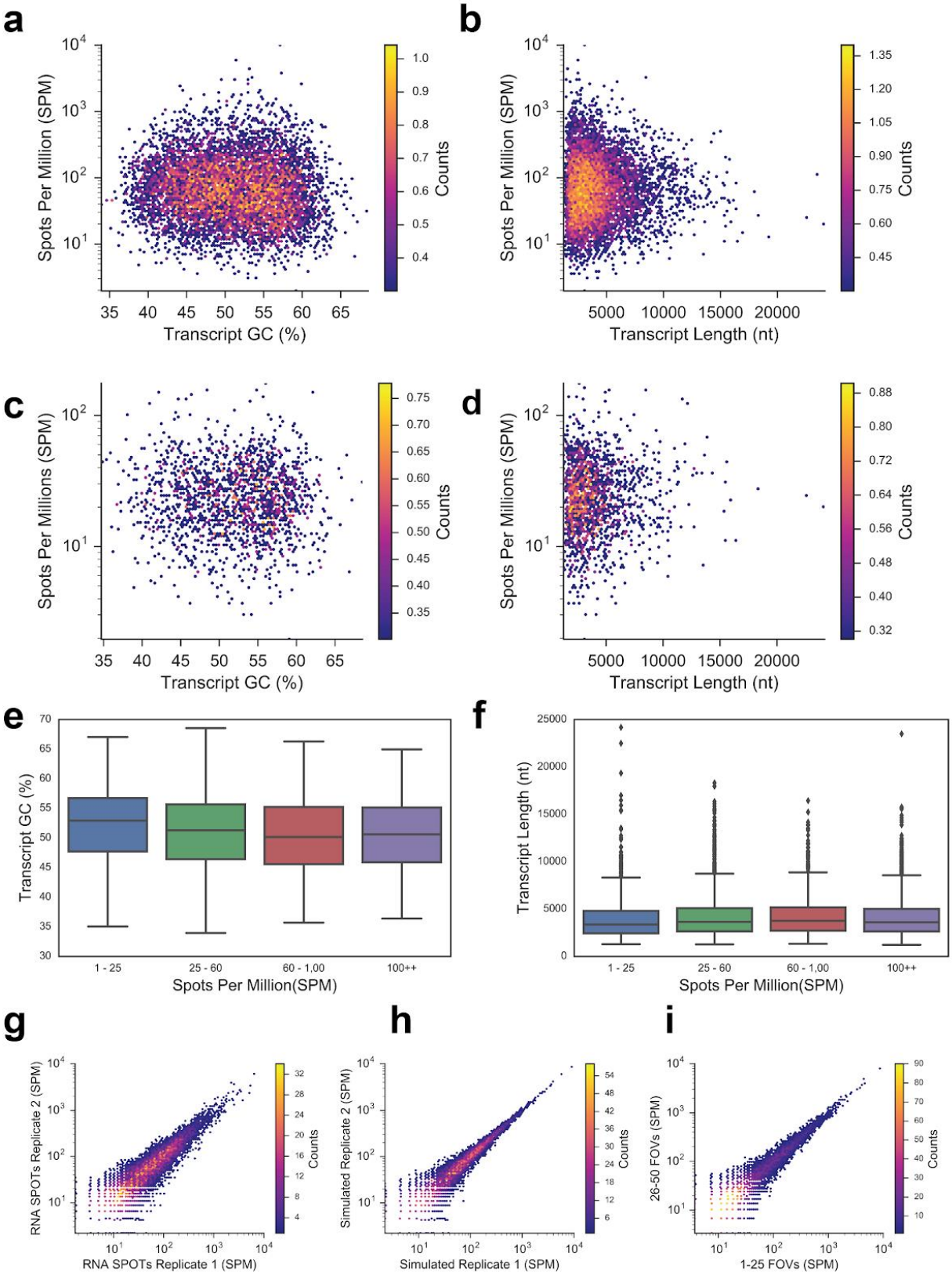
Supplementary Figure 4 (previous page). Assessment of primary probes non-specific binding. (a) Raw images of 532 channel with the presence of mRNA on coverslips through LNA poly(d)T capturing. (b) No bright fluorescent signals is observed in the absence of mRNA on coverslips as a control. The left image has the same contrast as (a) while the right image contrast has been increased 4.5 fold to illustrate better the non-specific fluorescent signals. (c) Quantitative measurement of fluorescent intensity in channel 647 with and without the presence of mRNA. A threshold can be set to distinguish between the two populations to identify the real signals. (d) & (e) same as (c) but for channel 594 and channel 532.



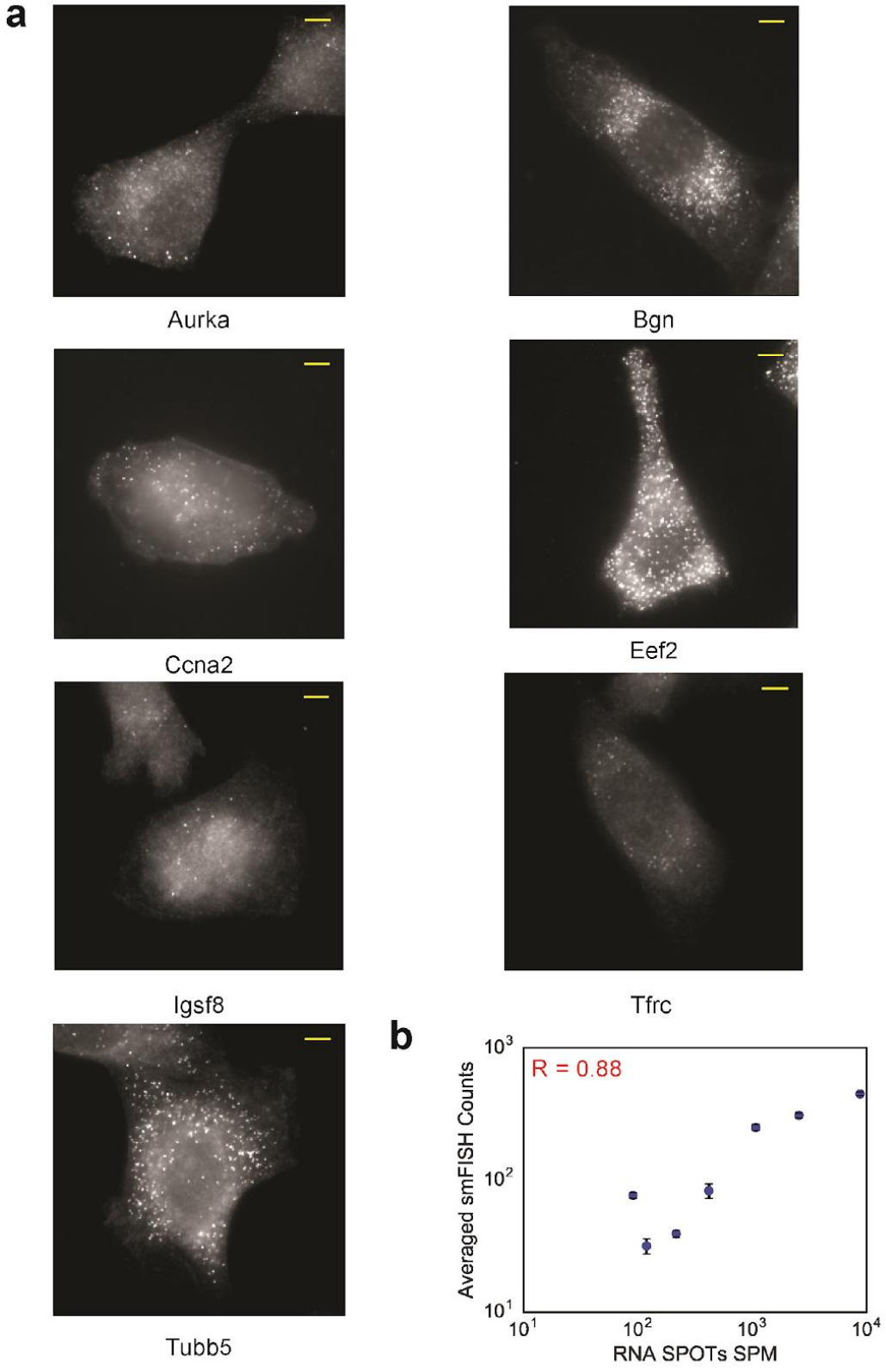
Supplementary Figure 5. RNA SPOTs at lower depth. (a) Correlation between RNA-seq FPKM and RNA SPOTs SPM from another replicate is high when a total of 376,781 spots are counted. SPM, spots per million; FPKM, fragments per kilobase per million reads. (b) High reproducibility of RNA SPOTs between the two replicates in profiling ES-E14 cell gene expression (n1=376,781 spots, n2=1,688,747 spots).


a
RNA-Seq (FPKM)
RNA SPOTs (SPM)
y = 0.17x-2.53
Counts
b
RNA-Seq (FPKM)
RNA SPOTs (SPM)
Counts
c
RNA SPOTs Replicate 2
RNA SPOTs Replicate 1
Counts
d
RNA SPOTs Replicate 2
RNA SPOTs Replicate 1
Counts
e
RNA-Seq (FPKM)
RNA SPOTs (SPM)
y = 0.18x-4.96
Counts
f
RNA-Seq (FPKM)
RNA SPOTs (SPM)
Counts

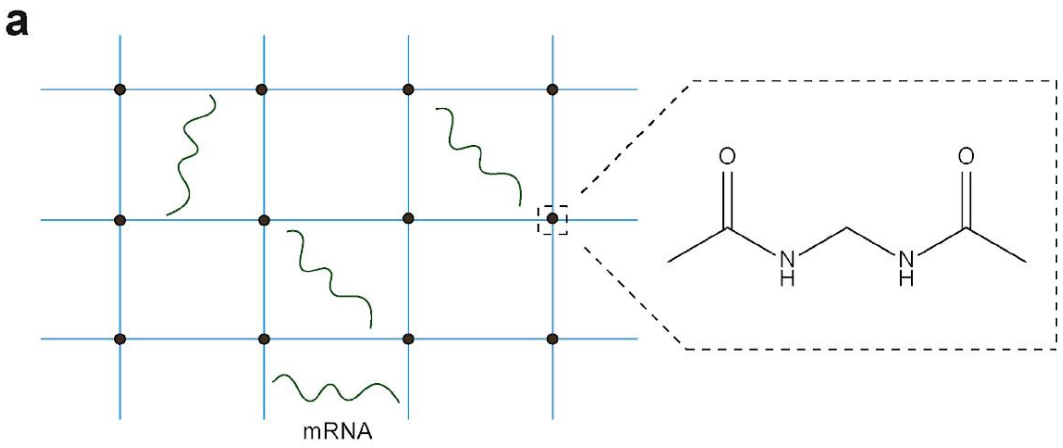
Supplementary Figure 6 (*previous page)*. Linear plots for Figure 2. (a) Correlation between RNA SPOTs and RNA-Seq for NIH/3T3 cells. (b) Zoomed-in boxed region in (a). (c) Reproducibility between two SPOTs replicates. The dashed line corresponds to the $y = x$ line. (d) Zoomed-in boxed region in (c). (e) Correlation between RNA SPOTs and RNA-Seq for ES-E14 cells. (f) Zoomed-in boxed region in (e).

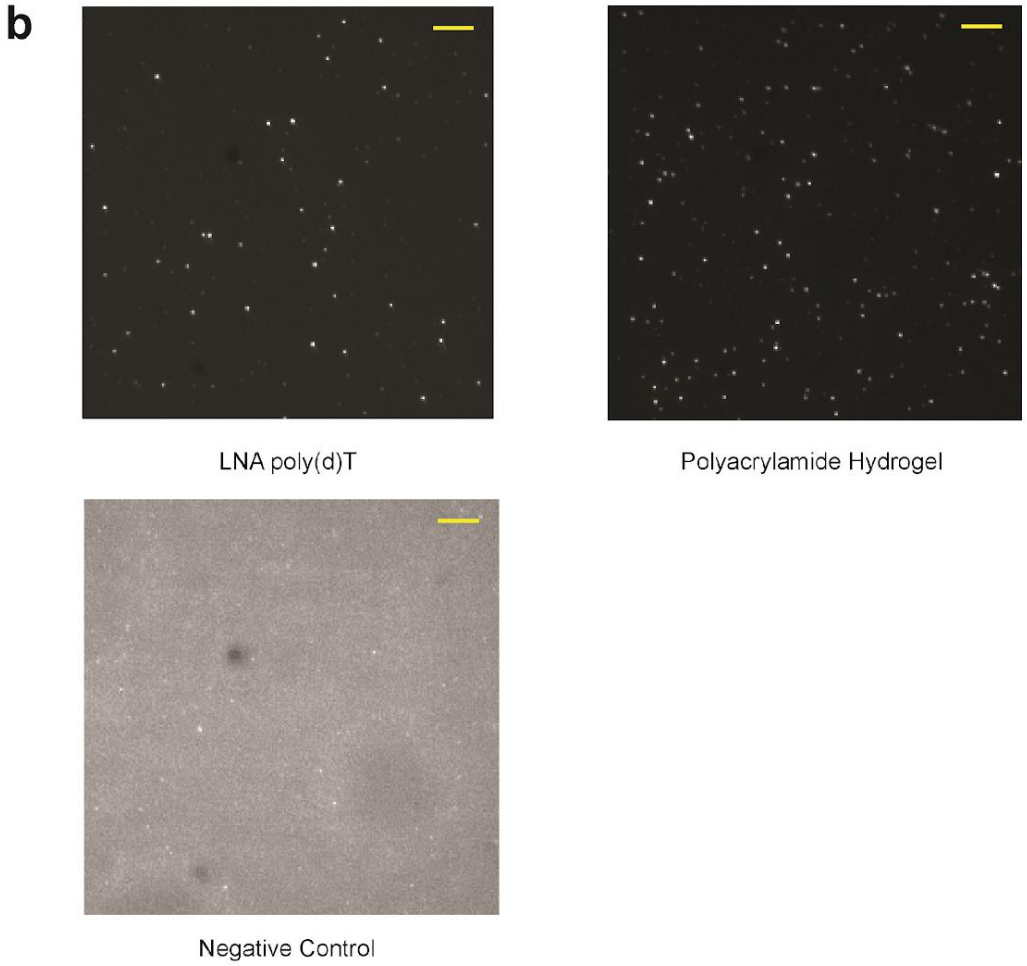

a
Spots Per Million (SPM)
Transcript GC (%)
Counts
b
Spots Per Million (SPM)
Transcript Length (nt)
Counts
c
Spots Per Millions (SPM)
Transcript GC (%)
Counts
d
Spots Per Millions (SPM)
Transcript Length (nt)
Counts
e
Transcript GC (%)
1 - 25
25 - 60
60 - 1,00
100++
Spots Per Million(SPM)
f
Transcript Length (nt)
1 - 25
25 - 60
60 - 1,00
100++
Spots Per Million(SPM)
g
RNA SPOTs Replicate 2 (SPM)
RNA SPOTs Replicate 1 (SPM)
Counts
h
Simulated Replicate 2 (SPM)
Simulated Replicate 1 (SPM)
Counts
i
26-50 FOVs (SPM)
1-25 FOVs (SPM)
Counts

Supplementary Figure 7 ***(previous page)*****.** RNA SPOTs has little bias in GC content and transcript length. (a) Hexbin plot of NIH/3T3 mean SPM (n=2) shows no obvious trend with transcript GC content. (b) Transcript length does not bias RNA SPOTs detection. (c) Same as (a) but for genes with < 1 FPKM. (d) same as (b) but for genes with < 1 FPKM. (e) Boxplots of different groups of genes with different expression levels against transcript GC content. n = 1360, 2323, 1473, 2645 for each group from left to right. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers. (f) Boxplot of SPM against transcript length. (g) Replicate plot of SPOTs as shown in Fig 2b. (h) Simulated replicates with Poisson noise. The total number of simulated SPOTs (n_1=447,094, n_2=448,249) was set to match the experimental replicates. (i). SPOTs data from two sets of field of views (FOVs) from the E14 experiment 3, x-axis contains 25 FOVs (n_1=269,459 spots), y-axis contains another 25 FOVs (n_2=295,403 spots).


a
Aurka
Bgn
Ccna2
Eef2
Igsf8
Tfrc
Tubb5
b
R = 0.88
Averaged smFISH Counts
RNA SPOTs SPM
10^1
10^2
10^3
10^4

Supplementary Figure 8 (*previous page*). smFISH measurement in single cells correlates with RNA SPOTs measurement in NIH/3T3 cells. (a) Raw images of the 7 genes measured by smFISH in NIH/3T3 cells. (Scale bars: 5μm.) (b) The averaged RNA smFISH counts agrees with RNA SPOTs SPM (spots per million) with a Pearson correlation coefficient of 0.88, indicating RNA SPOTs quantitation is accurate. Error bars represent the standard error of the mean (SEM) across different single cells.


a
O
O
N
H
N
H
mRNA


b
LNA poly(d)T
Polyacrylamide Hydrogel
Negative Control

Supplementary Figure 9 ***(previous page).*** mRNA can be immobilized by polyacrylamide hydrogel on a bind-silane treated coverslips. (a) mRNA is trapped in the hydrogel mesh once acrylamide and bis-acrylamide monomers crosslink completely on the coverslip. (b) smFISH detection of *ACTB* once the total RNA is captured on a coverslip through LNA poly(d)T capturing (left) or polyacrylamide hydrogel (right). Negative control (channel 488) shows that the fluorescent signals are not coming from nonspecific sources. (Scale bars: 5μm.)

Supplementary Table 1. Primary probe sequences for 10,212 genes (Provided as a separate Excel file)

Supplementary Table 2. 12-base code book for 10,212 genes (Provided as a separate Excel file)

Supplementary Table 3. Readout probes sequences (Provided as a separate Excel file)

Supplementary Table 4. Spots per million in NIH/3T3 and ES-E14 cells (Provided as a separate Excel file)

Supplementary Table 5. Summary of experiments in RNA SPOTs

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Cell lines	NIH/3T3	NIH/3T3	ES-E14	ES-E14
mRNA used	50ng	50ng	50ng	5ng
FOVs collected	125	202	134	143
Decoded spots	453,679	581,772	1,688,747	376,781

Supplementary Table 6. Differential gene expression between NIH/3T3 and ES-E14 cells (Provided as a separate Excel file.)

2.5 Methods

Primary probe design. Gene specific primary probes were designed as previously described with some modifications [Shah 2016]. Probe sets were crafted separately for each gene and then refined as a full set to mitigate cross-hybridization in the experiment. Individual probe sets were first crafted using exons only from within the CDS region of the gene. For genes that did not yield enough targeted probes from the CDS region only, exons from both the CDS and 5' UTR regions were used. The masked genome and annotation database from UCSC were used to look up the gene sequences. Consensus regions of all spliced isoforms were identified. 25-nt sequences of the gene sequences were extracted from these exons, and their GC contents were calculated. Probe sequences that fell outside of the allowed GC range (45-70% in this case) were immediately dropped. In addition, we dropped any probe sequences which contained 5 or more consecutive nucleotide bases of the same kind. A local BLAST query was run on each remaining probe against a BLAST database that was constructed from GENCODE reversed introns and mRNA sequences. BLAST hits on any sequences other than the target gene with a 15-nt match were considered off-target hits. We compiled a collection of RNA-seq data from ENCODE and computed a copy number table for all the genes across different samples. This off target copy number table was used to evaluate the off target hits. Any probe that hit an expected total off-target copy number exceeding 10,000 FPKM was dropped. Probes were sequentially dropped from genes until any off-target gene was hit by no more than 6 probes from entire pool. At this stage, all of the viable probes for the gene had been identified. For the final probe set, the best possible subset from the viable probes was selected such that none of the final probes were within 2 nucleotide bases of each other on the target sequence. The overlapping probes were grouped and sorted by distance from the target GC content (55% in this case). Overlapping probes were removed in order of descending distance from target GC, starting from the probe with the greatest distance, until no overlaps remained. To minimize cross hybridization between probe sets, a local BLAST database was constructed from all the viable probe sequences, and the probes were queried against it. All matches of 17-nt or longer between probes were removed by dropping the matched probe from the larger probe set. For this experiment, the targeted probe set size range was set to 28-32 probes. Any probe set with more than 32 probes was trimmed down by removing probes with the farthest GC content from 55%. To design the 20-nt readout sequences, a set of probe sequences were randomly generated with the 4 bases nucleotides. Readout probe sequences with range 45-60% GC were selected. We used BLAST to eliminate any sequences that matched with any contiguous homology

sequences longer than 14-nt to the mouse transcriptome. The reverse complements of these readout sequences were included in the primary probes according to the designed barcodes.

Primary probe construction. Primary probes were ordered as an oligoarray complex pools from Twist Bioscience and were constructed as previously described [Beliveau, 2012, Engreitz 2013]. Briefly, a 2-step limited PCR cycles were used to amplify the designated probe sequences from the oligo complex tool. Then, the amplified products were purified using QlAquick PCR Purification Kit (28104; Qiagen) according to the manufacturer's instructions. The PCR products were used as the template for in vitro transcription (E2040S; NEB) followed by reverse transcription (EP7051; Thermo Fischer) with the forward primer. After alkaline hydrolysis, the single stranded DNA (ssDNA) probes were purified by ethanol precipitation and resuspend in primary probe hybridization buffer comprising of 30% formamide (F9037; Sigma), 2x SSC (15557036; Thermo Fischer) , and 10% (w/v) Dextran Sulfate (D8906; Sigma). The probes were stored at -20°C.

Readout probe synthesis. 20-nt readout probes were ordered from Integrated DNA Technologies (IDT) as 3' thiol modified at its oxidized form. Alexa Fluor 647 Cadaverine (A30679; Invitrogen) and Alexa Fluor 594 Cadaverine (A30678; Invitrogen) were reacted with *N*-Succinimidyl 3-(2-pyridyldithio)propionate, SPDP (P3415; Sigma) at 1: 100 ratio in 1x PBS (AM9624, Ambion) at room temperature for at least 4 hours on a shaker. Then, the mixture was purified using PD MiniTrap G-10 (28-9180-10; GE Healthcare), and was evaporated in a vacuum concentrator. The dye-linker intermediate product was kept at -20°C until the conjugation with 3' thiol oligonucleotide probes. 10mM TCEP (77720; Thermo Scientific) was used to activate the 3' thiol readouts at 37°C for 30 minutes. Then the oligonucleotides were purified using illustra NAP-5 columns (17-0853-02; GE Healthcare), and the oligonucleotides were directly eluted in 1x PBS with 10mM EDTA (15575020; Thermo Fischer) and were mixed with the dye-linker intermediate product. The reaction was allowed to proceed at room temperature for 2 hours. Then, the mixture was ethanol precipitated, HPLC purified, resuspend into 500nM concentration in 1x Tris-EDTA buffer (93283; Sigma) and was kept at -20°C. To conjugate Cy3B fluorophore (PA63101; GE Healthcare) to the 3' thiol oligonucleotides, a (3-(2-pyridyldithio)propionyl hydrazide), PDPH (22301; Thermo Scientific) was used instead of the SPDP linker.

Coverslips functionalization. Coverslips were functionalized as previously described [Bose 2015] with some modifications. Briefly, coverslips (3421; Thermo Scientific) were sonicated in 100% ethanol for 20 minutes. After drying, the coverslips were cleaned with a plasma cleaner at HIGH (PDC-001, Harrick Plasma) for 5 minutes. Then, the coverslips

were immediately immersed in a 2% (v/v) trimethoxysilane aldehyde (PSX1050; UCT Specialties) solution made in pH 3.5 10% (v/v) acidic ethanol solution for 15 minutes at room temperature. After triple rinsing of the coverslips with ethanol, the coverslips were heat-cured at 90°C for 10 minutes. Then an oligonucleotide reaction mixture containing 2.5 μM 5'-aminated LNA-oligo(dT) (300100-02; Exiqon), cyanoborohydride coupling buffer (C4187; Sigma), and 1M sodium chloride (AM9759; Thermo Fischer) was sandwiched between two coverslips at room temperature in a humid hybridization chamber for 3 hours. The coverslips were then rinsed with Millipore water and dried with compressed air. A quenching reaction mixture made from 10%(v/v) 100mM pH7.5 Tris-HCl (15567027; Thermo Fischer) buffer in cyanoborohydride coupling buffer was added to the entire silanized surface of the coverslips to quench the remaining aldehyde functional groups at room temperature for 30 minutes. Finally, the coverslips were rinsed with water and dried with compressed air. All coverslips were made fresh before SPOTs experiment.

Cell cultures and RNA Preparation. ES-E14 cells were cultured as previously described[Singer 2014]. NIH/3T3 cells (ATCC) were cultured in DMEM (10569044; Gibco) supplemented with 10% FBS (S11150; Atlanta biologicals) and 1% penicillin (10378016; Gibco). Once the cell confluency reached 60-80%, the total RNA was extracted using RNeasy Mini Kit (74104; Qiagen) according to the manufacturer's instructions.

Hydrogel immobilization. Coverslips were first sonicated at 100% ethanol for 20 minutes, followed by plasma cleaning with a plasma cleaner at HIGH for 5 minutes. The coverslips were then immersed in the 2% PlusOne bind-silane(17-1330-01) solution made in ethanol for 30 minutes at room temperature. After rinsing the coverslips with ethanol for several times, the coverslips were dried at 90°C for 30 minutes. Purified total RNA was mixed in 4% acrylamide/bis solution (1610147; Bio-Rad) with fresh 25mM VA-044 initiator (27776-21-2; Wako Chemical) and the solution was degassed for 10 minutes on ice. A 12mm square coverslip (470019-000; VWR) was functionalized with GelSlick (Lonza; 50640). 1uL of the RNA hydrogel solution was added to the bind-silane functionalized coverslip and was spread out using the GelSlick functionalized square coverslip. The thickness of the hydrogel formed can be controlled by manipulating the volume added. The polymerization happened in a humid hybridization at 37°C for 2 hours. After polymerization was complete, the coverslips were immersed in 2x SSC for an hour or more to facilitate the removal of the top coverslips. smFISH measurement was then performed according to standard protocol.

Primary probe hybridization. A custom Secure Seal Flowcell, 2 x 28mm 3mm ID, 35 x 15 OD, 0.25mm thick (RD478685-M; Grace Bio-labs) was applied on the

functionalized poly(dT) coverslips. For NIH/3T3 cells experiments, 50 ng of total RNA in RNA binding buffer comprising of 1M LiCl (L9650; Sigma), 40mM pH7.5 Tris-HCl , 2mM EDTA, 0.1% Triton X-100 (93443; Sigma), and 20U of SUPERase IN RNase Inhibitor (AM2694, Ambion) was allowed to be captured at room temperature for 1 hour. For ES-E14 experiment 1 and 2, the amount of total RNA used was 50 ng and 5 ng respectively. Once the mRNA is immobilized on the coverslip, 20 uL of 1 nM/probe for a total of 323,156 probes in hybridization buffer containing 30% formamide (F9037; Sigma), 2x SSC (15557036; Thermo Fischer) , and 10% (w/v) Dextran Sulfate (D8906; Sigma) was hybridized to the targeted mRNA at 37°C for 24 hours in a humid hybridization chamber. After hybridization, the sample was washed for 30 minutes at room temperature with wash buffer containing 40% formamide, 2x SSC, and 0.1% Triton X-100 to remove non-specific binding of the primary probes. The sample preparation of primary probe hybridization ended with a 3 times washes with 2x SSC and was kept in 2x SSC until the next step.

RNA SPOTs imaging. Each readout probes hybridization mixture contained 10nM each for three unique readout probes either conjugated to Alexa 647, Alexa 594, or Cy3b in hybridization buffer comprising 10% formamide, 2x SSC, and 10% (w/v) Dextran Sulfate (D4911; Sigma). Each serial hybridization takes 15 minutes to achieve optimal fluorescent signals, followed by a 4-minutes high stringency wash containing 20% formamide and 2x SSC to remove non-specific binding of probes. Once the first hybridization is complete, the flow cell was connected to an automated fluidics delivery system made from two multichannel fluidics valves (EZ1213-820-4; IDEX Health & Science) and a peristaltic pump (NE-9000G-UP, New Era Pump Systems Inc.). The integration of the fluidics valves, peristaltic pump, and microscope imaging were controlled through a custom script written in Micromanager software. Once the flow cell is connected, ~100 to ~200 frame of views (FOVs) were imaged at 647-nm, 594-nm, 532-nm, and 488-nm channels with 500 ms exposure time under anti-bleaching buffer containing 20mM Tris-HCl pH 8 (15568025; Thermo Fischer), 50mM NaCl, 3mM Trolox (238813; Sigma), 0.8% glucose (G7528; Sigma), 3U/mL pyranose oxidase (P4234; Sigma) or 50U/mL of glucose oxidase (G2133; Sigma), and 20 U/mL SUPERase IN RNase Inhibitor. The anti-bleaching buffer was stored under a layer of mineral oil (M5904; Sigma) throughout the whole experiment. Imaging was done using a standard epifluorescence microscope (Nikon Ti Eclipse with custom built laser assembly), a Nikon 60x oil objective and a sCMOS camera (Zyla 4.2; Andor). Nikon Ti Eclipse PFS autofocus was activated to keep the plane focused during imaging. Once the imaging is complete, reduction buffer made from 50mM TCEP (646547; Sigma), 2x SSC, and 0.1% Triton X-100 was flowed into the flow cells and the solution was allowed to incubate for 5 minutes. Then, 2x SSC buffer supplemented with 20U/mL SUPERase IN RNase

Inhibitor was flowed into the flow cell in excess for 4minutes to completely remove the TCEP solutions. As our flow cell only takes ~22uL of solution, 200uL of subsequent serial hybridization solutions was flowed into the flow cell each time to ensure hybridization. The whole process was repeated until 20 rounds of hybridizations were imaged. Generally, a SPOTs experiment takes ~14 hours for imaging 100-200 FOVs. After the SPOTs imaging is complete, a few FOVs were imaged to use for threshold and illumination background corrections in image analysis. A multispectral beads slide was imaged at the end of experiment for chromatic aberration corrections.

Image Processing. To remove the effects of chromatic aberration, 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.

Image Registration. As the Alexa 488 channel labeled all the spots in the field of view, this channel was used to align all sets of images using a normalized 2D image cross-correlation.

Barcode calling. The potential RNA signals were then found by finding local maxima in the image above a predetermined pixel threshold in the registered images. 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 4 out of 5 rounds were used in the analysis. The number of each barcode was then counted and transcript numbers were assigned based on the number of on-target barcodes present. The remaining barcodes were used to assess the false

positives rate by running through the same process. All image processing and image analysis code can be obtained upon request.

smFISH. Unless stated, all smFISH measurements were conducted with 1nM/probe concentration with a total number of 24 probes targeting a gene in hybridization buffer comprising 10% formamide ,2x SSC and 10% (w/v) dextran sulfate at 37°C. The probes were conjugated to either Alexa 647, Alexa 594, or Cy3b dyes. NIH/3t3 cells were fixed with 4% paraformaldehyde (28908; Sigma) in 1x PBS at room temperature for 10 minutes. After washes with 1x PBS, the cells were permeabilized using 70% ethanol and kept in -20°C. The probe sequences for each gene were designed using Stellaris Biosearch Technologies and the probes were ordered from IDT with 5' amine modifications. The probes were conjugated to dye as previously described [Lubeck 2014]. After hybridization, the sample was washed with wash buffer supplemented with 30% formamide and 2x SSC at room temperature for 30 minutes. The samples were then stained with DAPI (D1306; Thermo Fischer) in 2x SSC, followed by imaging under anti-bleaching buffer. The cells were segmented and the copy numbers for each gene were counted using a custom Matlab script.

RNA-Seq. RNA-seq data were obtained from Gene Expression Omnibus (GEO) with an accession number of GSE98674. Briefly, the total RNA was purified using RNeasy Mini Kit following the manufacturer's instruction. The library was constructed using NEBNext ultra RNA-seq (E7530; NEB) according to the manufacturer's instructions and sequenced on Illumina HiSeq2500. Base calls were performed with RTA 1.13.48.0 followed by conversion to FASTQ with bcl2fastq 1.8.4. Alignment was performed using TopHat algorithm. Transcript assembly and FPKM estimates were done using Cufflinks algorithm.

Statistics and reproducibility. The technical replicates for RNA SPOTs of NIH/3T3 and ES-E14 are two in both cell cultures. The R values in the plots of technical replicates and SPOTs versus RNA-seq are Pearson's r correlation coefficient. For smFISH average measurements, the error bars represent the s.e.m. For differential gene expression analysis, two-tailed student t-test is carried out with n = 2 for mean SPM for both NIH/3T3 and ES-E14. P values smaller than 0.05 and $\log_2$ fold change greater and less than ±2 are used as a threshold for significance.

A Life Sciences Reporting Summary for this publication is available.

Data and software availability. The raw data for one field of view used to generate Figure 2 are available at Zenodo.org, doi:https://doi.org/10.5281/zenodo.1030239. Additional raw data from this study are

available from the corresponding author upon reasonable request. Custom-written scripts used in this study are available at https://github.com/CaiGroup/RNA-SPOTs and as Supplementary Software.

2.6 References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621-8.

2. Nagalakshmi, U. *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344–1349 (2008).

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

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

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

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

7. Takei Y, Shah S, Harvey S, Qi LS, Cai L. Multiplexed Dynamic Imaging of Genomic Loci by Combined CRISPR Imaging and DNA Sequential FISH. Biophys J. 2017 May 9;112(9):1773-1776.

8. Frieda KL, Linton JM, Hormoz S, Choi J, Chow KK, Singer ZS, Budde MW, Elowitz MB, Cai L. Synthetic recording and in situ readout of lineage information in single cells. Nature. 2017 Jan 5;541(7635):107-111.

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

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

11. Beliveau BJ, Joyce ER, Apostolopoulos N, Yilmaz F, Fonseka CY, McCole RB, Chang Y, Li JB, Senaratne TN, Williams BR, Rouillard J-M, Wu, C-t. A versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Nat. Acad. Sci. USA 2012 109:21301-6. PMID: 23236188; PMCID: PMC3535588.

12. Daniel, Steven G., et al. "FastTag nucleic acid labeling system: a versatile method for incorporating haptens, fluorochromes and affinity ligands into DNA." *RNA and oligonucleotides. Biotechniques* 24 (1998): 484-489.

13. Moffitt JR, Hao J, Wang G, Chen KH, Babcock HP, Zhuang X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc Natl Acad Sci U S A. 2016 Sep 27;113(39):11046-51.
14. Singer ZS, Yong J, Tischler J, Hackett JA, Altinok A, Surani MA, Cai L, Elowitz MB. Dynamic heterogeneity and DNA methylation in embryonic stem cells. Mol Cell. 2014 Jul 17;55(2):319-31.

15. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 2008 Mar;26(3):317-25.

16. Choi HM, Beck VA, Pierce NA. Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. ACS Nano. 2014 May 27;8(5):4284-94.

17. Bose S, Wan Z, Carr A, Rizvi AH, Vieira G, Pe'er D, Sims PA. Scalable microfluidics for single-cell RNA printing and sequencing. Genome Biol. 2015 Jun 6;16:120.

18. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, Adey A, Waterston RH, Trapnell C, Shendure J. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017 Aug 18;357(6352):661-667.

19. Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, Sladek R. TFCat: the curated catalog of mouse and human transcription factors. Genome Biol. 2009;10(3):R29. doi: 10.1186/gb-2009-10-3-r29.

20. Donner Y, Feng T, Benoist C, Koller D. Imputing gene expression from selectively reduced probe sets. Nat Methods. 2012 Nov;9(11):1120-5.

21. Beliveau BJ, Joyce ER, Apostolopoulos N, Yilmaz F, Fonseka CY, McCole RB, Chang Y, Li JB, Senaratne TN, Williams BR, Rouillard J-M, Wu, C-t. A versatile design

and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Nat. Acad. Sci. USA 2012 109:21301-6. PMID: 23236188; PMCID: PMC3535588.

22. Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, Surka C, Kadri S, Xing J, Goren A, Lander ES, Plath K, and Guttman M. (2013). The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X-chromosome. *Science*

Chapter 3

TRANSCRIPTOME-SCALE SUPER-RESOLVED IMAGING IN TISSUES BY RNA SEQFISH+

Eng, Chee-Huat Linus, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, et al. 2019. "Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA seqFISH+." *Nature* 568 (7751): 235–39. https://doi.org/10.1038/s41586-019-1049-y.

3.1 Abstract

Imaging the transcriptome *in situ* with high accuracy has been a major challenge in single cell biology, particularly hindered by the limits of optical resolution and the density of transcripts in single cells[1–5]. Here, we demonstrate seqFISH+, which can image the mRNAs for 10,000 genes in single cells with high accuracy and sub-diffraction-limit resolution, in the mouse brain cortex, subventricular zone, and the olfactory bulb, using a standard confocal microscope. The transcriptome level profiling of seqFISH+ allows unbiased identification of cell classes and their spatial organization in tissues. In addition, seqFISH+ reveals subcellular mRNA localization patterns in cells and ligand-receptor pairs across neighboring cells. This technology demonstrates the ability to generate spatial cell atlases and to perform discovery-driven studies of biological processes *in situ*.

3.2 Introduction

Spatial genomics, the analysis of the transcriptome and other genomic information directly in the native context of tissues, is crucial to many fields in biology, including neuroscience and developmental biology. Pioneering work in single molecule Fluorescence *in situ* Hybridization (smFISH) showed that individual mRNA molecules could be accurately detected in cells[6,7]. Development of sequential FISH (seqFISH) to impart a temporal barcode on RNAs through multiple rounds of hybridization allowed many molecules to be multiplexed[1–3]. Recently, we showed that seqFISH scales to the genome level *in vitro*[8] and for nascent transcription active sites[9].

However, the major challenge preventing global profiling mRNA in cells is the optical density of transcripts in cells: each mRNA occupies a diffraction limited spot in the image and there are tens to hundreds of thousands of mRNAs per cell depending on the cell type. Thus, optical crowding prevents mRNAs from being resolved and has bottlenecked all implementations of spatial profiling experiments [3-5]. For example, *in situ* sequencing methods, detected only ~500 transcripts per cell[4,5,10] because of the lower efficiency and larger dot size of rolling circle amplification, whereas seqFISH detected thousands of

transcripts per cell[3]. We have previously proposed to combine super-resolution microscopy with FISH[11] to overcome this crowding problem. However, existing super-resolution localization microscopy[12,13] relies on detection of single dye molecules, which emit a limited number of photons and only work robustly in optically thin (<1 μm) samples.

To enable discovery-driven approaches *in situ*, it is essential to scale up the spatial multiplexed methods to the genome level. To date, spatial methods have always relied on existing genomics methods, such as scRNAseq, to identify target genes, and serve to only map cell types identified from scRNAseq. At the level of hundreds and even a thousand genes, spatial methods cannot be used as *de novo* discovery-driven tool, which is a major drawback of the technology. In addition, many genes are expressed in a spatially dependent fashion independent of cell types[14] that is not recovered in the dissociated cell analysis.

3.3 Results

Here, we demonstrate seqFISH+, which achieves super-resolution imaging and multiplexing of 10,000 genes in single cells using sequential hybridizations and imaging with a standard confocal microscope. The key to seqFISH+ is expanding the barcode base palette from 4-5 colors, as used in seqFISH[1,3] and *in situ* sequencing experiments[4,5], to a much larger palette of "pseudocolors" (Figure 1a) achieved by sequential hybridization. By using 60 pseudocolor channels, we effectively dilute mRNA molecules into 60 separate images and allows each mRNA dot to be localized below the diffraction limit[12,15,16] before recombining the images to reconstruct a super-resolution image. We separate the 60 pseudocolors into 3 fluorescent channels (Alexa 488, Cy3b and Alexa 647) and generate barcodes only within each channel to avoid chromatic aberrations between channels. 20^3=8000 genes can be barcoded in each channel for a total of 24,000 genes by repeating this pseudocolor imaging 4 times with one round used for error-correction[3].

As imaging time is the main bottleneck in spatial transcriptomics experiments, seqFISH+ is 8-fold faster in imaging time compared to implementing seqFISH with expansion microscopy[17] (Figure 1b). An equivalent 60-fold expansion of the sample would require 4 colors x 8 barcoding rounds x 60 volume expansion = 1920 images per field of view (FOV) to cover 4^7=16,384 genes. In contrast, seqFISH+ acquires 60 pseudocolors x 4 barcoding rounds = 240 images per FOV to cover 24,000 genes, an 8-fold reduction in imaging time. Furthermore, a large number of pseudocolors and a shorter barcode (4 units) decreases the errors that accumulate over barcode rounds.

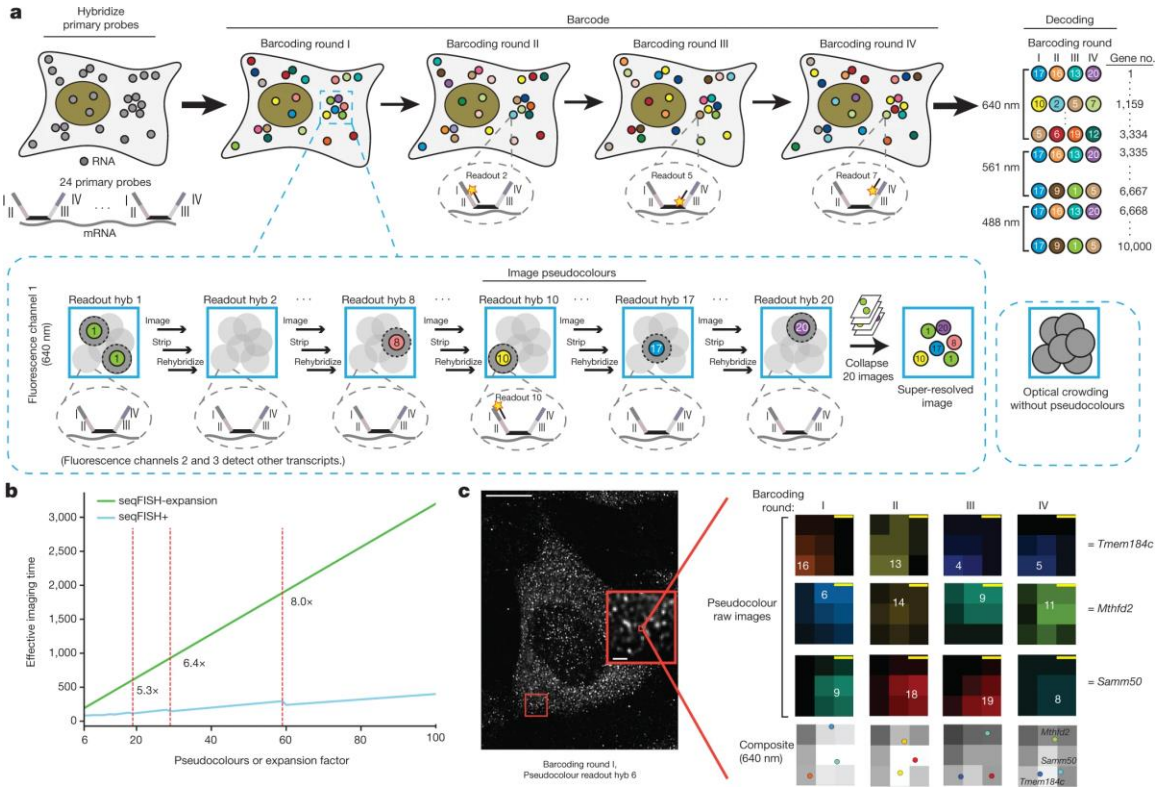


Figure 1. seqFISH+ resolves optical crowding and enables transcriptome profiling *in situ*. a, Schematics of seqFISH+. Primary probes (24 per gene) against 10,000 genes are hybridized in cells. Overhang sequences (I-IV) on the primary probes correspond to 4 barcoding rounds (orange panel). Only 1/20th of the total genes in each fluorescent channel are labeled by readout probes in each pseudocolor readout round, lowering the density of transcripts in each image. mRNA dots in each pseudocolor can then be localized by Gaussian fitting and collapsed into a super-resolved image (blue panel). Each gene is barcoded within only one fluorescent channel (Methods). **b,** Compared to seqFISH with expansion microscopy (seqFISH-Expansion, green line) in covering 24,000 genes, seqFISH+ with 60 pseudocolors (blue line) is 8 fold faster in imaging time. (Methods). **c,** Image of a NIH3T3 cell from one round of hybridization (n = 227 cells; scale bar = 10 μm). Zoomed in inset shows individual mRNAs (scale bar = 1 μm). Different mRNAs are decoded within a diffraction limited region, magnified from the inset (scale bar = 100 nm). The number in each panel corresponds to the pseudocolor round that each mRNA was detected, with no dots detected during the other pseudocolor rounds in this channel (640 nm).

To demonstrate transcriptome level profiling in cells, we first applied seqFISH+ to cleared NIH3T3 fibroblast cells (Figure 1c, Extended Data Figure 1,2)[18–20]. We randomly selected 10,000 genes while avoiding highly abundant housekeeping genes, such as ribosomal proteins. These 10,000 genes add up to a total of >125,000 FPKM values with a wide range of expression levels from 0 to 995.1 FPKM. All 24,000 genes in the fibroblast transcriptome add up to ~420,000 FPKM[21], only a 3 fold higher density from the 10,000 gene experiment, which can be accommodated with the current scheme, or with more channels or pseudocolors.

Overall, 35,492±12,222 (mean±s.d.) transcripts are detected per cell (Figure 2a). The 10,000 seqFISH+ data are highly reproducible and strongly correlated with RNA-seq (R=0.80)[21], RNA SPOTs (R=0.80)[8], and smFISH (R=0.87) (Figure 2b-d, Extended Data Figure 3a,b). Each of the three fluorescent channels was decoded independently and correlated well with RNA-seq and smFISH (Extended Data Figure 3a,c). The false positive rate per cell is 0.22±0.07 (mean±s.d.) per barcode (Extended Data Figure 3d,e). Comparison with 60 genes from smFISH showed that the seqFISH+ detection efficiency is 49%, which is highly sensitive compared to single cell RNAseq.

seqFISH+ allows us to visualize the subcellular localization patterns for tens of thousands of RNA molecules *in situ* in single cells. Three major clusters were observed to be nuclear/peri-nuclear, cytoplasmic and protrusion enriched. Many new protrusion localized genes are found in addition to the ones identified previously[22,23]. We further observed three distinct subclusters in the perinuclear/nuclear localized transcripts with genes in each of these subclusters enriched in distinct functional roles (Extended Data Figure 3f-j).

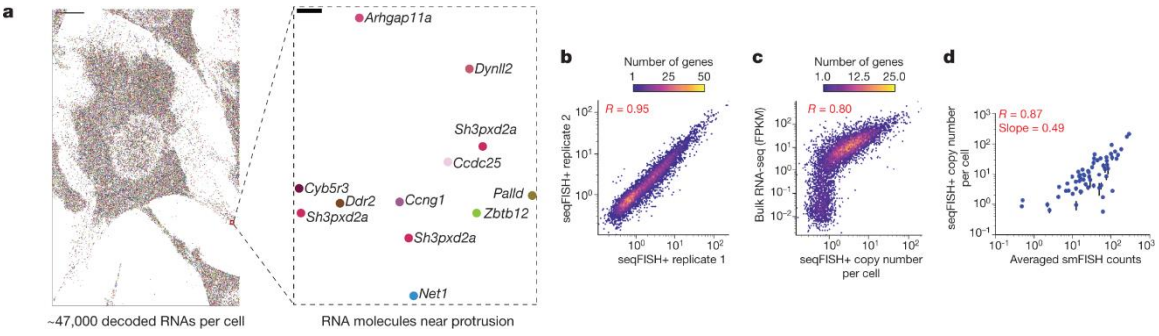


Figure 2 seqFISH+ profiles 10,000 genes in cells with high efficiency. a, Approximately 47,000 mRNAs (colored dots) were identified in a NIH3T3 cell from a single z-section (scale bar = 10 μm). Inset shows the transcripts decoded in cell protrusions (n = 227 cells; scale bar = 100 nm). **b,** seqFISH+ replicates in NIH3T3 cells are highly reproducible (n_1 = 103 cells; n_2 = 124 cells). seqFISH+ correlates well with (**c**) RNA-seq (n = 9875 genes)

Figure 2 (continued from above) and (**d**) single molecule FISH (n= 60 genes; p-value = 2.26 x 10^{-19}). The efficiency of seqFISH+ is about 49% compared to smFISH. Error bars in (**d**) represents standard error of the mean. (**b-d,** p-values < 0.0001, Pearson's r, two-tailed p values).

To demonstrate seqFISH+ works robustly in tissues, we used the same 10,000 gene probe set to image cells in the mouse brain cortex, the sub-ventricular zone (SVZ) (Figure 3a), and the olfactory bulb in two separate brain sections. We collected 10,000-gene-profiles for 2963 cells (Figure 3b-e), covering an area of approximately 0.5 mm^2. In the cortex, cells contained on average 5615±3307 (mean±s.d.) transcripts from 3338±1489 (mean±s.d.) detected genes (Extended Data Figure 4a,b). We imaged only a single z optical plane (0.75 μm) to save imaging time. Full 3D imaging of cells with seqFISH+ is available for 5-10x "deeper" sampling of the transcriptome.

With an unsupervised clustering analysis[24], the seqFISH+ cell clusters show clear layer structures (Figure 3h) and are strongly correlated to the clusters in a scRNAseq[25] dataset (Methods, Extended Data Figure 4c-f, 5). Similar layer patterns are observed with Hidden Markov Random Field (HMRF) analysis[14] where the expression patterns of neighboring cells were taken into account (Extended Data Figure 4g-i, 6).

With the seqFISH+ data, we can explore the subcellular localization patterns of 10,000 mRNAs directly in the brain in a cell type specific fashion (Supplementary Table 3). In many cells types, the transcripts for *Snrnp70*, a small nuclear riboprotein, and *Nr4a1*, a nuclear receptor, are found in the nuclear/perinuclear regions. In contrast, *Atp1b2*, a Na+/K+ ATPase, and *Kif5a*, a kinesin, are observed to be near the cell peripheries in many cell types including excitatory, inhibitory neurons as well as glia cells. In addition, many transcripts in astrocytes, such as *Gja1* and *Htra1*, localize to the cell periphery and processes, which we confirmed by smFISH (Figure 3f,g, Extended Data Figure 7).

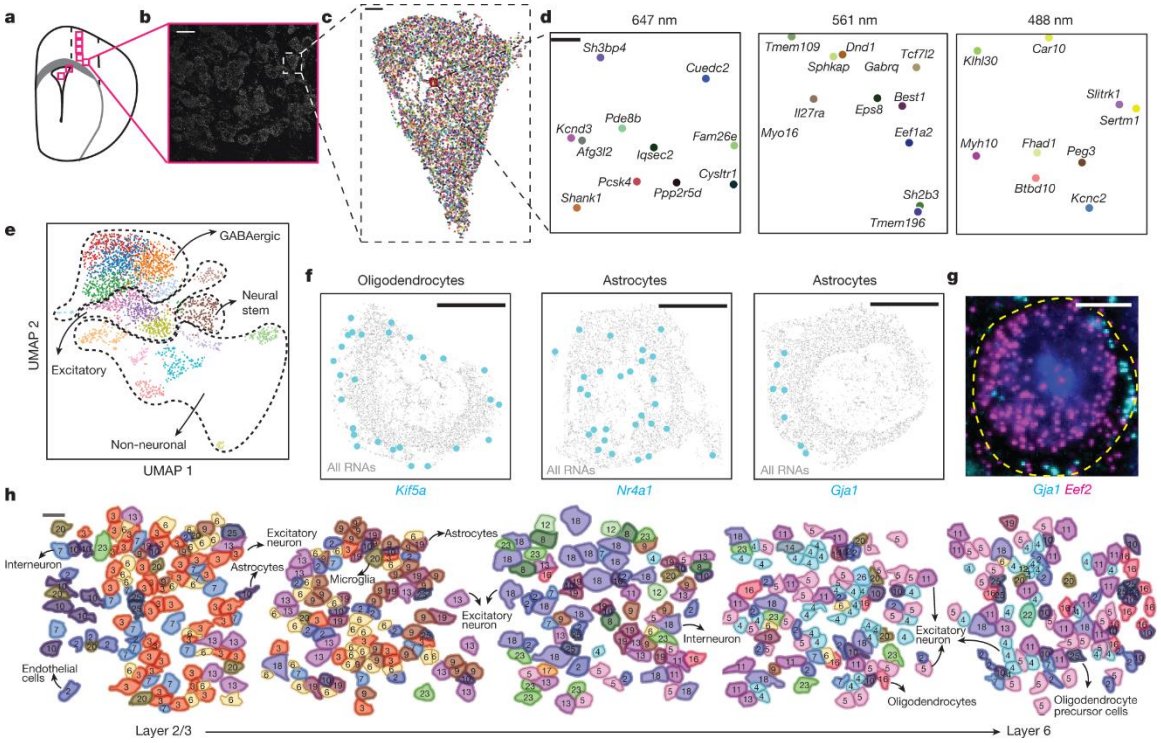


Figure 3. seqFISH+ robustly characterize cell classes and subcellular RNA localization in brain slices. a, Schematic of the regions (red boxes) imaged. **b,** Cells in a single FOV of the primary motor cortex (scale bar = 20 μm). **c,** Reconstruction of the 9,418 mRNAs (colored dots) detected in a cell (scale bar = 2 μm). **d,** Decoded transcripts for a magnified region (n= 523 cells, scale bar= 100nm). **e,** Uniform Manifold Approximation and Projection (UMAP) representation of the seqFISH+ data in the cortex, SVZ, and olfactory bulb (n=2963 cells). **f,** Reconstructed seqFISH+ images show subcellular localization patterns for mRNAs (Cyan) in different cell types. (n = 62 astrocytes and 28 oligodendrocytes; scale bar = 2 μm). **g,** smFISH of *Gja1* in cortical astrocytes shows periphery localization compared to the uniform distribution of *Eef2* mRNAs. (n=10 FOVs,40x objective;scale bar = 5μm). **h**. Each cortex layer consists of a distinct cell class composition (see annotations, Supplementary Table 2). (scale bar = 20 μm).

We next explored the spatial organization of the SVZ. We identified neural stem cells (NSCs, Clusters 8,16) expressing astrocyte markers *Gja1* and *Htra1*, transit-amplifying progenitors (TAPs, Cluster 15) expressing *Ascl1, Mcm5* and *Mki67,* and neuroblasts (NBs) expressing *Dlx1* and *Sp9*, consistent with previous studies[26]. We further quantified the spatial organization of the different cell types in the SVZ (Figure 4a, Extended Data Figure

8), and found that class 12 and 17 neuroblasts are preferentially in contact, whereas TAP cells tend to associate with other TAP cells. It would be exciting to further investigate the RNA velocity trajectories[27] of these cells *in situ* with intron seqFISH[9] as well as their lineage relationships with MEMOIR[28].

Next, we examined the spatial organization of the olfactory bulb. Our clustering analysis revealed distinct classes of GABAergic interneurons, olfactory ensheathing cells (OECs), astrocytes, microglia, and endothelial cells (Figure 4b,c), consistent with literature[29]. In the granule cell layer (GCL) at the center of the olfactory bulb, several cell classes are observed, with an interior core consisting of immature neuroblast-like cells expressing *Dlx1* and *Dlx2* encased by a distinct outer layer of the GCL composed of more mature interneurons (Figure 4b and Extended Data Figure 9,10). An excitatory cluster of cells expressing *Reln*, *Slc17a7* are observed in the mitral cell layer (MCL) as mitral cells and in the external plexiform layer (EPL) and glomerulus as tufted cells. We also found several clusters of *Th*+ dopaminergic neurons (Figure 4b-d, Supplementary Table 2) which were previously not known. For example, Cluster 1 cells express both *Vgf*, a neuropeptide, as well as tyrosine hydroxylase (*Th*), and are distributed both in the glomerulus and the GCL. Similarly, *Trh* is enriched in a distinct set of *Th*+ cells (Cluster 3), which are predominantly in the glomerulus, whereas Clusters 5 and 22 dopaminergic neurons are in the GCL. We validated these clusters by smFISH imaging (Figure 4d, Extended Data Figure 9,10).

Finally, we analyzed ligand-receptor pairs that are enriched in neighboring cells, which are not available in the dissociated cell analysis. These proposed potential cell-cell interactions are hypothesized on the basis of mRNA and not protein. In endothelial cells adjacent to microglia in the olfactory bulb, Endoglin (*Eng*, a type III TGF-β receptor) and Activin A-receptor (*Acvrl1 or Alk1*, a type I TGF-β receptor) mRNAs are expressed, with TGF-β ligand (*Tgfb1*) mRNA expressed by the microglia. Microglia-endothelial neighbor cells express, *Lrp1 (Tgfbr5)* and *Pdgfb*, in the cortex, indicating that signaling pathways may be used in a tissue specific fashion. Beyond ligand receptor interactions, we found broadly that gene expression patterns in a particular cell type are highly dependent on the local tissue context of neighboring cells (Figure 4e,f, Supplementary Table 4).

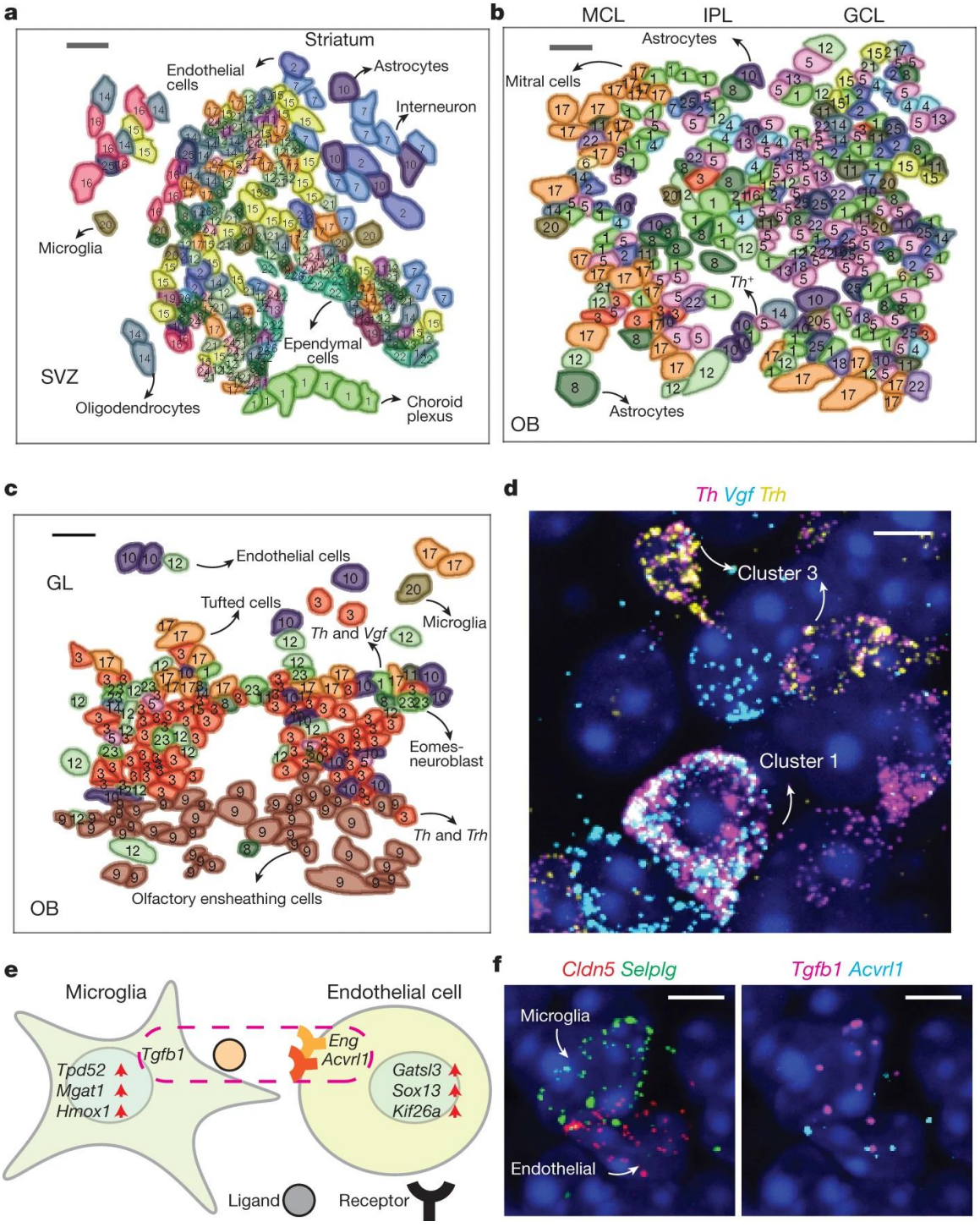

a
Striatum
Endothelial cells
Astrocytes
Interneuron
Microglia
Ependymal cells
SVZ
Oligodendrocytes
Choroid plexus
b
MCL
IPL
GCL
Astrocytes
Mitral cells
Th+
Astrocytes
OB
c
GL
Endothelial cells
Tufted cells
Microglia
Th and Vgf
Eomes-neuroblast
Th and Trh
Olfactory ensheathing cells
OB
d
Th Vgf Trh
Cluster 3
Cluster 1
e
Microglia
Endothelial cell
Tgfb1
Eng
Acvrl1
Tpd52
Mgat1
Hmox1
Gatsl3
Sox13
Kif26a
Ligand
Receptor
f
Cldn5 Selplg
Tgfb1 Acvrl1
Microglia
Endothelial

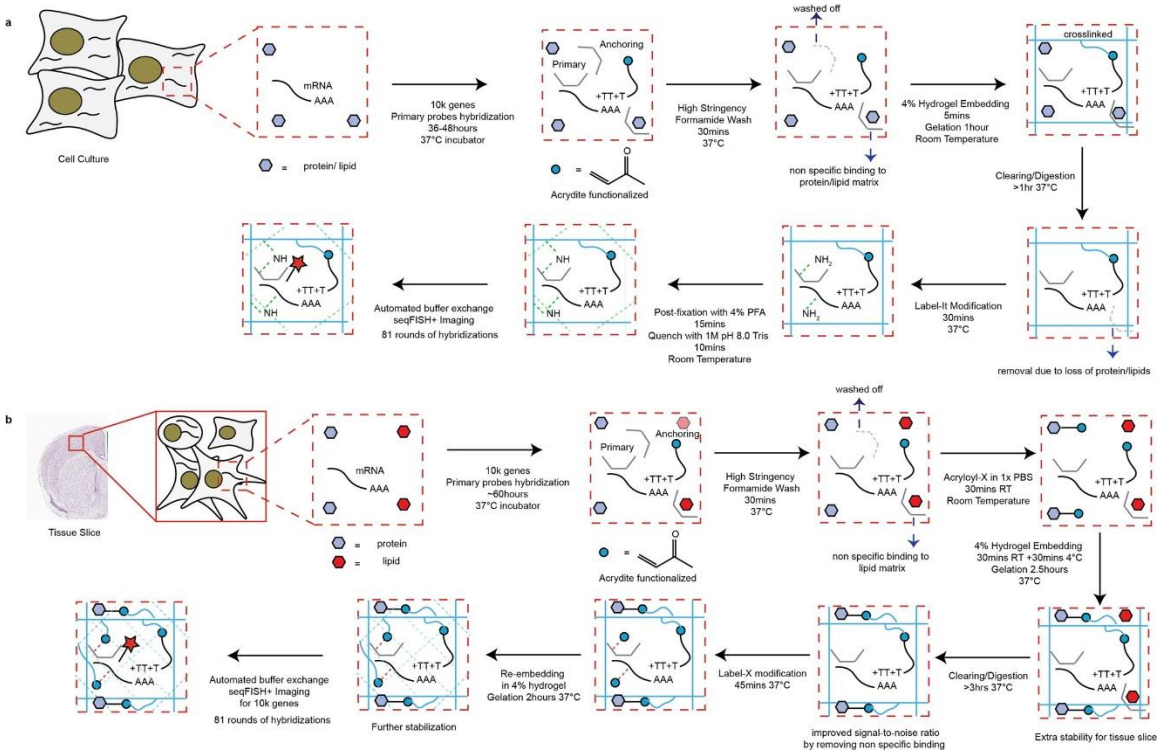
Figure 4 (*previous page*). seqFISH+ reveals ligand receptor repertoires in neighboring cells and spatial organization in tissues. a, Spatial organization of distinct cell clusters in the SVZ. **b,** Spatially-resolved cell cluster maps of the mitral cell layer(MCL), granule cell layer(GCL), and **c,** glomerular layer(GL) (scale bars: 20 μm). Remaining FOVs are shown in Extended Data Figure 10. The cluster numbers in the SVZ and OB are different (Supplementary Table 2). **d,** Distinct populations of *Th*+ dopaminergic neurons in the OB with differential expression of *Vgf* and *Trh*, shown with smFISH, confirming seqFISH+ clustering analysis. **e,** Schematic showing ligand-receptor pairs in neighboring microglia-endothelial cells. In microglia next to endothelial cells, certain genes, such as *Tpd52*, are enriched compared to microglia neighboring other cell types. **f,** mRNAs of *Tgfb1* ligand and *Acvrl1* receptor are visualized in adjacent microglia-endothelial cells by smFISH. (**d&f,** n = 10 FOVs, 40x objective; scale bars = 5 μm)

3.4 Discussions

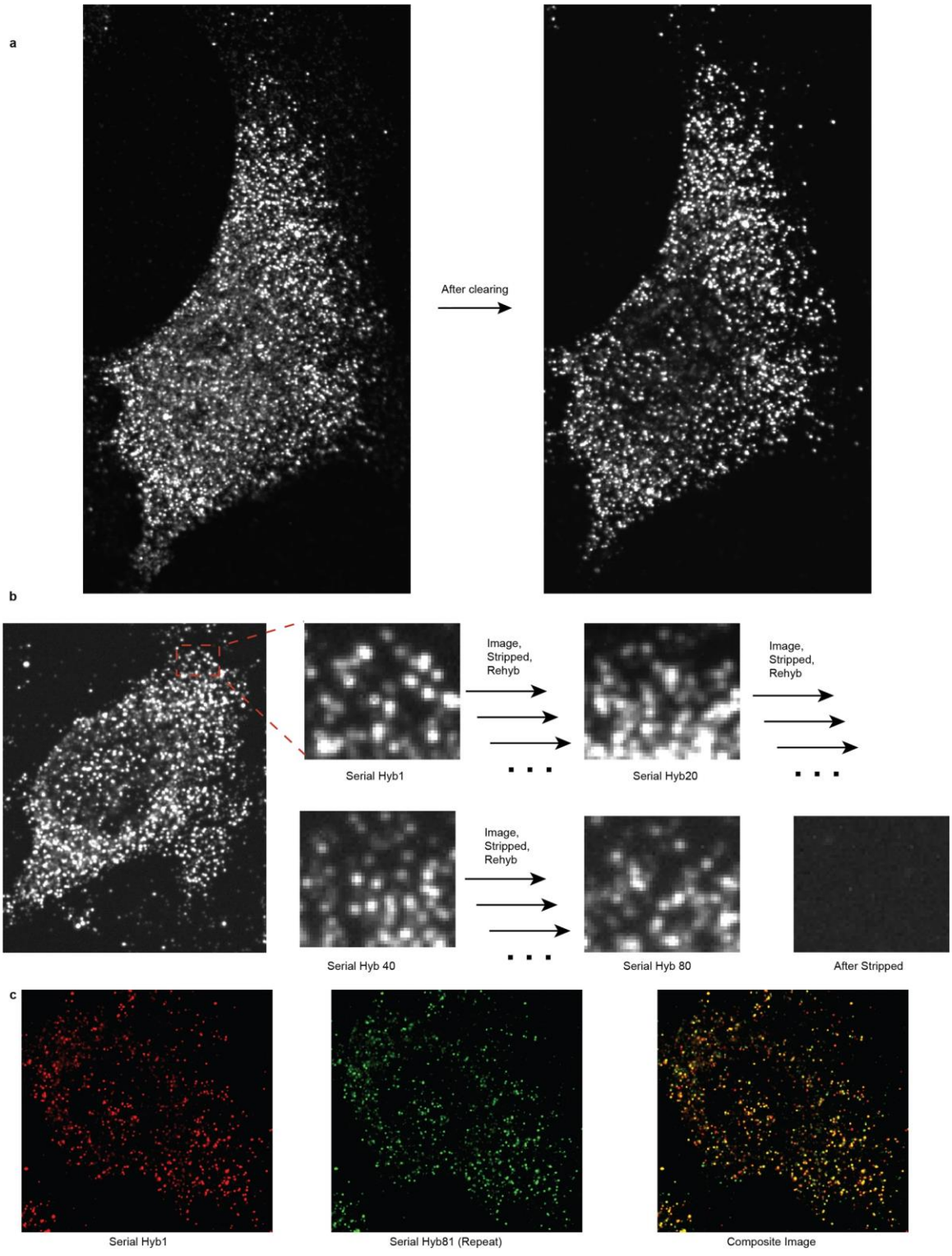
These experiments demonstrate that seqFISH+ can robustly profile transcriptomes in tissues, overcoming optical crowding and removing the last conceptual roadblock in generating spatial single cell atlases in tissues. seqFISH+ provides 10-fold or more improvement over existing methods in the number of mRNAs profiled and the total number of RNA barcodes detected per cell. seqFISH+ also allows super-resolved imaging with commercial confocal microscopes and can be generalized to chromosome[30] and protein imaging.

With the genome coverage and spatial resolution of seqFISH+, it is now possible to perform discovery-driven studies directly *in situ*. In particular, elucidating signaling interactions between cells is a crucial first step towards understanding developmental processes and cell fate decisions, along with explorations of the combinatorial signaling logic[21]. Lastly, the genomics coverage of seqFISH+ will allow discovery of novel targets that are cell type specific in disease samples as well as enable precise spatial-genomics and single-cell based diagnostics test.

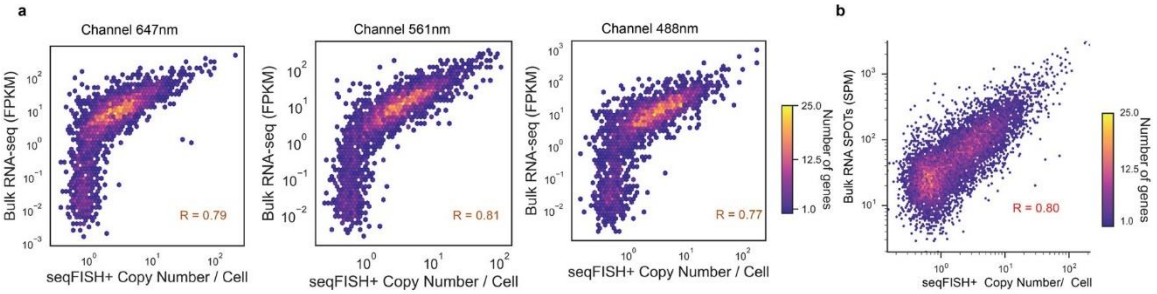
3.5 Supplementary Data and Figures

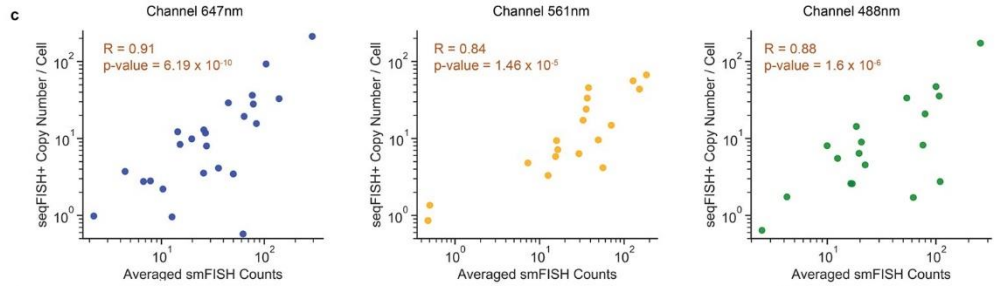


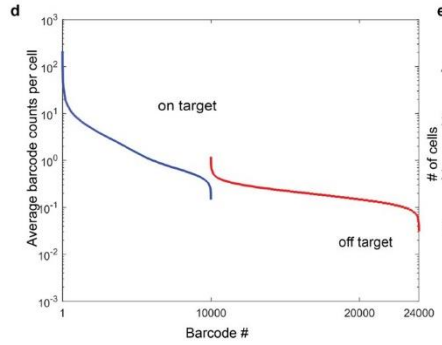
Extended Data Figure 1. Clearing and probe anchoring protocols for the seqFISH+ experiments in (**a**) NIH3T3 cells and (**b**) the mouse brain slices.

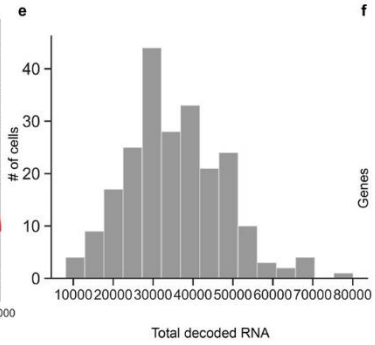

a
After clearing
b
Image, Stripped, Rehyb
Serial Hyb1
Image, Stripped, Rehyb
Serial Hyb20
Image, Stripped, Rehyb
Serial Hyb 40
Serial Hyb 80
After Stripped
c
Serial Hyb1
Serial Hyb81 (Repeat)
Composite Image

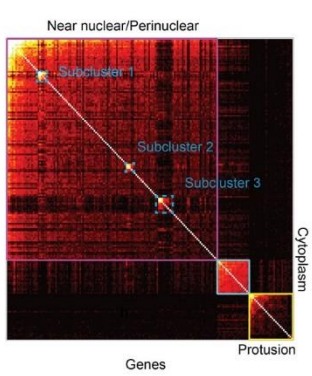
Extended Data Figure 2 (*previous page*). Clearing removes background nonspecific bound dots. **a,** Raw images of a NIH3T3 cell before and after clearing. Significant decrease in background is observed in cleared sample. Image is acquired on a spinning disk confocal microscope. **b,** In each round of hybridization for the 10,000 gene experiment, diffraction limited dots are clearly separated, indicating the pseudocolor scheme effectively dilutes the density of the sample. Signal is completely removed between different rounds of hybridization, with no "cross-talk" between the pseudocolors. Stripping is accomplished by 55% formamide wash, which is highly efficient. **c,** After the completion of each seqFISH+ experiment, readout probes used in hyb1 is re-hybridized in round 81. The colocalization rates between Hyb1 and 81 are 76% (647 channel), 73% (561 channel) and 80% (488 channel) within a 2-pixel radius. The colocalization between the two images indicates that most of the primary probes remain bound through 80 rounds of hybridization and imaging, although some loss of RNA and signal is seen across 80 rounds of hybridization (**a-c,** n = 227 cells).



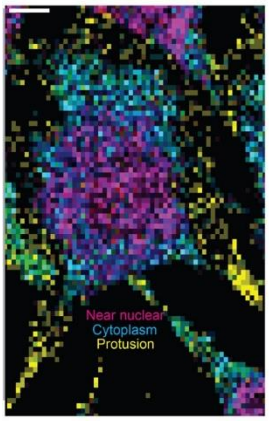








g



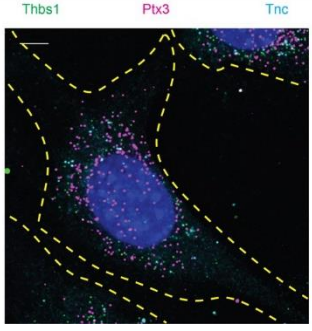
h

Protrusion	Nuclear/Perinuclear	Cytoplasm
Cyb5r3	Col1a1	Ddb1
Sh3pxd2a	Fn1	Myh9
Ddr2	Fbln2	Actn1
Net1	Col6a2	Tagln2
Trak2	Bgn	Kpnb1
Kif1c	Nid1	Hnrnpf
Kctd10	Lox	Ppp1ca
Dynll2	P4hb	Hnrnpl
Arhgap11a	Aebp1	Pcbp1
Gxylt1	Emp1	Tagln
H6pd	Col5a1	Fscn1
Gdf11	Sdc4	Psat1
Dync1li2	Postn	Cald1
Palld	Col3a1	Snd1
Ppfia1	Pdia6	Uba1
Naa50	Col5a2	Hnrnpm
Ptgfr	Itgb1	Cap1
Zeb1	Calu	Ssrp1
Arhgap32	Pdia3	Ugdh
Scd1	Cyr61	Caprin1

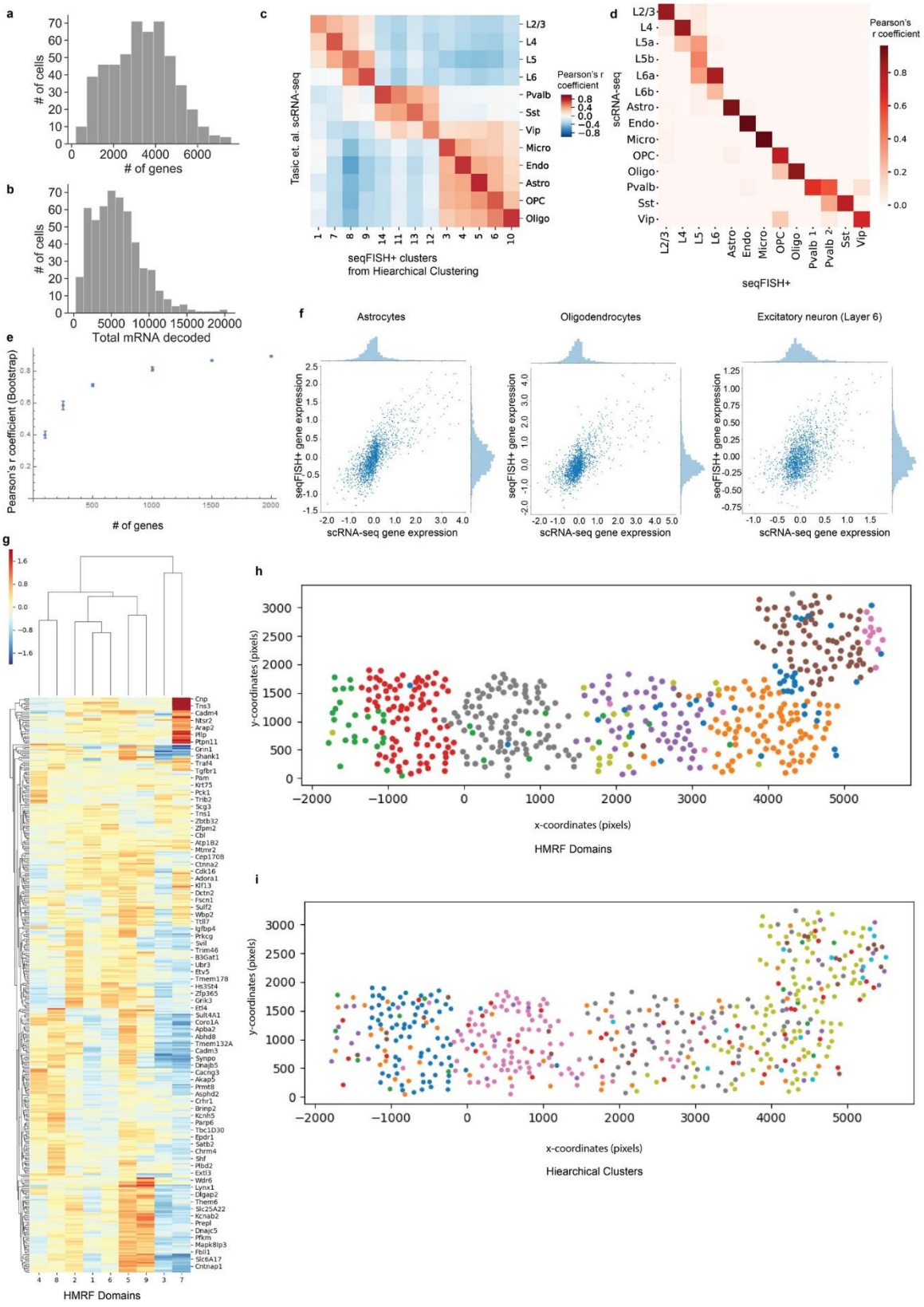
i

Subcluster 1	Subcluster 2	Subcluster 3
Cyr61	Flna	Bicd2
Thbs1	Flnb	Nin
Ctgf	Plec	Pcnt
Serpine1	Flnc	Numa1
Ptx3	Iqgap1	Spdl1
Tnc		Ctif
		Cep350
		Trip11
		Gorasp2

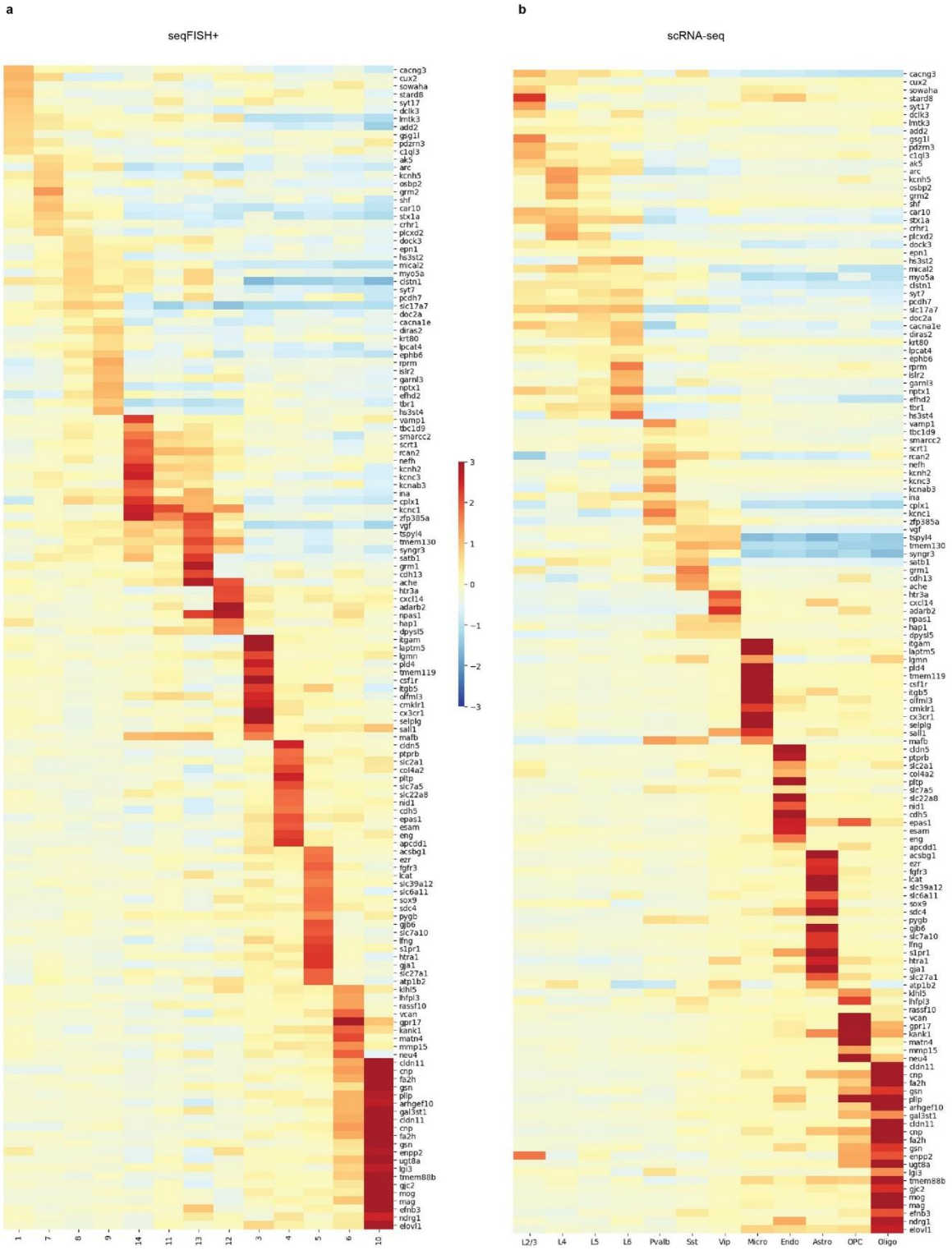
j



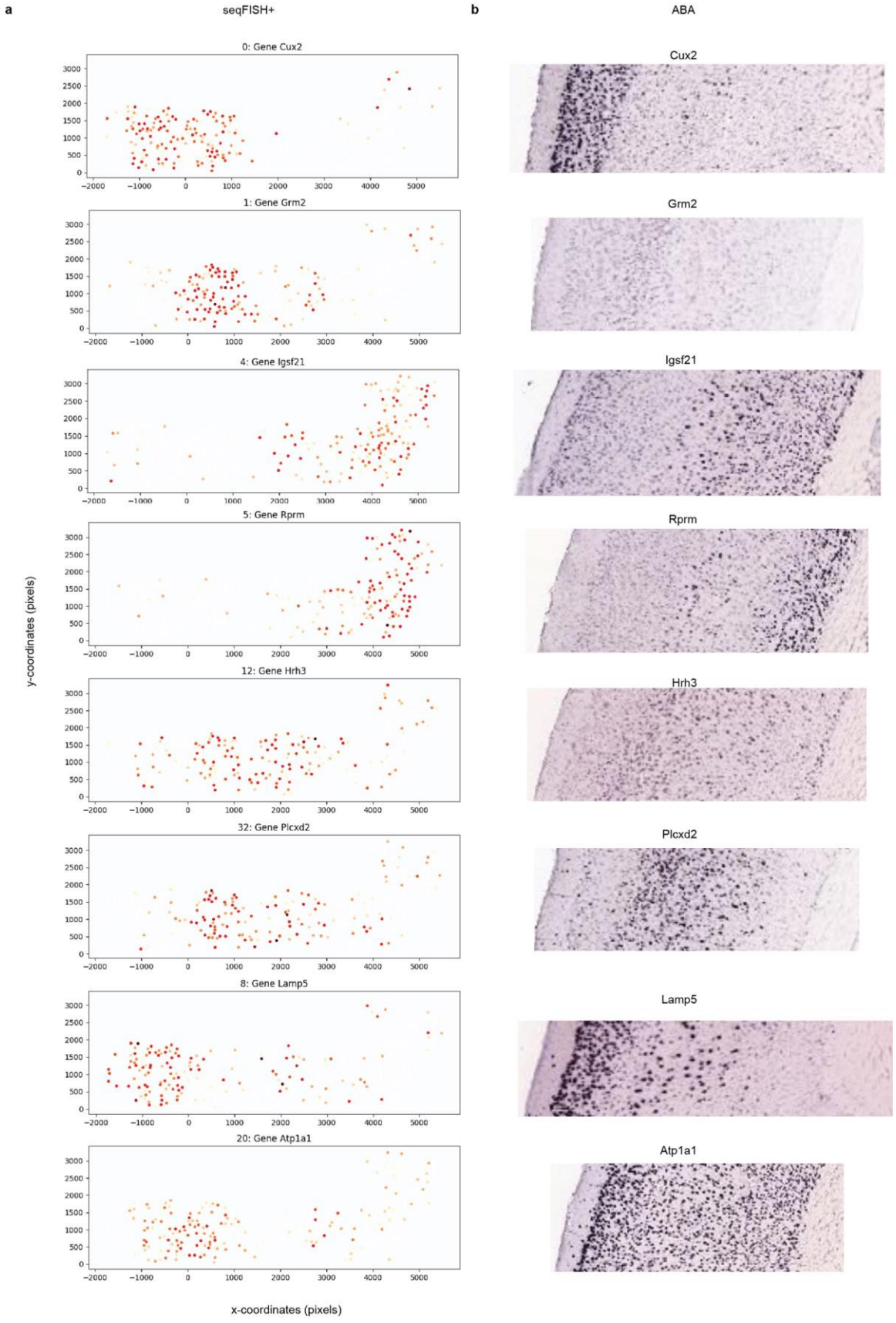
Extended Data Figure 3. seqFISH+ works efficiently across all three fluorescent channels and identifies localization patterns of transcripts in NIH3T3 cells. **a,** Correlation plots between seqFISH+ and bulk RNAseq in three fluorescent channels. Barcodes are coded entirely within each channel, with n = 3334, 3333, and 3333 barcodes in each channel respectively. Barcodes in all channels are decoded and called out efficiently. **b,** seqFISH+ result correlates strongly with RNA SPOTs measurement in NIH3T3 cells. SPM= Spots Per Million. **c,** Correlation between seqFISH+ and smFISH for each fluorescent channel (from left to right: n = 24, 18, 18 genes). All correlations were computed by Pearson's r coefficient correlation with two-tailed p values reported. **d,** The callout frequency of on-target 10,000 barcodes versus the remaining 14,000 off target barcodes. Off target barcodes are called out at a rate of 0.22±0.07 (mean±s.d) per barcode. **e,** Histogram of the total number of mRNAs detected per NIH3T3 cell. On average, 35,492±12,222 transcripts are detected per cell. **f,** Genes are clustered based on their co-occurrence in 10x10 pixel window. Three major clusters are nuclear/perinuclear, cytoplasmic, and protrusions. **g**, mRNAs show preferential spatial localization patterns: nuclear, cytoplasm and protrusion (n = 227 cells). The image is binned into 1 μm x 1 μm windows and colored based on the genes enriched in each bin (scale bar = 10 μm). **h,** Example of genes enriched in each spatial cluster. **i,** Genes in the subclusters within the nuclear localized group. Subcluster 1 contains genes encode for extracellular matrix proteins. Subcluster 2 genes are involved in actin cytoskeleton while subcluster 3 genes are involved in microtubule networks. **j,** Representative smFISH image (single z-slice) of three genes in subcluster 1 shows nuclear/perinuclear localization (n = 20 FOVs, 40x objective). Scale bar: 10μm.


a
# of cells
# of genes
b
# of cells
Total mRNA decoded
c
Tasic et. al. scRNA-seq
seqFISH+ clusters
from Hiearchical Clustering
Pearson's r coefficient
L2/3
L4
L5
L6
Pvalb
Sst
Vip
Micro
Endo
Astro
OPC
Oligo
d
scRNA-seq
seqFISH+
Pearson's r coefficient
e
Pearson's r coefficient (Bootstrap)
# of genes
f
Astrocytes
Oligodendrocytes
Excitatory neuron (Layer 6)
seqFISH+ gene expression
scRNA-seq gene expression
g
HMRF Domains
h
y-coordinates (pixels)
x-coordinates (pixels)
HMRF Domains
i
y-coordinates (pixels)
x-coordinates (pixels)
Hiearchical Clusters

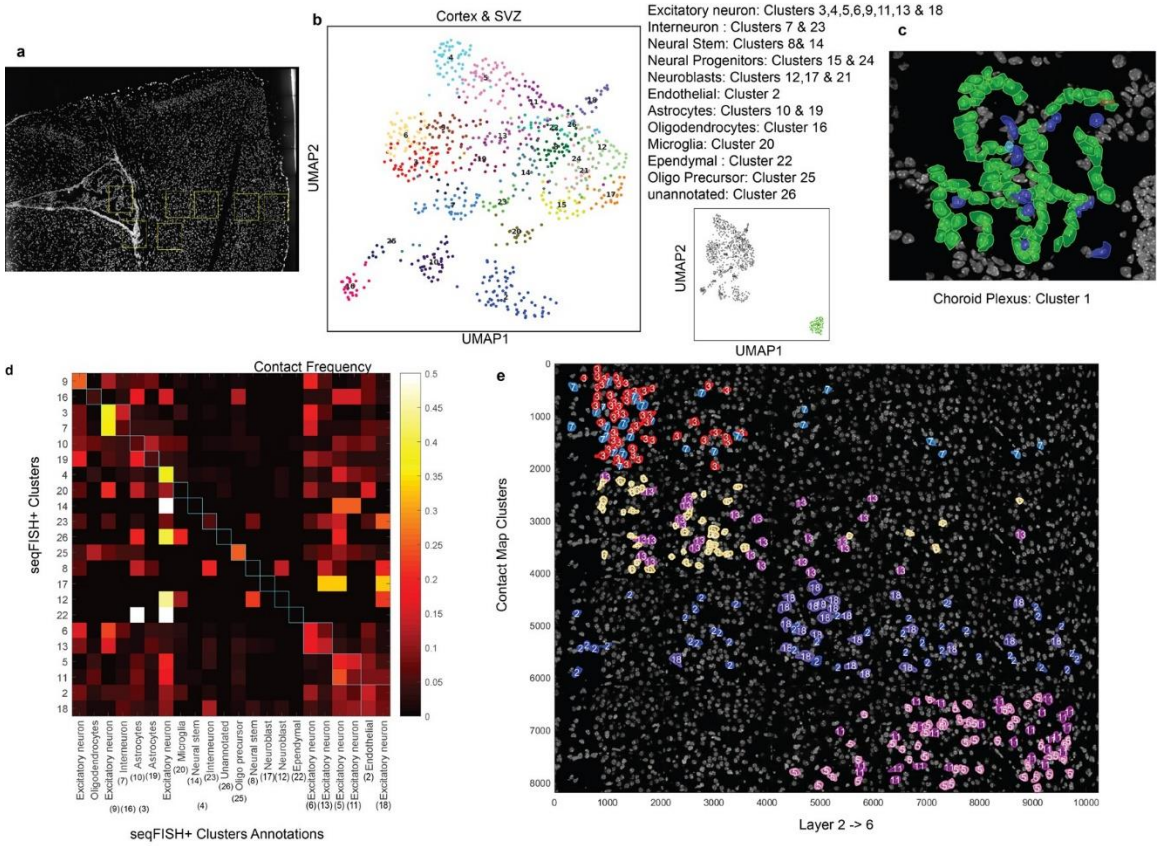
Extended Data Figure 4. scRNAseq comparison with seqFISH+, bootstrap, and HMRF analysis. **a,** Histogram of the number of genes and **b,** total RNA barcodes detected per cell by seqFISH+ in the cortex. **c,** Unsupervised clustering of seqFISH+ correlates well with scRNAseq. (n = 1857 genes; Pearson's r coefficient correlation) **d,** Supervised mapping of seqFISH+ analyzed cortex cell clusters with those from single cell RNA-seq clusters. (n = 1253 genes; p-value < 0.005). **e,** The number of genes were downsampled from the 2511 genes that expressed at least 5 copies in a cell. For each downsampled dataset, the cell-to-cell correlation matrix is calculated and correlated with the cell-to-cell correlation matrix for the 2511 gene dataset. 5 trials are simulated for each downsampled gene level. Error bars denote mean +/- standard deviation. Even when downsampled to 100 genes, about 40% of the cell to cell correlation is retained, because the expression pattern of many genes is correlated. **f,** Scatterplots of seqFISH+ with scRNAseq in different cell types. Each dot represents a gene and their mean expression z-score values in either seqFISH+ or scRNAseq in astrocytes, oligodendrocytes and excitatory neurons. In general, seqFISH+ and scRNAseq are in good agreement (n = 598 genes each). **g,** HMRF detects spatial domains that contain cells with similar expression patterns regardless of cell type. Domain specific genes are shown. **h,** Spatial domains in the cortex. **i,** Mapping of the hierarchical clusters onto the cortex. X-Y coordinates are in pixels (103 nm per pixel). Each camera field of view is 2000 pixels.


a
seqFISH+
b
scRNA-seq
cacng3
cux2
sowaha
stard8
syt17
dclk3
lmtk3
add2
gsg1l
pdzrn3
c1ql3
ak5
arc
kcnh5
osbp2
grm2
shf
car10
stx1a
crhr1
plcxd2
dock3
epn1
hs3st2
mical2
myo5a
cldn1
syt7
pcdh7
slc17a7
doc2a
cacna1e
dirас2
krt80
lpcat4
ephb6
rprm
islr2
garnl3
nptx1
efhd2
tbr1
hs3st4
vamp1
tbc1d9
smarcc2
scrt1
rcan2
nefh
kcnh2
kcnc3
kcnab3
ina
cplx1
kcnc1
zfp385a
vgf
tspyl4
tmem130
syngr3
satb1
grm1
cdh13
ache
htr3a
cxcl14
adarb2
npas1
hap1
dpysl5
itgam
laptm5
lgmn
pld4
tmem119
csf1r
itgb5
olfml3
cmklr1
cx3cr1
selplg
sall1
mafb
cldn5
ptprb
slc2a1
col4a2
pltp
slc7a5
slc22a8
nid1
cdh5
epas1
esam
eng
apcdd1
acsbg1
ezr
fgfr3
lcat
slc39a12
slc6a11
sox9
sdc4
pygb
gjb6
slc7a10
ifng
s1pr1
htra1
gja1
slc27a1
atp1b2
klhl5
lhfpl3
rassf10
vcan
gpr17
kank1
matn4
mmp15
neu4
cldn11
cnp
fa2h
gsn
plp
arhgef10
gal3st1
enpp2
ugt8a
lgi3
tmem88b
gjc2
mog
mag
efnb3
ndrg1
elovl1
3
2
1
0
−1
−2
−3
1
7
8
9
14
11
13
12
3
4
5
6
10
L2/3
L4
L5
L6
Pvalb
Sst
Vip
Micro
Endo
Astro
OPC
Oligo

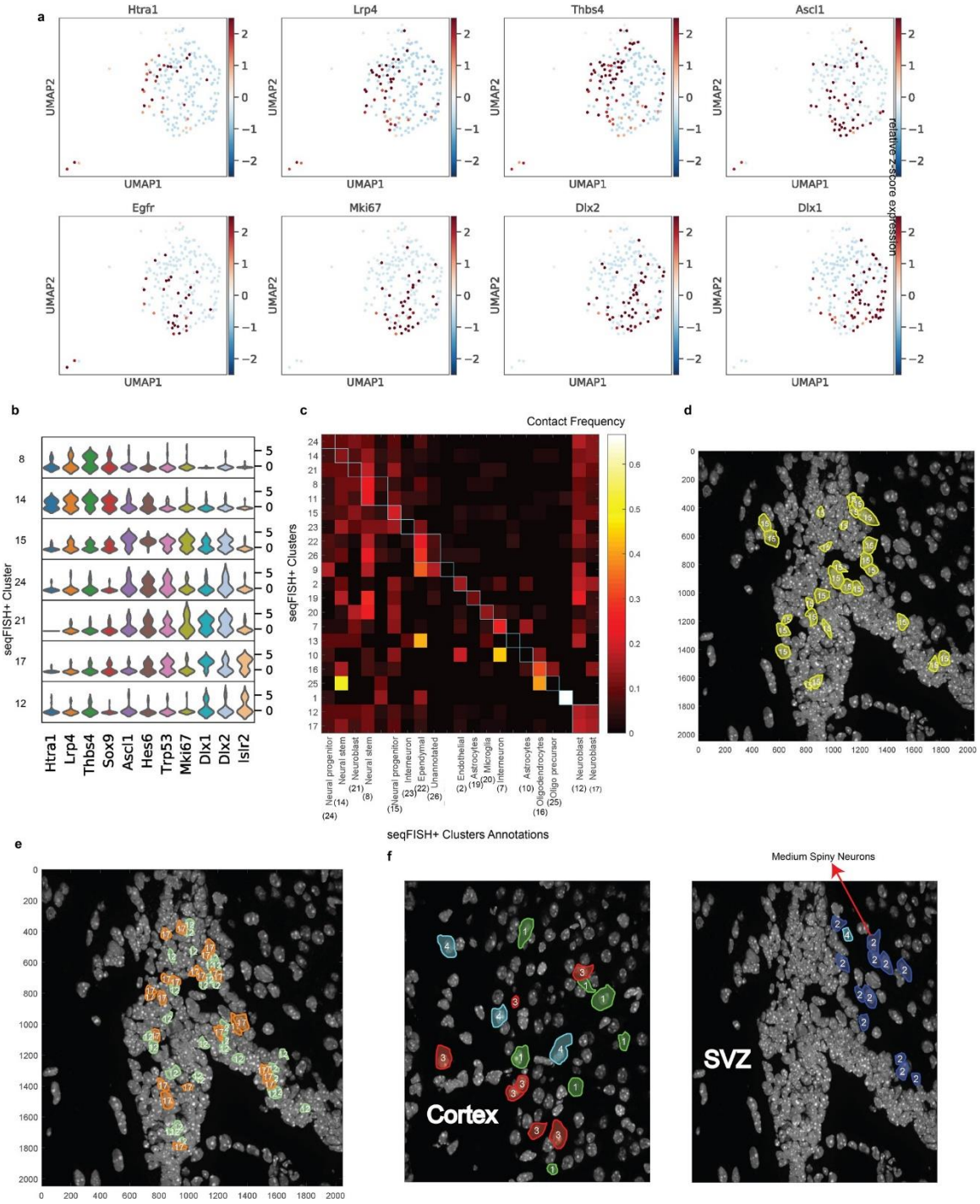
Extended Data Figure 5 (*previous page*). Differential gene expressions between the cell type clusters in both (**a**) seqFISH+ and (**b**) scRNA-seq. The expression patterns of seqFISH+ clusters are similar to scRNA-seq clusters (n = 143 genes)


a
seqFISH+
b
ABA
0: Gene Cux2
1: Gene Grm2
4: Gene Igsf21
5: Gene Rprm
12: Gene Hrh3
32: Gene Plcxd2
8: Gene Lamp5
20: Gene Atp1a1
Cux2
Grm2
Igsf21
Rprm
Hrh3
Plcxd2
Lamp5
Atp1a1
y-coordinates (pixels)
x-coordinates (pixels)

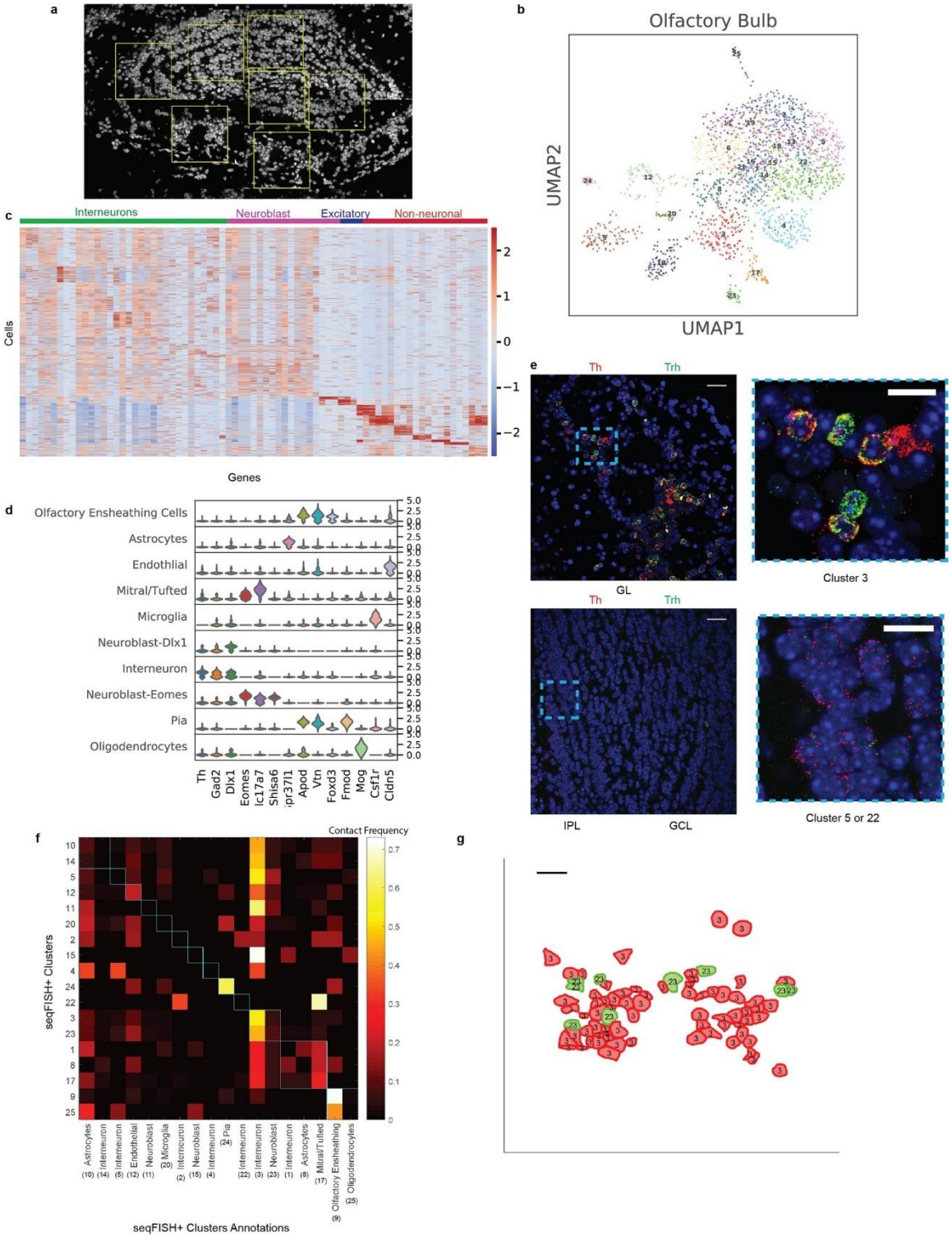
Extended Data Figure 6 (*previous page*). Comparison of the spatial expression patterns across the primary motor cortex in the (**a**) seqFISH+ data versus the (**b**) Allen Brain Atlas. X-Y coordinates are in pixels (103 nm per pixel). Layers I-VI are shown from left to right.


a
b
Cortex & SVZ
UMAP2
UMAP1
Excitatory neuron: Clusters 3,4,5,6,9,11,13 & 18
Interneuron : Clusters 7 & 23
Neural Stem: Clusters 8& 14
Neural Progenitors: Clusters 15 & 24
Neuroblasts: Clusters 12,17 & 21
Endothelial: Cluster 2
Astrocytes: Clusters 10 & 19
Oligodendrocytes: Cluster 16
Microglia: Cluster 20
Ependymal : Cluster 22
Oligo Precursor: Cluster 25
unannotated: Cluster 26
c
Choroid Plexus: Cluster 1
d
Contact Frequency
seqFISH+ Clusters
seqFISH+ Clusters Annotations
e
Contact Map Clusters
Layer 2 -> 6
f
Htra1 Astrocytes
Htra1
g
Atp1b2 Astrocytes
Atp1b2
h
Nr4a1 Excitatory
i
Kif5a Excitatory
Kif5a Interneuron

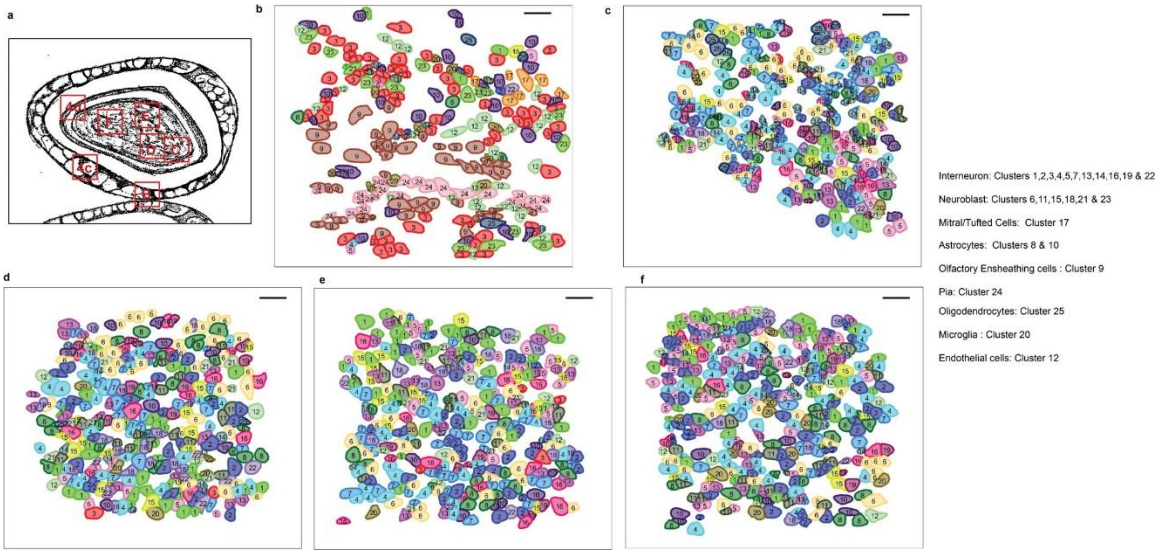
Extended Data Figure 7 (*previous page*). Additional analysis of cortex and subcellular localization patterns in different cell types. **a,** Slide explorer image of the cortex and SVZ FOVs imaged in the first brain slice (n=913 cells). Schematic is shown in Fig 3a. **b,** UMAP representation of cortex and SVZ cells. **c,** Mapping of the choroid plexus cells, which are exclusively present in the ventricle (n =109 cells). **d,** Frequency of contacts between the different cell class in the cortex, normalized for the abundances of cells in each clusters. **e,** Each strip represents cells that cluster together, which breaks into layers in the cortex, consistent with expectation, as cells within a layer preferential interact with each other (n = 523 cells). **f,** *Htra1* transcripts are preferentially localized to the periphery of the astrocytes in the cortex. Left panel shows a reconstructed image from the 10,000 gene seqFISH+ experiment. *Htra1* transcripts are shown in cyan, and all other transcripts are shown in black. Scale bar is 2μm. Middle and right panels show single z-slice of smFISH images of *Htra1* in cortical astrocytes (Scale bar: 5μm). **g,** *Atp1b2* localization in seqFISH+ (left; scale bar: 2μm) and single z-slice smFISH images (middle and right; scale bars: 5μm) . Many *Htra1* and *Atp1b2* transcripts are localized to astrocytic processes (**f,g**, n= 62 astrocytes). SmFISH images were background subtracted for better display of RNA molecules (n= 10 FOVs, 40x objective). **h,** *Nr4a1* localization patterns are distinct from *Htra1* and *Atp1b2* and are more nuclear localized across different cell types. An excitatory neuron is shown from the seqFISH+ reconstructions (n = 337 excitatory neurons; scale bars: 2μm). **i,** *Kif5a*, a kinesin, also exhibits periphery and process localizations in different cell types (n = 60 interneurons; scale bar: 2μm).


a
Htra1
Lrp4
Thbs4
Ascl1
Egfr
Mki67
Dlx2
Dlx1
UMAP1
UMAP2
relative z-score expression
b
seqFISH+ Cluster
Htra1
Lrp4
Thbs4
Sox9
Ascl1
Hes6
Trp53
Mki67
Dlx1
Dlx2
Islr2
c
Contact Frequency
seqFISH+ Clusters
Neural progenitor (14)
Neural stem (24)
Neuroblast (21)
Neural stem (8)
Neural progenitor (15)
Interneuron (23)
Ependymal (22)
Unannotated (26)
Endothelial (2)
Astrocytes (19)
Microglia (20)
Interneuron (7)
Astrocytes (10)
Oligodendrocytes (25)
Oligo precursor (16)
Neuroblast (12)
Neuroblast (17)
seqFISH+ Clusters Annotations
d
e
f
Medium Spiny Neurons
Cortex
SVZ

Extended Data Figure 8 (*previous page*). Additional analysis of the subventricular zone (SVZ). **a,** Expression of individual genes in the SVZ in the UMAP representation (n = 281 cells). **b,** Violin plots denotes z-scored gene expression patterns for Louvain clusters corresponding to NSC to neuroblasts in the SVZ, (n = 281 cells). **c,** Spatial proximity analysis of the cell clusters in the mouse subventricular zone(SVZ). Frequency of contacts between the different cell class in the SVZ, normalized for the abundances of cells in each clusters. **d,** Neural progenitors appear to be in spatial proximity with each other. **e,** Two neuroblasts cell clusters are found to be in spatial proximity in the SVZ (**c-d**, n = 281 cells). **f,** Subclusters of type 7 cells in the cortex (left). Medium spiny neurons that expressed *Adora2, Pde10a,* and *Rasd2* marker genes form a separate cluster that is detected only in the striatum (right) (n = 42 cells in cluster 7).


a
b
Olfactory Bulb
UMAP2
UMAP1
c
Interneurons
Neuroblast
Excitatory
Non-neuronal
Cells
Genes
d
Olfactory Ensheathing Cells
Astrocytes
Endothlial
Mitral/Tufted
Microglia
Neuroblast-Dlx1
Interneuron
Neuroblast-Eomes
Pia
Oligodendrocytes
Th
Gad2
Dlx1
Eomes
Ic17a7
Shisa6
Ipr3711
Apod
Vtn
Foxd3
Fmod
Mog
Csf1r
Cldn5
e
Th
Trh
GL
Cluster 3
IPL
GCL
Cluster 5 or 22
f
Contact Frequency
seqFISH+ Clusters
seqFISH+ Clusters Annotations
g

Extended Data Figure 9. Additional analysis of the olfactory bulb (OB). **a,** Slide explorer image of the OB FOVs imaged in the second brain slice. **b,** UMAP analysis of OB cells. **c,** Z-scored gene expression patterns heatmap of cells in the olfactory bulb. **d,** Violin plots show z-scored marker genes expression patterns in the different classes of cells detected in the OB. (**a-d**, n = 2050 cells) **e,** Representative smFISH images of *Th* and *Trh*. Images were maximum z projected. In the glomeruli layer (GL), cluster 3 cells express both *Th* and *Trh*, whereas in the GCL, only *Th* are expressed (cluster 5 and 22 cells). (n= 10 FOVs, 40x objective). Scale bars: 13µm (left image); 6.5µm (right image). **f,** Frequency of contacts between the different cell class in the glomerulus, normalized for the abundances of cells in each cluster. **g,** Cell clusters #3 (*Th*+ interneurons) and #23 (neuroblast) are in close proximity in the mapped image (**f-g,** scale bars: 20µm).



Extended Data Figure 10. Spatial organization of the olfactory bulb. **a,** Schematics of the field of views imaged in the OB. Spatial mapping of the cell clusters in the Glomerulus Layer (**b**) and Granule Cell Layer (**c-f**) in the OB. Note the neuroblast cells tend to reside in the interior of the GCL (upper parts of **c** and **d** and lower parts of **e** and **f**), whereas more mature interneurons are present in the outer layer. This is consistent with the migration of neuroblasts from the SVZ through the rostral migratory stream into the granule cell layer. Scale bars : 20µm.

Supplementary Table 1. Codebook for 10,000 genes. Base 20 pseudocolor coding scheme for each of the 10,000 genes in the three fluorescent channels.

Supplementary Table 2. Genes enriched in each of the cell clusters identified in the cortex and olfactory bulb data. The top 20 genes in z-score are shown. Cluster annotations are also listed. The same cluster numbers are used in the main and extended data figures.

Supplementary Table 3. mRNA localization patterns in the cortex. Cells are divided up into the annotated clusters. In each cluster, mRNAs that are periphery localized or near nuclear localized are tabulated.

Supplementary Table 4. Ligand-receptor pairs and gene enrichments in neighboring cells. Ligand receptor pairs that are expressed above z-score of 1 are shown in the cortex and the olfactory bulb. p-values are determined from randomly permuting cell labels (n=1000). The enrichment tab shows genes that are expressed more strongly in cluster 1 cells that are neighboring cluster 2 cells than all cluster 1 cells. The expression values are z-scores and p-values are determined from permuting cell labels (n=100).

3.6 Methods

Data Reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments outcome assessment.

Experiment Design

Primary probe design. Gene-specific primary probes were designed as previously described with some modifications[8]. To obtain probe sets for 10,000 different genes, 28-nt sequences of each gene were extracted first using the exons from within the CDS region. For genes that did not yield enough target sequences from the CDS region, exons from both the CDS and UTRs were used. The masked genome and annotation from UCSC were used to look up the gene sequences. Probe sequences were required to fall within the GC content in the range of 45-65%. Any probe sequences that contained five or more consecutive bases of the same kind were dropped. Any genes which do not achieve a minimum number of 24 probes were dropped. A local BLAST query was run on each probe against the mouse transcriptome to ensure specificity. BLAST hits on any sequences other than the target gene with a 15-nt match were considered off targets. ENCODE RNA-seq data across different mouse samples were used to generate an off-target copy number table. Any probe that hit an expected total off-target copy number exceeding 10,000 FPKM was dropped to remove housekeeping genes, ribosomal genes,

and very highly expressed genes. To minimize cross-hybridization between probe sets, a local BLAST database was constructed from the probe sequences and probes with hits of 17-nt or longer were removed by dropping the matched probe from the larger probe set.

Readout probe design. 15-nt readout probes were designed as previously described[9]. Briefly, a set of probe sequences was randomly generated with the combinations of A, T, G, or C nucleotides. Readout probe sequences within a CG content range of 40-60% were selected. We BLAST against the mouse transcriptome to ensure the specificity of the readout probes. To minimize cross-hybridization of the readout probes, any probes with 10-contiguously matching sequences between readout probes were removed. The reverse complements of these readout probe sequences were included in the primary probes according to the designed barcodes.

Primary probe construction. Primary probes were ordered as oligoarray complex pools from Twist Bioscience and constructed as previously described with some modifications[8] . Briefly, limited PCR cycles were used to amplify the designated probe sequences from the oligo complex pool. Then, the amplified PCR products were purified using QIAquick PCR Purification Kit (28104; Qiagen) according to the manufacturer's instructions. The PCR products were used as the template for in vitro transcription (E2040S; NEB) followed by reverse transcription (EP7051; Thermo Fisher) with the forward primer containing a uracil nucleotide[31]. After reverse transcription, the probes were subjected to 1:30 dilution of Uracil-Specific Excision Reagent (USER) Enzyme (N5505S; NEB) treatment to remove the forward primer by cleaving off the uracil nucleotide next to it for ~24 hours at 37°C. Since the reverse complement of T7 sequences was used as the reverse primer, the final probe length in this probe set was ~93-nt. Then, the ssDNA probes were alkaline hydrolyzed by 1 M NaOH at 65°C for 15 minutes to degrade the RNA templates, followed by 1 M acetic acid neutralization. Next, to clean up the probes, we performed ethanol precipitation to remove stray nucleotides, phenol-chloroform extraction to remove protein, and Zeba Spin Desalting Columns (7K MWCO) (89882, Thermo Fisher) to remove any residual nucleotides and phenol contaminants. Then, the probes were mixed with 2 μM of Locked Nucleic Acid (LNA) polyT15 and 2 μM of LNA polyT30 before speed-vac to dry powder and resuspended in primary probe hybridization buffer comprised of 40% formamide (F9027, Sigma), 2x SSC (15557036, Thermo Fisher), and 10% (w/v) Dextran Sulfate (D8906; Sigma). The probes were stored at -20°C until use.

Readout probe synthesis. 15-nt readout probes were ordered from Integrated DNA Technologies (IDT) as 5' amine modified[9]. The construction of readout probe was similar to previously described. Briefly, 5 nmoles of DNA probes were mixed with 25 μg of Alexa Fluor 647 NHS ester or Cy3B or Alexa Fluor 488 NHS ester in 0.5 M sodium bicarbonate buffer containing 10% DMF. The reaction was allowed to go for at least 6

hours at 37°C. Then, the DNA probes were subjected to ethanol precipitation, HPLC purification, and column purification to remove all contaminants. Once resuspended in water, the readout probes were quantified using Nanodrop and a 500 nM working stock was made. All the readout probes were kept at -20°C.

Coverslip functionalization. For cell culture experiment, the coverslips were cleaned with a plasma cleaner at HIGH (PDC-001, Harrick Plasma) for 5 minutes followed by the immersion in 1% bind-silane solution (GE; 17-1330-01) made in pH3.5 10% (v/v) acidic ethanol solution for 30 minutes at room temperature. Then the coverslips were rinsed with 100% ethanol 3 times, and heat-dry in an oven for > 90°C for 30 minutes. Next, the coverslips were treated with 100 μg/uL of Poly-D-lysine (P6407; Sigma) in water for >1 hour at room temperature, followed by rinsing with water three times. The coverslips were then air-dried and kept at 4°C for no longer than 2 weeks. For the mouse brain slices experiment, the coverslips were cleaned by 1M HCl at room temperature for 1 hour, rinsed with water once, and followed by 1M NaOH solution treatment at room temperature for 1 hour. Then, the coverslips were rinsed three times with water, before immersion in 1% bind-silane solution for 1 hour at room temperature. The remaining steps are the same as the coverslip functionalization for cell culture.

seqFISH+ encoding strategy. We separate the 60 pseudocolors into 3 fluorescent channels (Alexa 488, Cy3b and Alexa 647) equally. In each channel, the 20- pseudocolor imaging was repeated 3 times hence achieving 20^3=8000 genes barcoding capacity. We did an extra round of pseudocolor imaging to obtain error-correctable barcodes, an error-correction scheme which we had previously introduced[3]. Thus, we obtained 8000 error-correctable barcodes x 3 fluorescent channels = 24,000 error-correctable barcoding capacity in total. One can easily use more fluorescent channels and/or more pseudocolors to achieve greater dilution of the mRNA density per imaging round. In this experiment, we encoded 3333, 3333, and 3334 genes in each of the fluorescent channels. This pseudocolor scheme evolved from the one used in RNA SPOTs[8] and intron seqFISH[9] by eliminating chromatic aberration and dramatically diluting the density to achieve profiling of mRNA at the transcriptome level *in situ*.

To visualize the different transcripts, 24 “primary” probes were designed against each target mRNA. The primary probes contain overhang sequences that code for the 4-unit base-20 barcode unique to each gene. Hybridization with fluorophore labeled “readout” probes allows the readout of these barcodes and fluorescently labels the subset of genes that contain the corresponding sequences. All of the genes are sampled every 20 rounds of readout hybridization and collapsed into super-resolved images. A total of 80 rounds of hybridizations enumerate the 4-unit barcode for each gene. Each round of stripping and readout hybridization is fast and completed in minutes.

After primary probes hybridization, the samples were subjected to hydrogel embedding and clearing before seqFISH+ imaging. The details are available on *cell culture experiment* , *tissue slices experiment*, and *seqFISH+ imaging*.

Cell culture experiment. NIH/3T3 cells (ATCC) were cultured as previously described[8] on the functionalized coverslips until ~80-90% confluence. Then the cells were washed with 1x PBS once, fixed with freshly made 4% formaldehyde (28906; Thermo Fisher) in 1x PBS (AM9624, Invitrogen) at room temperature for 10 minutes. The fixed cells were permeabilized with 70% ethanol for 1 hour at room temperature. The cell samples were dried and the 10,000 gene probes (~1 nM per probe for 24 probes per gene) were hybridized by spreading out using another coverslip. The hybridization was allowed to proceed for ~36-48 hours in a humid chamber at 37°C. We found hybridization for 48 hours yielded slightly brighter signals. After hybridization, the samples were washed with 40% formamide in 2x SSC at 37°C for 30 minutes, followed by 3 times rinsing with 1 mL 2x SSC. Next, the cell samples were incubated with 1:1000 dilution of Tetraspeck beads in 2x SSC at room temperature for 5-10 minutes. The density of the beads can be easily adjusted by varying the dilution factor or incubation time. Then, the samples were rinsed with 2x SSC and incubated with degassed 4% acrylamide (1610154; Bio-Rad) solution in 2x SSC for 5 minutes at room temperature. To initiate polymerization, the 4% acrylamide solution was aspirated, then 10 μL of 4% hydrogel solution containing 4% acrylamide (1:19), 2x SSC, 0.2% ammonium persulfate (APS) (A3078; Sigma) and 0.2% N,N,N′,N′-Tetramethylethylenediamine (TEMED) (T7024; Sigma) was dropped on the sample, and sandwiched by a coverslip functionalized by GelSlick (Lonza;50640). The polymerization step was allowed to happen at room temperature for 1 hour in a homemade nitrogen gas chamber. After that, the two coverslips were gently separated, and the excess gel was cut away with a razor. A custom-made flow cell (RD478685-M; Grace Bio-labs) was attached to the coverslips covering the region of cells embedded in hydrogel. The hydrogel embedded cell samples were cleared as previously described for >1 hour at 37°C[19]. The digestion buffer consists of 1:100 Proteinase K (P8107S; NEB), 50 mM pH 8 Tris HCl (AM9856; Invitrogen), 1 mM EDTA (15575020; Invitrogen), 0.5% Triton-X 100, and 500 mM NaCl (S5150, Sigma). Then, the samples were rinsed with 2x SSC multiple times and subjected to Label-IT modification(1:10) (MIR 3900; Mirus Bio) at 37°C for 30 minutes. After that, the cell samples were post-fixed with 4% PFA in 1x PBS to stabilize the DNA, RNA, and the overall cell sample for 15 mins at room temperature. The reaction was quenched by 1 M pH8.0 Tris HCl at room temperature for 10 minutes. The cell samples were either imaged immediately or kept in 4x SSC supplemented with 2 U/μL of SUPERase In RNase Inhibitor (AM2696; Invitrogen) at 4°C for no longer than 6 hours.

Animals. All animal care and experiments were carried out in accordance to Caltech Institutional Animal Care and Use Committee (IACUC) and NIH guidelines. Wild-type mice C57BL/6J P23 (male) and P40 (male) were used for the cortex and olfactory bulb seqFISH+ experiments, respectively. For smFISH experiments, adult wild-type mice C57BL/6J aged 10 weeks (female) were used for the RNA localization experiment in the cortex and ligand-receptor interaction experiment in the olfactory bulb. For cell clusters validation in the olfactory bulb, a section from P40 mice was used.

Tissue slices experiment. Brain extraction was performed as previously described[3]. In brief, mice were perfused for 8 minutes with perfusion buffer (10 U/ml heparin, 0.5% NaNO2 (w/v) in 0.1 M PBS at 4°C). Mice were then perfused with fresh 4% PFA in 0.1 M PBS buffer at 4°C 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 4°C 30% RNAse-free Sucrose (Amresco 0335-2.5KG) in 1x PBS until the brain sank. After the brain sank, the brain was frozen in a dry ice of isopropanol bath in OCT media and stored at -80°C. 5 μm sections were cut using a cryotome and immediately placed on the functionalized coverslips. The thin tissue slices were stored at -80°C. To perform hybridization on the tissue slices, the tissue slices were first permeabilized in 70% ethanol at 4°C for >1 hour. Then, the tissue slices were cleared with 8% SDS (AM9822; Invitrogen) in 1x PBS for 30 minutes at room temperature. Primary probes were hybridized to the tissue slices by spreading out the hybridization buffer solution with a coverslip. The hybridization was allowed to proceed for ~60 hours at 37°C. After primary probe hybridization, the tissue slices were washed with 40% formamide at 37°C for 30 minutes. After rinsing with 2X SSC 3 times and 1X PBS once , the sample was subjected to 0.1mg/mL Acryoloyl-X SE (A20770; Thermo Fisher) in 1X PBS treatment for 30 minutes at room temperature. After that, the tissue slices were incubated with 4% acrylamide (1:19 crosslinking) hydrogel solution in 2X SSC for 30 minutes at room temperature. Then the hydrogel solution was aspirated and 20 μL of 4% hydrogel solution containing 0.05% APS and 0.05% TEMED in 2x SSC was dropped onto the tissue slice and sandwiched by Gel-Slick functionalized slide. The samples were transferred to 4°C in a homemade nitrogen gas chamber for 30 minutes before transferring to 37°C for 2.5 hours to complete polymerization. After polymerization, the hydrogel embedded tissue slices were cleared with digestion buffer as mentioned above , except it includes 1% SDS, for >3 hours at 37°C. After digestion, the tissue slices were rinsed by 2X SSC multiple times and subjected to 0.1mg/mL Label-X modification for 45 minutes at 37°C. The preparation of Label-X stock was as previously described[19]. To further stabilize the DNA probes, RNA molecules, and the tissue slices overall structure, the tissue slices were re-embedded in hydrogel solution as the previous step, except the gelation time can be shortened to 2 hours. The tissue slice samples were

either imaged immediately or kept in 4X SSC supplemented with 2 U/μL of SUPERase In RNase Inhibitor at 4°C for no longer than 6 hours.

seqFISH+ Imaging. Imaging platform and automated fluidics delivery system were similar to those previously described with some modifications. In brief, the flow cell on the sample was first connected to the automated fluidics system. Then the region of interests(ROI) was registered using nuclei signals stained with 10 μg/mL of DAPI (D8417; Sigma). For cell culture experiments, blank images containing beads only were first imaged before the first round of serial hybridization. Each serial hybridization buffer contained three unique sequences with different concentrations of 15-nt readouts conjugated to either Alexa Fluor 647(50 nM) , Cy3B(50 nM), or Alexa Fluor 488(100 nM) in EC buffer made from 10% Ethylene Carbonate (E26258; Sigma), 10% Dextran Sulfate (D4911; Sigma) , 4X SSC and 1:100 dilution of SUPERase In RNase Inhibitor. The 100 μL of serial hybridization buffers for 80 rounds of seqFISH+ imaging with a repeat for round 1 (in total 81 rounds) were pipetted into a 96 well-plate. During each serial hybridization, the automated sampler will move to the well of the designated hyb buffer and flow the 100 μL hyb solution through a multichannel fluidic valves (EZ1213-820-4; IDEX Health & Science) to the flow cell (required ~25 μL) using a syringe pump (63133-01, Hamilton Company). The serial hyb solution was incubated for 17 minutes for cell culture experiments and 20 minutes for tissue slice experiments at room temperature. After serial hybridization, the sample was washed with ~300 μL of 10% formamide wash buffer (10% formamide and 0.1% Triton X-100 in 2X SSC) to remove excess readout probes and non specific binding. Then, the sample was rinsed with ~200 μL of 4X SSC supplemented with 1:1000 dilution of SUPERase In RNase Inhibitor before stained with DAPI solution (10 μg/mL of DAPI, 4X SSC, and 1:1000 dilution of SUPERase In RNase Inhibitor) for ~15 seconds. Next, an anti-bleaching buffer solution made of 10% (w/v) glucose, 1:100 diluted catalase (Sigma C3155), 0.5 mg/mL Glucose oxidase (Sigma G2133) , 0.02 U/μL SUPERase In RNase Inhibitor , 50 mM pH8 Tris-HCl in 4x SSC was flowed through the samples. Imaging was done with the microscope (Leica, DMi8) equipped with a confocal scanner unit (Yokogawa CSU-W1), a sCMOS camera (Andor Zyla 4.2 Plus), 63 × oil objective lens (Leica 1.40 NA), and a motorized stage (ASI MS2000). Lasers from CNI and filter sets from Semrock were used. Snapshots were acquired with 0.35 μm z steps for two z slices per FOV across 647-nm, 561-nm, 488-nm and 405-nm fluorescent channels. After imaging, stripping buffer made from 55% formamide and 0.1% Triton-X 100 in 2x SSC was flowed through for 1 minute, followed by an incubation time of 1 minute before rinsing with 4X SSC solution. In general, the 15-nt readouts were stripped off within seconds, and a 2-minute wash ensured the removal of any residual signal. The serial hybridization, imaging, and signal extinguishing steps were repeated for 80-rounds. Then, stainings buffer for segmentation

purposes consists of 10 μg/mL of DAPI, 50nM LNA T20-Alexa 647, and 1: 100 dilution of Nissl stainings (N21480; Invitrogen) in 1x PBS was flowed in and allowed to incubate for 30 mins at room temperature before imaging. The integration of automated fluidics delivery system and imaging was controlled by a custom written script in Micro-Manager[32]

smFISH. Single molecule FISH (smFISH) experiments were done as previously described[8]. In brief, 60 genes were randomly chosen from the 10,000 gene list across a broad range of expression levels. The same probe sequences were used for these 60 genes, except each primary probe contained two binding sites of the readout probes. The fixed cells were hybridized with the primary probes(10nM/probes) in 40% hyb buffer(40% formamide, 10% Dextran Sulfate and 2x SSC) at 37°C for overnight. The sample was washed with 40% wash buffer for 30 minutes at 37°C and subjected to the same hydrogel embedding and clearing as the cell culture experiment before imaging. The imaging platform is the same as the one in seqFISH+ experiment. A single z-slice across hundreds of cells was imaged and the sum of the gene counts per cell was analyzed by using a custom written Matlab script. For smFISH experiments in the tissue, sample was hybridized with 10nM/probe in 40% hyb buffer at 37°C for >16 hours. The sample was washed with 40% wash buffer for 30 minutes at 37°C and subjected to the same hydrogel embedding and clearing as the tissue experiment before imaging. Since the imaging time is short, the Acryoloyl-X functionalization and post hydrogel anchoring steps were omitted. 5 z-slices with z-step of 1μm were taken across multiple FOVs with the imaging platform in the seqFISH+ experiment, except a 40x oil objective was used (Leica 1.40 NA). Images were background subtracted and maximum z-projected for clearer display of RNA dots.

Image Analysis

All image analysis was performed in Matlab. Unless a specific Matlab function is referenced, custom code was used.

Image Registration. Each round of imaging included imaging with the 405-nm channel which included the DAPI stain of the cell along with imaging in the 647-nm, 561-nm and 488-nm channels of TetraSpeck beads' (T7279, Thermo Fischer) and seqFISH+ probes. In addition, a pre-hybridization image was used to find all beads before the readouts were hybridized. Bead locations were fit to a 2D Gaussian. An initial estimate of the transformation matrix between the DAPI image for each serial hybridization round and the only beads image was found using imregcorr (Matlab). Using this estimate transformation, the bead coordinates were transformed to each serial hybridization image,

where the location of the bead was again fit to a 2D Gaussian. A final transformation matrix between each hybridization image and the only-beads image was then found by applying fitgeotrans (Matlab) to the sets of Gaussian fit bead locations. For the tissue samples no beads were used and registration was based on DAPI alone.

Image processing. Each image was deconvolved, using a bead (7x7pixels) as an estimate for the point spread function. Cell segmentation was performed manually using ImageJ's ROI tool.

Barcode Calling. The potential RNA signals were then found by finding local maxima in the image above with a predetermined pixel threshold in the registered and deconvolved images. Dot locations were then further resolved using radialcenter.m [33]. Once all potential points in all serial hybridizations of one fluorescent channel were obtained, they were organized by pseudocolor and barcoding round. Dots were matched to potential barcode partners in all other pseudo channels of all other barcoding rounds using a 1 pixel search radius (or for the tissue samples a 1.4 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 barcoding round 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 areas and transcript numbers were assigned based on the number of on-target barcodes present in the cell. Centroids for each called barcode were also recorded and assigned to cells. The same procedure was repeated for 647, 561 and 488 channels. The remaining unused barcodes were used as an off-target evaluation by repeating the same procedure as described.

Data Analysis

RNA-seq/RNA SPOTs. Pearson's r correlation was performed to compare seqFISH+ data to RNA-seq (GEO: GSE98674), RNA SPOTs[8], and smFISH measurement using Matlab or Python function.

Spatial clustering of genes for NIH3T3 cells. The same barcode calling procedure described above was repeated without cell segmentation to remove the possibility of clipping potentially interesting regions of the cell. RNA locations were coarse grained to 10x10 pixels, resulting in a matrix of dimension total number of coarse grained pixels by the number of genes. Coarse pixels with no RNA were removed from the analysis. RNA with fewer than 10 copies per field of view were dropped. Genes were then correlated with Pearson's r correlation and hierarchical clustering was performed on the resulting correlation matrix. Clusters of less than 10 genes were dropped.

Hierarchical clustering of brain seqFISH+ data. The 10,000 genes were divided into 3 approximately equal subsets (with 3334, 3333, and 3333 genes, respectively) based on the group in which genes are barcoded. Genes were normalized separately within each subset, by dividing the gene counts in per cell by the total counts per cell within each subset. We then multiplied the result by the scaling factor of 2,000 which is approximately the median count. Next, we selected the subset of cells that were in the motor cortex. We computed log(1+normalized counts).

To select genes for clustering, we first computed statistics for the following criteria for each gene: 1) number of cells with nonzero expression, 2) average gene expression of all cells, 3) average expression of top 5% cells with highest expression, 4) average of top 10% cells with highest expression, 5) average of top 2% cells with highest expression, and 6) average gene expression of all nonzero cells. For each criterion, we selected the top 25% of genes that were ranked based on the criterion. We next obtained the union of all 6 gene lists, forming an initial 3877 gene-set. The reasoning is that the union of genes would contain both genes needed to cluster common cell types (which would be expressed in a large population of cells, captured by criterion 2) and rare cell types (which would be expressed in a small population, captured by criteria 3, 4, and 5). The 3877-gene expression data matrix was next transformed by z-scoring per cell and per gene. Principle component analysis (PCA) was performed and jackstraw procedure was adopted in order to further select the most relevant genes for clustering. Specifically, the jackstraw procedure[34] permutes the expression of a small number of genes in order to identify significant genes with significantly higher loading than permuted case ($P<0.001$). Using the top 9 components, we found a total of 1916 significant genes to be used for final clustering.

To this 1916-gene matrix we applied hierarchical clustering with Ward's linkage and with (1 - Pearson correlation) as the distance measure. Using the sigClust R package[35], which evaluates the significance of each branching in the dendrogram, we found significant tree

splits and produced 10-cluster and 16-cluster annotations corresponding to different cluster granularity. Each split was significant according to sigClust FWER corrected P < 0.05. We further performed an additional round of clustering within the interneuron annotated clusters, repeated gene-selection procedure, and replaced the broad interneuron cluster with the subclusters. All together, we derived 13-cluster and 18-cluster annotations.

Unsupervised comparison with scRNAseq data. Mouse visual cortex scRNAseq data was obtained from Tasic et al[25]. We used the cell-type annotations from the original study, representing 9 major, 22 fine, and 49 minor cell-types. For comparison, we focused on the 1857 genes that were commonly profiled by scRNAseq and seqFISH+ and processed the scRNAseq data in the same way as seqFISH+. The degree of similarity was evaluated by using the Pearson correlation (Extended Data Figure 4a).

Supervised mapping of cell types from scRNAseq to seqFISH+. Cell-type mapping was done as described before[14]. Briefly, MAST[36] was used to identify differentially expressed genes across annotated cell types in Tasic *et. al*. scRNA-seq dataset, using P=0.005 as the cutoff. 1253 of the differentially expressed genes were also profiled by seqFISH+ therefore retained for cell-type mapping. Then, we performed a quantile-normalization on the expression vectors of each gene in both the seqFISH+, scRNA-seq data to normalize cross-platform differences[14]. Multi-class support-vector machine models were trained on the scRNAseq cell types using linear kernels, and setting the tuning parameter C to 1e-5, shown in Figure 3g. The cross-validation accuracy of prediction of the 22 annotated cell types was 91% with these 1253 differentially expressed genes.

Spatial gene identification. Briefly, we computed a spatial score per gene as previously described[14]. Cells were divided into two sets based on gene g: L1, contains cells with highest 90th percentile by expression, and L0, the remaining cells. The spatial score measures whether the cells in L1 are spatially adjacent to each other and is quantified by the silhouette coefficient. The silhouette coefficient was computed using the calc_silhouette_per_gene() function in the smfish Hmrf Python package [14] (https://bitbucket.org/qzhudfci/smfishhmrf-py), setting dissimilarity matrix to rank-transformed Euclidean distance, examine_top=0.1, permutation_test=True, and permutations=1000. Rank-transformed distance was computed with rank_transform_matrix() function with reverse=False, rbp_p=0.99 where rbp_p is a rank-weighting parameter. We select all spatial genes with significant silhouette coefficient (P<0.01 permutation test). To further enrich for spatial signals within these genes, we performed a PCA analysis and then jackstraw procedure[34] to arrive at a set of 988 spatial

genes significantly correlated to the principle components. We performed HMRF analysis on the top 100, 200, and 400 of 988 genes.

Spatial domain identification via HMRF procedure. HMRF is a probabilistic spatial clustering method that we developed previously to identify spatial domains based on spatial relationships and gene expression per cell. We constructed a neighborhood graph by adopting a fixed radius corresponding to top 1-percentile of pairwise physical distances between cells, resulting in an average of 5 neighbors per cell. HMRF was run with the following parameters: tolerance=1e-10, k=9, and convergence_error=1e-8. To search for an optimal value of beta, we scanned through all integer values between 2 and 100 and ran the HMRF model for each setting. The value that resulted in minimal change of log-likelihood was selected as the final beta.

Louvain clustering. Unless specified, all functions of pre-processing and Louvain clustering was performed in Python using the package SCANPY[37]. We followed a standard procedure as suggested in the SCANPY reimplementation of Seurat's tutorial to analyze seqFISH+ data with some modifications. For clustering all cells from mouse cortex, subventricular zone(SVZ) , choroid plexus, and olfactory bulb, we first normalize the counts per cell, and then choose highly variable genes with >0.4 min_dispersion, 0.01 min_mean, with max_mean =3. This yields 3509 genes. Then we take the logarithm of the data, regress out the total count effect per cell and scale the data to unit variance. We compute the PCAs and using top PCs to compute the neighborhood graph before performing Louvain clustering. We use the rank_gene_groups function with raw data and the top 20 genes enrichment in each cluster were used to identify the clusters based on marker genes annotation from single cell RNA-seq / DropSeq data[29,38]. We found that both Hierarchical clustering and Louvain clustering yield similar results despite different methods.

To spatially map back the clusters on the raw image, we perform Louvain clustering on cortex, SVZ, and choroid plexus data, and olfactory bulb data separately. Genes with max count greater than 4 across all cells were chosen for cortex and SVZ (include choroid plexus cells) data. Next, we filtered out cells with less than 200 genes expressed from analysis. The counts were normalized per cell and a minimum dispersion of greater than 0 with min_mean of 0.05 were chosen to filter out the variable genes. This yield 1813 genes for subsequent analysis. For the olfactory bulb, genes with max count greater than 2 across all cells were first chosen. Then the counts were normalized per cells. To obtain the highly variable genes, a threshold of min_mean=0.05, and min_dispersion of 0.2 were chosen. This yields 1972 genes for subsequent analysis. After choosing the highly variable genes, the data was subjected to PCA reduction, computed neighborhood graph

with top PCs, and Louvain clustering. The top 20 enrichment genes were obtained using rank_genes_groups function and the clusters were identified according to published literature. Sub-clustering of the main cluster was performed by repeating the process described above. The visualization of these clusters to two dimensions using Uniform Manifold Approximation and Projection (UMAP) was done with SCANPY function. These cluster numbers were mapped back to the original data to visualize the spatial heterogeneity of different cell types across different part of the tissues.

Calculation of the time acceleration of seqFISH+ vs expansion seqFISH. For expansion seqFISH, we assume that to code ~20,000 genes, the coding scheme is with 4 colors and 8 rounds of hybridization (4^7=16,384 genes) with 1 round of error correction. Thus, the total number of effective imaging per field of view (FOV) is equal to the expansion factor x 4 x 8. For 60-fold expansion, this is 60x4x8=1920 images. For seqFISH+, we assume a coding scheme with 3 separate fluorescent channels, with 8000 genes coded in each channel for a total of 24000 genes. Pseudocolors are used to code for 8000 genes. For example, if the number of pseudocolors is 20 per fluorescent channel, then 4 rounds of barcoding (including 1 round of error correction) is need. The effective imaging per FOV is then 20 x 4 x 3 =240 images, a 8-fold acceleration compared to expansion seqFISH. As another example, if the pseudocolor per channel is 10, then 5 rounds of barcoding is need to cover 8000 genes per channel. Then a total of 10 x 5 x 3 =150 images. However, this coding scheme only provides 10 x 3= 30 fold decrease in the RNA density. If an equivalent of 30-fold expansion was implemented, then 30 x 4 x 8 = 960 images are needed per FOV for an acceleration rate of 960/150 = 6.4 fold.

Bootstrap analysis. We calculate the cell-to-cell correlation matrix with the number of genes were downsampled from the 2511 genes that expressed at least 5 copies in a cell. For each downsampled dataset, 100, 250, 500, 1000, 1500, and 2000 genes were selected randomly. The Pearson's correlation coefficient of each of the cell-to-cell correlation matrix is computed with the cell-to-cell correlation matrix for the 2511 gene dataset. 5 trials are simulated for each downsampled gene level. Error bars denote standard deviation.

Neighbor cell analysis. The spatial coordinates for the cell centroids were used to create a nearest neighbor network (k = 4), whereby nodes represent individual cells and edges are observed proximities between 2 cells. Edges between identical or different annotated cell types were respectively labeled as homo- and heterotypic. To identify enriched or depleted proximities between two identical or different cell types the observed number of edges between any two cell types was compared to a random permutation (n = 100)

distribution by reshuffling the cell labels. Associated p-values were calculated by observing how often the simulated values were higher or lower as the observed value for respectively enriched or depleted proximities.

Gene expression enrichment for cell types in close proximity was calculated as the average expression for that gene in all the cells of these two cell types that were in close proximity according to the spatial network. The number of observed edges between two cell types and z-scores for each gene were further used to filter and identify enriched gene expression in any combination of two proximal cell types.

To determine the ligand-receptor pairs in neighboring cells, we extracted genes that have z-scores of 1 or greater, are expressed in at least 25% of the cells in the interacting pairs, and have at least 4 or more instances of being neighbors. We then match up the ligand-receptor pairs from literature[39], which is shown in Supplementary Table 4. To identify statistically enriched ligand-receptor pairs we compared the calculated ligand-receptor scores with that of a random permutation (n = 1000) distribution by reshuffling the cell labels. p-value < 0.05 is deemed to be significant.

RNA localization analysis. To determine the subcellular localization patterns of mRNAs in the cortex, all cells are first separated into the 26 cell clusters (Extended Data Table 2). Within each cell class, the top 200 highly expressed genes are selected for localization analysis. In each cell, the average distance of all of the transcripts for each of the 200 genes from the center of the mass of all of the transcripts for all the genes are calculated. This metric corresponds to whether the gene is likely to be found close or far from the cell center. Only cells with 4 or more copies of that RNA are included in the calculation. The average distance from the center for each cell is normalized by the size of the cell, determined as the square root of the area span by the convex hull of all the mRNA dots in that cell. To select the genes that are localized far from the center of the cell, a threshold of 0.45 for the localization score is used and the average expression level is set at greater than 2.5 copies detected per cell. We selected genes that are close to the cell center using a localization score of 0.35 or lower and the expression level of greater than 2.5 copies per cell. The results are shown in Supplementary Table 3.

Contact maps. The minimum distance between the pixels defining the edge of all pairs of cells in a field of view were tabulated. To count the number of times cells of each type were in contact with cells of each other type, the following procedure was followed. Cells within 15 pixels of a given cell were considered in contact, and the appropriate entry in a square matrix of length equal to the number of cell types was incremented. The counts

were then normalized such that each row sums to 1. Hierarchical clustering was then performed to cluster cell types.

Code Availability. The custom written scripts used in this study are available at https://github.com/CaiGroup/seqFISH-PLUS

Data Availability. RNA-seq data were obtained from GEO accession number GSE98674. RNA SPOTs data were obtained from a previous study[8]. Source data from this study are available at https://github.com/CaiGroup/seqFISH-PLUS. All data obtained during this study are available from the corresponding author upon reasonable request.

3.7 References

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

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

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

4. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).

5. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).

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

7. 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).

8. Eng, C.-H. L., Shah, S., Thomassie, J. & Cai, L. Profiling the transcriptome with RNA SPOTs. *Nat. Methods* **14**, 1153–1155 (2017).

9. Shah, S. *et al.* Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell* **174**, 363–376.e16 (2018).

10. Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857 (2013).

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

12. Betzig, E. *et al.* Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).

13. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–795 (2006).

14. Zhu, Q., Shah, S., Dries, R., Cai, L. & Yuan, G.-C. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4260

15. Thompson, R. E., Larson, D. R. & Webb, W. W. Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* **82**, 2775–2783 (2002).

16. Yildiz, A., Tomishige, M., Vale, R. D. & Selvin, P. R. Kinesin walks hand-over-hand. *Science* **303**, 676–678 (2004).

17. Chen, F., Tillberg, P. W. & Boyden, E. S. Optical imaging. Expansion microscopy. *Science* **347**, 543–548 (2015).

18. Yang, B. *et al.* Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell* **158**, 945–958 (2014).

19. Chen, F. *et al.* Nanoscale imaging of RNA with expansion microscopy. *Nat. Methods* **13**, 679–684 (2016).

20. Moffitt, J. R. *et al.* High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14456–14461 (2016).

21. Antebi, Y. E. *et al.* Combinatorial Signal Perception in the BMP Pathway. *Cell* **170**, 1184–1196.e24 (2017).

22. Mili, S., Moissoglu, K. & Macara, I. G. Genome-wide screen reveals APC-associated RNAs enriched in cell protrusions. *Nature* **453**, 115–119 (2008).

23. Wang, T., Hamilla, S., Cam, M., Aranda-Espinoza, H. & Mili, S. Extracellular matrix stiffness and cell contractility control RNA localization to promote cell migration. *Nat. Commun.* **8**, 896 (2017).

24. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* **3**, 861 (2018).

25. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).

26. Shah, P. T. *et al.* Single-Cell Transcriptomics and Fate Mapping of Ependymal Cells Reveals an Absence of Neural Stem Cell Function. *Cell* **173**, 1045–1057.e9 (2018).

27. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

28. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).

29. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999–1014.e22 (2018).

30. Takei, Y., Shah, S., Harvey, S., Qi, L. S. & Cai, L. Multiplexed Dynamic Imaging of Genomic Loci by Combined CRISPR Imaging and DNA Sequential FISH. *Biophys. J.* **112**, 1773–1776 (2017).

31. Wang, G., Moffitt, J. R. & Zhuang, X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci. Rep.* **8**, 4847 (2018).

32. Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. Computer control of microscopes using μManager. *Curr. Protoc. Mol. Biol.* **Chapter 14**, Unit14.20 (2010).

33. Parthasarathy, R. Rapid, accurate particle tracking by calculation of radial symmetry centers. *Nat. Methods* **9**, 724–726 (2012).

34. Chung, N. C. & Storey, J. D. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* **31**, 545–554 (2015).

35. Huang, H., Liu, Y., Yuan, M. & Marron, J. S. Statistical Significance of Clustering using Soft Thresholding. *J. Comput. Graph. Stat.* **24**, 975–993 (2015).

36. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

37. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

38. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015–1030.e16 (2018).

39. Ramilowski, J. A. *et al.* A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).