Mapping Cell Types, Dynamics, and Connections in Neural Circuits

by

Samuel Gordon Rodriques

B.Sc. in Physics
Haverford College, 2013

M.Phil. in Engineering
Churchill College, University of Cambridge, 2014

Submitted to the Department of Physics in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Physics

at the

Massachusetts Institute of Technology

June 2019

Signature of Author: _____

Samuel G. Rodriques

Certified by: _____

Jeff Gore, Associate Professor of Physics, Supervisor

Accepted by: _____

Nergis Mavalvala, Associate Department Head

Mapping Cell Types, Dynamics, and Connections in Neural Circuits

by

Samuel Gordon Rodriques

Submitted to the Department of Physics on May 17th, 2019, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physics.

Abstract

Neuroscience is limited by the difficulty of recording neural activity, identifying cell types, and mapping connectivity in high throughput. In this thesis, I present several scalable technologies aimed at improving our ability to characterize the activity, composition, and connectivity of neural circuits. My primary contributions include the design for a nanofabricated electrical recording device and a new approach to nanofabrication within swellable hydrogels; a high-throughput method for mapping the locations of cell types in tissue; an approach to direct sequencing of proteins at the single molecule level; an approach to directly recording neural activity into the sequence of RNA, enabling it to be detected by DNA sequencing; and a method for molecular barcoding of neurons, with the goal of enabling a high-throughput approach to neural circuit mapping. I conclude with a consideration of the limitations of the academic incentive structure as concerns the development and deployment of new technologies, and propose a structure for basic science research, complementary to the academic structure, based on the systematic establishment of well-funded, highly focused research projects with clear goals, an incentive to rapidly disseminate information, and limited lifetimes.

Thesis Supervisor: Jeff Gore
Title: Associate Professor of Physics

Thesis Co-supervisor: Edward S. Boyden
Title: Associate Professor of Media Arts and Sciences

Contents

Acknowledgements

I have completed my Ph.D. only with the assistance and generosity of a number of people who I have had the pleasure to work with. Foremost among these is Ed Boyden, who kept me motivated through years of difficult work with the promise that we would eventually uncover a path to revolutionize neuroscience. We just might have.

Adam Marblestone has stood out as one of my primary influences, participating in one way or another in the origination of 5 of the 6 ideas presented here. Adam's thesis is titled "Designing Scalable Biological Interfaces," and it briefly occurred to me that an appropriate name for my thesis would be "Implementing Scalable Biological Interfaces." I feel extraordinarily lucky to have had the opportunity to work with him. I hope that I will have another such opportunity soon.

Fei Chen was a guiding light to me throughout my graduate school career, and was kind enough to take me under his wing in late 2016. I have learned an extraordinary amount of science from him in the years since. Fei is one of the most creative people I know, with a voluminous knowledge of molecular biology, a true 21st-century tinkerer. I look forward to continuing to work with him in the future, and to seeing what wild ideas he comes up with next.

I had the tremendous good fortune to work with Evan Macosko through my work on Slide-seq. Evan has vision and a rare clarity of thought, and has served as a mentor to me for the past two years. I greatly appreciate all of his help and guidance.

My graduate studies have been supported through the Hertz Fellowship, and I reflect often on how lucky I am to have been selected as one. The Hertz Foundation is run by a small group of administrators and donors with an intense, collective devotion to the foundation's mission: to provide scientists and engineers in training with the *freedom to innovate*. Somehow, through an exceptional selection process and sheer determination, they have succeeded in making nearly all Hertz fellows feel part of a community dedicated to innovation, and it is the sense of belonging to this community, more than anything else, that has made the difference in my graduate career. More times than I can count, I have been confronted with the possibility that one of my mentors would disapprove of my decision to pursue a particular project or line of reasoning, and have simply said to myself, "well who cares what they think? I'm a *Hertz fellow*, and the mission is *freedom to innovate.*" I thank all of the members of the foundation and the major donors, Robbee Kosak, Jay Davis, Tom Weaver, Lowell Wood, David Galas, Philip Welkhoff, Kathy Young, Mandy O'Connor, Linda Swift, Linda Kubiak, Ray Sidney, Harold Newman, Louis Lerman, and many others, for putting that reminder in the back of my mind and emboldening me to pursue the best science I possibly could. Chief among those, I owe a tremendous debt of gratitude to Rosemarie Havranek and Nathan Myhrvold, who funded my Hertz fellowship. I am particularly lucky to count Rosemarie as a friend – she is a remarkable woman, and I am extremely grateful for her generosity.

In addition, my studies have been supported by the National Science Foundation, through the generosity of the American taxpayers. In times of such intense national division, it can be easy to lose sight of the fact that millions of Americans still pool their resources to support science, engineering, and education, in the hope of building a better world and a better future. I have been a beneficiary of that generosity, and I hope that the work that I have produced will in turn benefit the lives of the people who have supported me.

I would like to thank my committee members, Jeff Gore, Jeremy England, and Leonid Mirny, and my academic advisor, Ibrahim Cisse, for providing me with professional and scientific guidance over the past several years.

I am entirely incapable of expressing my gratitude for having had the opportunity to work with Linlin Chen. Linlin reached out to me in September of 2016 to ask if she could do a UROP with me, and I have benefited tremendously from the opportunity to mentor her. I can't wait to watch her trajectory as she starts graduate school at Caltech.

Dan Oran and I have been friends since I formed the D&D club at the Cambridge School of Weston in my first year of high school, at the age of 13. His transition from professional photographer, working out of an art gallery in Philadelphia, to pioneering nanotechnology researcher, has been nothing short of remarkable. I have been very grateful for the opportunity to count him both as a friend and as a colleague, and I hope to have more opportunities to do so in the future.

Working with Bob Stickels and Carly Martin on Slide-seq was one of the most fun collaborations that I had the good fortune to participate in in graduate school. Bob is always generous and charitable, and has an incredible sense of determination and a penchant for discovering random gems lying on the manifold of biological protocols. Carly is one of the most brilliant and talented RAs I have worked with, and it has been great to watch her discover her own formidable ability both in the wet lab and in analysis. I am thrilled that she got into MIT, and look forward to seeing what great things she will accomplish next.

I benefited greatly from the mentorship of many others in the Boyden lab, including Shahar Alon, Kate Adamala, and Kiryl Piatkevich, who have served as sounding boards for my ideas on many occasions. Demian Park has generously provided me with more wells of neuron culture than I can count. Oz Wassie was my rotation mentor and taught me how to pipette and how to do cell culture, and was patient and understanding when I subsequently and unceremoniously abandoned our project. Daniel Schmidt kindly and patiently taught me to run PAGE gels on an evening when neither of us wanted to still be in lab. Daniel Estandian and Noah Jakimo taught me how to clone, and Daniel and Grace Hyunh taught me how to perform surgeries. Dan Goodwin taught me more things than I can possibly enumerate, and I hope that I am able to provide him with even a

fraction of the wisdom and mentorship in science that he has provided me in navigating the many issues of life.

I have had the great good fortune to work with many other exceptional scientists in the course of my projects, including Anubhav Sinha, Evan Murray, Katriona Guthrie-Honea, Emma Costa, Changyang Linghu, Ruixuan Gao, Shoh Asano, Daniel Barabasi, Mark Skylar-Scott, Sophia Liu, Joe Scherrer, Aleksandrina Goeva, and Josh Welch. I have had the pleasure of interacting with Dawen Cai and Konrad Kording on many occasions. In addition, I would like to express my appreciation to many of the scientists at the Broad Institute, including Aviv Regev and Eric Lander, for their support. In addition, Lowell Wood and Christian Wentz have been constant sources of inspiration and fun.

I am grateful to Daniel Estandian, one of my former rotation students, for the work he did with me to try to map the connectome, and to Noah Jakimo, for working with me on Brainbar and helping me learn how to collaborate effectively. I am grateful to Nick Barry for the many ways he has helped me in science and in life, and for the work he did with me on the zinc finger and TAL effector project. I am especially grateful to Bobae An, a brilliant postdoc with whom it has been a pleasure to work, for all of her contributions towards the connectome. We will get it.

It was wonderful to have the opportunity to work with Ellen Zhong on the Tickertape project, as a rotation student. She made outstanding contributions to the project and taught me a great deal about machine learning, programming, and climbing.

With Jesse Engreitz and Charlie Fulco, I initiated a project in the Chen lab that lives on as HYPR-seq. I am very grateful to them for their help with that method.

Through my interactions with Louis Kang, I have been forced to reflect on what it means to have impact, and how best to go about realizing the progress I want to see in the world. I am very grateful to him for those interactions, and am excited to work with him in the future.

I would like to thank my many UROPs and rotation students, including Dana Gretton, Fadi Atieh, Preston Ge, Sophia Liu, Joe Scherrer, Daniel Estandian, Ellen Zhong, and Linlin Chen, who have worked and learned from me as I have worked and learned from them.

I am very lucky to count Andrew Payne as a great friend. He is a dreamer and a chaser of big problems, like me, and the many nights we spent together at Mead Hall served as a source of inspiration to me.

I have had several experiences throughout grad school that have helped me to remind me why I do science. Chief among those experiences, I would like to thank the MIT Assassin's Guild; as well as Polina Binder and all of the folks at Dead Hand Path (5:45 and J) for helping me to rediscover what I love about the universe.

I thank Madeleine Laitz from the bottom of my heart for the afternoons and evenings we spent dreaming and goofing together, and for always understanding exactly what I was going through during an especially challenging time. I thank David Turban and Fio Brady for sending me their love and support across the pond. To my many other wonderful friends, thank you for everything, I am lucky to have you.

I would like to thank my parents, Robbie Burnstine and Lou Rodriques, for providing me with a path to overcome my learning disabilities, pursue my dreams, and arrive at where I am today. I would like to thank my brother, Adam Rodriques, and my aunt and uncle, Michael Rodriques and Renee Aubrey for their loving support. I also owe an enormous debt of gratitude to Hannah Marshall, wherever she is, for her love, commitment, and support in the earlier years of my studies.

Chapter 1
Tool Development for Neuroscience

Biological systems are comprised of many components. In the other branches of science, systems consisting of multiple components are tractable when the components are identical and weakly interacting, lending themselves to statistical techniques. Practically all biological systems are statistical ensembles of non-identical, strongly-interacting components, and thus do not submit themselves to statistical analysis. Individual components can be studied in isolation, but knowledge of the function of a component *in vitro* rarely informs its function *in vivo*. The problem is exacerbated in the brain, where there are many more different kinds of cells than in other tissues (*1*, *2*), interactions are highly non-local (*3*, *4*), and the time-scales involved are much faster.

Given the failing of statistical techniques, many researchers believe that our ability to model and predict biological systems will be improved if we develop better tools to observe more components of the system simultaneously (*5*–*7*). Examples include increasingly multiplexed tools for measuring the distribution of RNA in space (*8*–*11*), for measuring the distributions of proteins in space (*12*–*15*), for measuring the projections of neurons (*16*–*19*), and for mapping neural activity (*20*–*23*). Through this work, progress has been made towards achieving "complete" descriptions of biological systems. Particularly in genomics, modern droplet-based approaches now enable the expression levels of every gene in the genome to be quantified in individual cells (*24*, *25*). In neuroscience, however, technologies for observing the activity of neurons are still 5 to 6 orders of magnitude away from a whole-brain recording system for the mouse brain; methods for measuring connections between neurons are likewise ~3-4 orders of magnitude away from being able to map the entire mouse brain; and measures for observing the spatial organization of gene expression in tissues are still limited to ~1% of the genome.

In this dissertation, I present designs or implementations of six new tools, each of which aims at increasing our ability to model, measure, or perturb biological systems (and especially neural systems) in a different way. Chapters 2 through 5 appear in print already, as described below. All of the work I describe here was performed with extensive assistance from many coauthors and collaborators. Detailed acknowledgements are included at the beginning of each chapter, in a foreword.

The first technology I describe concerns the measurement of the activity of neurons in the brain. Neuronal activity can be measured using electrical recording devices, such as electrodes (*26*).

However, electrode arrays suffer from a tradeoff between the number of channels on the array and the cross-sectional area. Recording electrodes can be made as narrow as 10 microns in diameter, but typically only have a single recording site, whereas electrode arrays with ~100 micron diameter can have hundreds of recording sites, but implantation into brain tissue leads to tissue damage up to and including hemorrhage (*27*, *28*). Indeed, as I later demonstrated in the course of my Slide-seq work, traumatic brain injury leads to sustained perturbations in neural activity, so activity recorded from traditional recording electrodes might not even be a reasonable reflection of ordinary activity (*29*). Moreover, because of the risk of tissue damage, recording electrodes cannot generally be applied in human brain tissue, and the number of electrodes that would need to be implanted into the brain in order to record from a large majority of the neurons in the brain is excessive (*30*). The field would benefit substantially from improved recording devices that fit more recording sites into a smaller frame. In Chapter 2, which appears in print as (*31*), I lay out a design for a new recording device for electrophysiology that breaks the tradeoff between the number of recording sites and the cross-sectional area of the electrode array by using an optical readout, rather than an electrical readout. The device directly converts electric fields from neurons into changes in the refractive index of a semiconductor waveguide, and then uses optical reflectometry to detect the refractive index as a function of position along the waveguide, packing 1000 recording sites into a $100\ \mu m^2$ device, greatly surpassing current designs (*32*–*35*).

In the process of considering the fabrication of electrical recording devices, it rapidly became clear to me that existing, 2D fabrication technologies may be insufficient for fabricating the kinds of complex nanotechnologies that will be necessary to study brain function. Moreover, althou In Chapter 3, which appears in print as (*36*), I present a method to reverse the process of Expansion Microscopy, a super-resolution technique that had been invented in the lab previously (*37*), in order to shrink structures for nanofabrication purposes. Uniquely among all nanofabrication technologies, Implosion Fabrication allows for direct 3D laser writing of metal structures with nanoscale feature sizes. Moreover, it uses a 3D molecular scaffold to position materials in space, allowing for the fabrication of structures with arbitrary 3D geometries. Implosion Fabrication is unlike any other existing nanofabrication technologies, and points the way to new and improved 3D nanofabrication tools.

Beyond the measurement of neural activity, a major challenge in neuroscience concerns mapping the many kinds of cells in the brain. In contrast to many other tissues, the brain consists of strictly organized structures consisting of thousands of cell types (*24*, *25*, *38*). Existing methods for mapping cell types in space rely on imaging, either using antibodies to characterize the protein content of cells or using in-situ hybridization to characterize the RNA content. However, all such techniques typically require special optimization for each sample or tissue type. Given the rapidly decreasing cost of RNA sequencing, a technique that can directly infer the spatial organization of tissue from RNA sequencing data would greatly reduce the barrier to accessing spatial gene

expression data. In Chapter 4, I present Slide-seq (*29*), a high-throughput tool for mapping the spatial distribution of cell types in tissue. Slide-seq enables direct capture of RNA from tissue onto a barcoded surface, allowing the positions of the RNAs to be reconstructed. From Slide-seq data, using new algorithms that we developed, one can infer the positions of different know cell types in space, and discover new patterns of spatial gene expression. Slide-seq leverages the throughput advantages of single-cell RNA sequencing technologies, and combines them with spatial resolution more typical to in-situ hybridization techniques, to provide a fundamentally new kind of data in the genomics toolkit.

Ultimately, however, a full description of the spatial organization of tissue will include a description of the spatial distribution of proteins, as well as RNA. More to the point, most of the functions in cells are performed by proteins, so understanding the protein composition of a cell is crucial to understand its behavior. RNA sequencing methods such as Slide-seq are often used to study the protein composition of cells by proxy, under the assumption that the RNA composition of a cell and its protein composition are highly correlated. However, this assumption is not true in general (*39*), necessitating equally powerful methods for visualizing the protein composition of cells. Although antibodies have been applied for this purpose (*40–42*), antibodies are also thought to be a major source of the reproducibility crisis in biology (*43*). In Chapter 5, I present a theoretical analysis of an approach for direct protein sequencing (*44*), which would provide an alternative, antibody-free method for highly multiplexed protein detection.

Returning to the question of neural activity recording, and inspired by the power of RNA sequencing methods, in Chapter 6 I present a method for recording transcriptional activity into the sequence of RNA with temporal resolution, allowing the history of RNA transcription in a cell to be inferred by RNA sequencing. The goal of the project, inspired by earlier work DNA tickertapes (*45*), was to allow for the activity of neurons to be encoded into the sequence of RNA, enabling neural activity to be measured in high throughput using RNA sequencers. This method was motivated in a similar way to the method presented in Chapter 2: although new methods allow measurement of activity from tens of thousands of neurons simultaneously, with theoretical access to hundreds of thousands (*46*), this still amounts only to 0.01%-0.1% of the neurons in the brain, and detection of activity in deep neural populations in freely behaving mice remains challenging (*47–49*). Using the RNA recorder presented in Chapter 6, one could in principle record from tens or hundreds of millions of neurons simultaneously for costs on the order of $10,000 using currently available sequencing technology. Chapter 6 does not yet appear in print, but we expect it to come out before the end of 2019.

Beyond mapping the activity and cell types of the brain, a major challenge in neuroscience is to map the *structure* of the brain. The brain consists of neurons connected by chemical and electrical synapses and molecular signaling pathways, and the set of all synaptic connections between neurons in the brain is typically referred to as the "connectome." The impact of the mammalian

connectome is difficult to estimate, but will likely be far-reaching: both the connectome and the lineage of C. Elegans have become widely used and cited resources for generating hypotheses about the interactions between neurons in those circuits (*50–52*). Moreover, the discovery of new cell populations defined by their connectivity regularly upends neuroscience research and prevents a holistic analysis of brain circuits (*53–55*), and the connectome would provide a systematic way to catalog all such populations. Stereotyped motifs of connectivity between different cell types may support generalized computations or transformations of information, with potential impact on AI. Finally, many brain disorders, such as schizophrenia, autism, and Alzheimer's disease, are known to involve changes in synaptic frequency and structure (*56–58*), and a tool for capturing the connectome would enable those phenomena to be studied systematically, potentially opening the door to new therapeutics.

At the same time, as we have come to better appreciate the diversity of cell types and synaptic mechanisms in the brain (*24*), it has become increasingly apparent that molecular annotation (i.e., a description of the molecules present in cells and synapses) will be a vital part of any connectomic effort. Indeed, this has also been the lesson from the C. Elegans connectome, where a lack of molecular annotation impeded our understanding of electrical synapses for more than three decades after the synaptic connectome was originally published (*59*). Ion channels govern how neurons generate their electrical pulses, and synaptic transmitters and receptors govern how neurons exchange information and transform presynaptic electrical pulses into postsynaptic ones; other proteins that generate or receive other messages are also known to be important for neural computation.

Up until now, the connections between neurons have been mapped using electron microscopy, but the scale and detail of the brain circuits that have been mapped using electron microscopy has not increased significantly since the first EM connectome was published in 1986 (*60*, *61*). Progress in this area has been fundamentally limited by the fact that electron microscopy cannot be utilized to visualize biomolecules such as proteins, DNA, or RNA. This limitation has two consequences: firstly, the proteins and nucleic acids present in a neuron can be used to distinguish that neuron from its neighbors, but because electron microscopy cannot visualize these molecules, neurons in an EM dataset must be reconstructed using machine vision algorithms (*62*), which have unacceptably high error rates and require hundreds of millions of hours of work by human annotators per cubic millimeter (*61*, *63–72*). Secondly, although electron microscopy can be used to visualize synapses, the function of the synapses, and of neurons, is defined by their molecular composition. By discarding molecular information, EM connectomics is fundamentally incapable of inferring the computational function of a neural circuit.

In Chapter 7, I lay out strategies for an optical approach to connectomics using molecular barcodes, which would enable the synaptic organization and the molecular composition of neural circuits to be mapped simultaneously. Using molecular barcodes provides a potential path to

circumvent the computational reconstruction challenge, while the optical readout allows the method to be combined with antibody staining techniques to reveal cell types and the distributions of proteins in the tissue. Because it is capable of using molecular information to distinguish neurons, our technology could map an entire mouse brain with molecular and connectivity information in only 3 years for $60M. It could ultimately be used to map large parts of the brains of primates, and perhaps even humans.

However, realizing the full impact of any of the technologies described here will require them to be scaled up. In the case of technologies like Slide-seq, they must be made broadly available to the academic community to realize their full impact, whereas in the case of connectomic barcoding technologies, they must be scaled up to the whole brain level. In both cases, the incentive structure in academia is insufficient to support the necessary scalable research efforts. In Chapter 8, which does not appear in print, I reflect on the phenomenon that the vast majority of tools developed in biology achieve extremely limited impact, and connect this phenomenon back to limitations on the system of incentives present in academia. I propose the creation of new focused research organizations (FROs) that would pursue research that requires more resources and focus that one can achieve in academia, but that is not yet ready for a for-profit venture.

The remaining chapters, 9 through 13, are appendices and supplementary information.

Chapter 2
Optical Reflectometry for Recording Neural Activity

The damage that implanted electrodes cause to neural tissue is one of the greatest challenges facing neural activity recording today. To overcome the challenge of implantation, Lowell Wood proposed to Adam Marblestone that recording electrodes could be delivered to the brain through the vasculature, navigated to locations of interest similarly to catheters. To record neural activity, however, it would be necessary to navigate the fiber into narrow blood vessels on the order of ~10 microns in diameter. No one had ever packed multiple recording sites onto a fiber of only 10 microns in diameter (*33*), so in order to put many recording sites onto our fiber, we would need to find alternative fabrication strategies, or an alternative method for detecting neural activity along the length of the fiber. Lowell suggested that optical reflectometry could be used to detect changes in the refractive index along the length of the fiber. If electrical activity could be transduced to detectable changes in the refractive index, then a fiber reflectometer would enable the electric field to be read out in 10 to 20 micron intervals along 10 centimeters of fiber, without any of the complications of wiring, parasitic capacitance, or thermal noise that one encounters in electrical systems. Adam worked on the project, but couldn't identify a method for transducing the electrical activity to a change in the refractive index with high enough sensitivity to be detected using an optical reflectometer.

In my first week after entering the Boyden Lab, in August 2014, Adam pitched me on the project. I identified the various electro-optic effects as candidates, including the Pockels, Kerr, and free-carrier dispersion effects. The free-carrier dispersion effect benefits from relying on the amount of charge present in the material, rather than the electric field; thus, it can be amplified for a given voltage using a very high capacitance. Adam and I did the primary analysis together, and the manuscript was submitted in early 2015, although it took us a year to complete revisions. The remainder of this chapter now appears as Ref. (*31*).

On its face, the device, if fabricated, would be extremely impactful for neuroscience. Compared to existing, highly multiplexed electrophysiology devices (*32*, *73*, *74*), the device proposed here would have an immensely simplified electrical backend (requiring only a single amplifier, rather than many amplifiers and a signal multiplexing scheme), and a ~10x reduced cross-section, at the cost of lower sensitivity and time resolution, a tradeoff that would prove useful in many applications. However, we lacked experience in the necessary fabrication methods, in materials, and in the reflectometry readout, and never pursued it beyond the design stage.

Summary

We introduce the design and theoretical analysis of a fiber-optic architecture for neural recording without contrast agents, which transduces neural electrical signals into a multiplexed optical readout. Our sensor design is inspired by electro-optic modulators, which modulate the refractive index of a waveguide by applying a voltage across an electro-optic core material. We estimate that this design would allow recording of the activities of individual neurons located at points along a 10-cm length of optical fiber with 40 μm axial resolution and sensitivity down to 100 μV using commercially available optical reflectometers as readout devices. Neural recording sites detect a potential difference against a reference and apply this potential to a capacitor. The waveguide serves as one of the plates of the capacitor, so charge accumulation across the capacitor results in an optical effect. A key concept of the design is that the sensitivity can be improved by increasing the capacitance. To maximize the capacitance, we utilize a microscopic layer of material with high relative permittivity. If suitable materials can be found—possessing high capacitance per unit area as well as favorable properties with respect to toxicity, optical attenuation, ohmic junctions, and surface capacitance—then such sensing fibers could, in principle, be scaled down to few-micron cross-sections for minimally invasive neural interfacing. We study these material requirements and propose potential material choices. Custom-designed multimaterial optical fibers, probed using a reflectometric readout, may, therefore, provide a powerful platform for neural sensing.

Introduction

The extracellular electrode is a classic neural recording technology. The electrode is essentially a conductive wire, insulated except at its tip, placed in the extracellular medium as close as possible to a neuron of interest, where it samples the local voltage relative to a common reference in the brain (*75*, *76*). This extracellular voltage differential is typically on the order of 100 μV in response to an action potential from a nearby neuron (*30*) and decays over a distance on the order of 100 μm. Note that the "transmembrane" voltage during an action potential is much larger, on the order of 100 mV.

The virtues of the electrode are twofold. First, the technique can reach single neuron precision by virtue of the electrode being inserted close to the measured neuron. Second, compared to optical methods, no exogenous contrast agents (i.e., genetically encoded fluorescent proteins, voltage sensitive nanoparticles, chemical dyes) are necessary: the endogenously generated electric currents in the brain are sensed directly in the form of a voltage. Ideally, for a neurotechnology to be medically valuable for a large number of human patients, it should not require modification of the neuron.

Yet, while multielectrode arrays allow the insertion of many electrodes into a brain, electrodes have limitations (*30*) in scaling to the simultaneous observation of large numbers of neurons. The

bandwidth of an electrical wire is limited by the cross-sectional area of the wire, due to the increase in RC time constant with increased resistance. Large numbers of high-speed electrical signals cannot be effectively multiplexed into a single electrical wire, hence, large numbers of wires must be routed out of the brain. Typically, in high-density multielectrode recording systems, one lithographically defined electrical trace is used per recording site. Creating such complex electrical wiring becomes increasingly difficult for long probe lengths, e.g., with lengths of centimeters.

In order to maintain the advantages of electrodes, single neuron precision based on endogenous neural signals while enabling improved scaling performance, we turn to photonics. Telecommunications has moved from electrical to optical data transmission because of the high bandwidths and low power losses enabled by optics in comparison to electrical conductors (*77*); the same may be helpful for neural readout technologies. Because optical radiation heats brain tissue and scatters off tissue inhomogeneities, a wired (i.e., fiber or waveguide based) optical solution may be desirable, i.e., using optical fibers to guide light so that it need not travel through the tissue itself. Second, to minimize volume displacement, signals from many neurons should be multiplexed into each optical fiber. Third, ideally, the sensing mechanism would rely only on endogenous signals, e.g., electrical or magnetic fields from the firing neurons, rather than imposing a need for exogenously introduced protein or nanoparticle contrast agents. With ~100,000 neurons per mm^3 in the cortex, or a median spacing of roughly one neuron per cube of size $(21.5\ \mu\mathrm{m})^3$, we require an axial resolution of sensing in the range of tens of micrometers. The system should be compatible with a variety of form factors, e.g., thin flexible fibers suitable for minimally invasive endovascular delivery (*78*, *79*), or rigid pillars suitable for direct penetration of the brain parenchyma (*80*).

Our proposed architecture is based on two powerful technologies developed by the photonics industry: fiber optic reflectometry, which enables optical fibers to act as distributed sensors (*81*–*84*), and electro-optic modulators based on the plasma dispersion effect, which generate large changes in the index of refraction of a waveguide in response to relatively small applied voltages (*85*–*89*). By combining reflectometry with electrooptic modulation, we propose that it would be possible to do spatially multiplexed neural recording in a single optical fiber.

Design Principles:

Reflectometers are capable of measuring changes in the index of refraction along the length of an optical fiber by sending optical pulses down the length of the fiber and recording the times and magnitudes of returning reflections (*82*). We propose to use reflectometry to sense neural activity at many points along the length of an optical fiber, as shown in Figure 2-1(a). The goal is to send a pulse of light into the fiber and to measure the reflections and their timing to determine the one-dimensional profile of neural activity along the length of the fiber. The local voltage at a given position along the fiber will modulate its local index of refraction via the free carrier dispersion

effect, giving rise to reflections. A reflectometer located outside the brain would then determine, at each time, the spatial profile of extracellular voltage along the length of the fiber.

Fiber-Optic Reflectometry

To determine the magnitude of the reflections generated by a change in local refractive index inside a fiber, note that when an electromagnetic planewave propagates in a material with refractive index n_1 and is normally incident on a material with refractive index n_2 the power reflected is given by the Fresnel equation:

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 \quad (1)$$

The waveguide under consideration will be divided into alternating segments of refractive indices n_1 and n_2, respectively [Figure 2-1(c)]. We define $\tilde{n} = (n_1 + n_2)/2$ and $\Delta n = n_2 - n_1$. At every interface between the two segments, a reflection is generated of magnitude:

$$R_0 = \left(\frac{\Delta n_0}{2\tilde{n}}\right)^2 \quad (2)$$

FDTD simulations (*90*) of the waveguide structure using the MEEP (*91*) software package [Figure 2-1(e)] confirm that the baseline reflections are of the predicted order of magnitude per this simple model. Assuming now that an event (i.e., local neural activity) causes n_2 to increase by a small amount $\Delta n \ll \Delta n_0$, the resulting reflections generated by the interface are given by

$$R = \left(\frac{\Delta n_0 + \Delta n}{2\tilde{n}}\right)^2 = R_0 + \frac{\Delta n \Delta n_0}{2\tilde{n}^2} + O(\Delta n^2) \quad (3)$$

The change in the reflections generated at the interface due to an event is thus

$$\Delta R = \frac{\Delta n \Delta n_0}{2\tilde{n}^2} \quad (4)$$

Reflection intensity sensing. Reflectometers are limited both in the minimum value of R that they can sense (termed the sensitivity) and in the minimum value of ΔR that they can sense (termed the resolution). In our device, the baseline power reflected at the boundaries between n_1 and n_2 will be much greater than the sensitivity of the reflectometer. Thus, the ability of the reflectometer to measure a change in the index of refraction is limited by its resolution, which is in turn fundamentally limited by photon shot noise. For the simple case of a time domain reflectometer, the number of photons registered at the detector due to a reflector of magnitude R_0 is given by

$$N_{\text{reflected}} = \text{QE} \cdot \frac{P}{h \cdot c/\lambda} \cdot R_0 \cdot \frac{1}{\text{BW}} \quad (5)$$

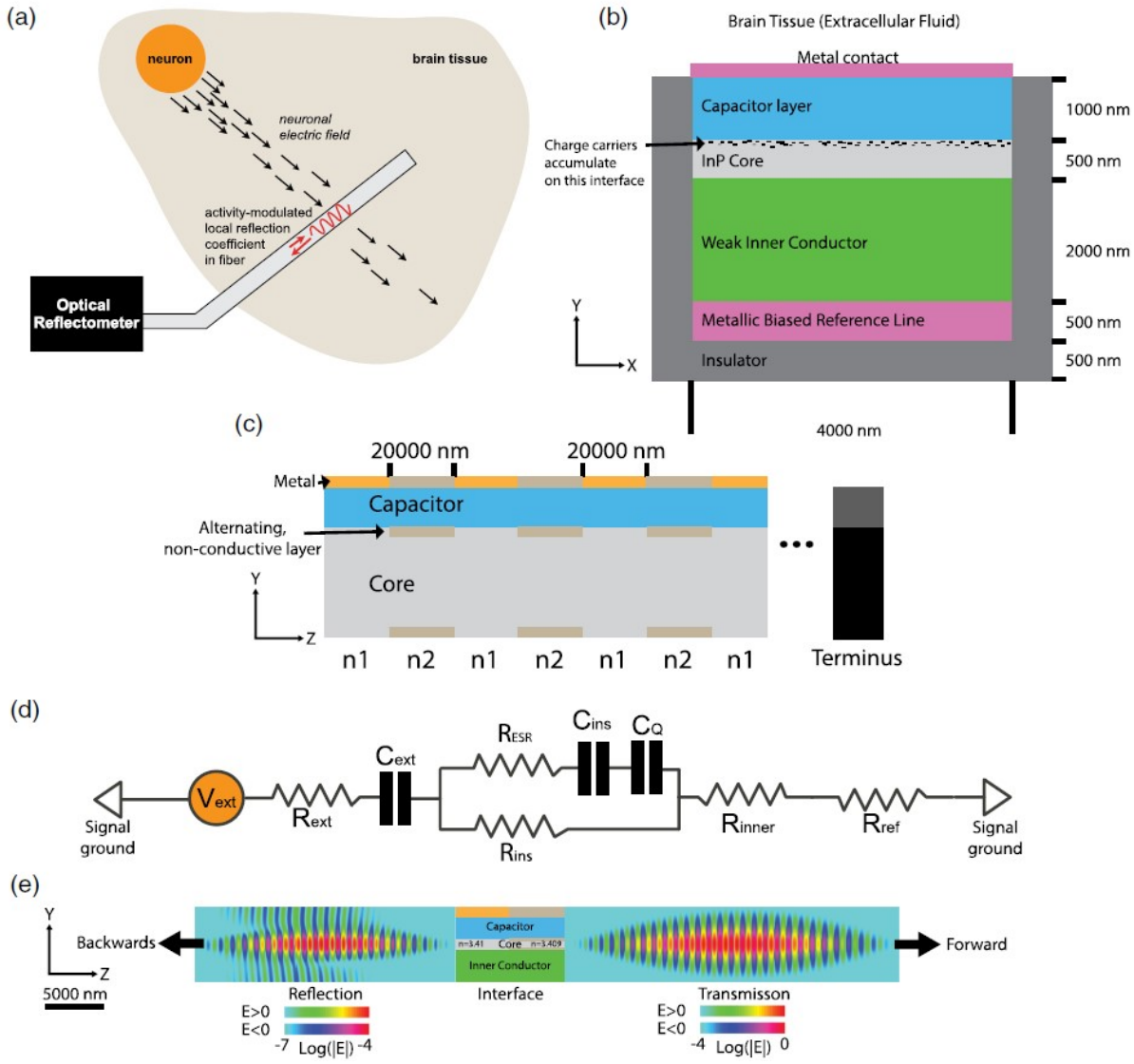

Figure 2-1 (a) High-level architecture. An optical fiber inserted into the brain acts as a distributed sensor for neuronal activity, which is read out by an optical reflectometer. **(b)** Axial cross-section of the probe. When a voltage is applied across the capacitor layer, free-charge carriers in the inner conductor and core build up on the surface of the capacitor layer and alter the refractive index in the core. A high capacitance is desired to improve sensitivity. **(c)** Longitudinal cross-section of the reflectometric probe. Alternating segments of higher and lower refractive index create baseline reflections at their interfaces, the intensities of which are modulated by the local extracellular voltage. The difference between n1 and n2 is generated by a thin layer of nonconductive material with a different index of refraction, which also serves to localize voltage-dependent refractive index changes to alternating segments. On the surface of the fiber, there are alternating sections of metal contact pads and oxide, to separate sensing and nonsensing regions. Caption continues on next page.

where P is the power entering the fiber, QE is the detector quantum efficiency, and BW is the sensing bandwidth. With a signal-to-noise ratio of $N/\sqrt{N}$ due to photon shot noise, the resolution

of the detector is given in decibels (dB) by

$$\mathrm{dB}_{\text{shot noise limit}} = 10\log_{10}\left(1+\frac{1}{\sqrt{N}}\right) = 10\log_{10}\left(1+\frac{1}{\sqrt{\mathrm{QE}\cdot\frac{P}{h\cdot c/\lambda}\cdot R_0\cdot\frac{1}{\mathrm{BW}}}}\right) \quad (6)$$

In all that follows, we will assume a bandwidth of 1 kHz, a quantum efficiency of 1, and a free-space wavelength $\lambda = 1550$ nm. Note that a higher bandwidth would be required to see the detailed shapes of individual action potentials, as may be required for spike sorting. For $P = 100$ mW, $\mathrm{QE} = 1$, $\mathrm{BW} = 1$ kHz, and $\Delta n_0 = 10^{-4}$, corresponding to $R_0 = 2.16\times10^{-1}$ (-96.6 dB), we then have a shot noise limited resolution of 0.01 dB, similar to existing reflectometers. More generally, for a signal ΔR to be sensed on top of a signal R_0, we must have

$$10\log_{10}\left(1+\frac{\Delta R}{R_0}\right) \geq 10\log_{10}\left(1+\sqrt{\frac{\mathrm{BW}}{\mathrm{QE}\cdot\frac{P}{h\cdot c/\lambda}\cdot R_0}}\right) \quad (7)$$

As we describe below, the device will be sensitive to changes in the index of refraction on the order of $\Delta n \sim 10^{-7}$ to 10^{-6}. Thus, because $\Delta n_0 \gg \Delta n$, the device operates in the linear regime of

(d) Equivalent circuit diagram of the device. The equivalent circuit of the device consists of a resistor representing each of the material layers between the neuron and the metal reference line, and three capacitors, one of which (C_{ext}) represents the interfacial capacitance, one of which (C_{ins}) represents the capacitance of the capacitor layer, and one of which (CQ) represents the capacitance due to the non-negligible charge centroid in the semiconducting core. The effective series resistance R_{ESR} of the insulating region capacitor can be neglected provided the capacitor has high quality factor Q at 1000 Hz, and the parallel resistance of the insulating capacitor layer R_{ins} can be neglected provided it is much larger than R_{ref}. R_{ref} is the resistance of the metallic reference line, R_{inner} is the resistance of the weak inner conductor layer and R_{ext} is the resistance of the brain–electrode interface. If, in addition, R_{ref} is chosen to be larger than the other resistances in the circuit, the capacitances C_{ins}, C_Q, and C_{ext} may be treated as series capacitances. (e) Optical simulation: We used the MIT Electromagnetic Equation Propagation (MEEP) package to simulate a waveguide with a silicon core divided into two regions. We used Si rather than InP as the simulated core material, because of the availability of well-validated tools for Si electrostatics simulation. In the first region, the core consisted of a 500 nm layer of silicon ($n = 3.410$). In the second region, the core consisted of a 460 nm-wide layer of silicon with two 20 nm layers of a material with $n = 3.40$ both above and below. The effective refractive index in the second region was thus 3.409, corresponding to $\Delta n = 10^{-3}$. The electric field profiles are shown on a logarithmic scale for the waves transmitted (right) and reflected (left) from the boundary between the regions, shortly after the reflection event. The left and right images have been normalized separately. The maximum value in the left image is $\sim 10^4$ times smaller than the maximum value in the right image, consistent with a value of R on the order of 10^{-8} for $\Delta n = 10^{-3}$.

Eq. (3), so Eq. (7) may be conveniently re-expressed as

$$\frac{\Delta n}{\tilde{n}} \geq \sqrt{\frac{\mathrm{BW}}{\mathrm{QE} \cdot \frac{P}{h \cdot c/\lambda}}} \qquad (8)$$

With the choices of the bandwidth, quantum efficiency, and wavelength given above, the resolution limit is strictly a function of power. The minimum resolvable Δn is shown as a function of P in Figure 2-2. Notably, Eq. (8) is independent of Δn_0 in the linear regime. The inset in Fig. 2 shows a schematic example of the expected output.

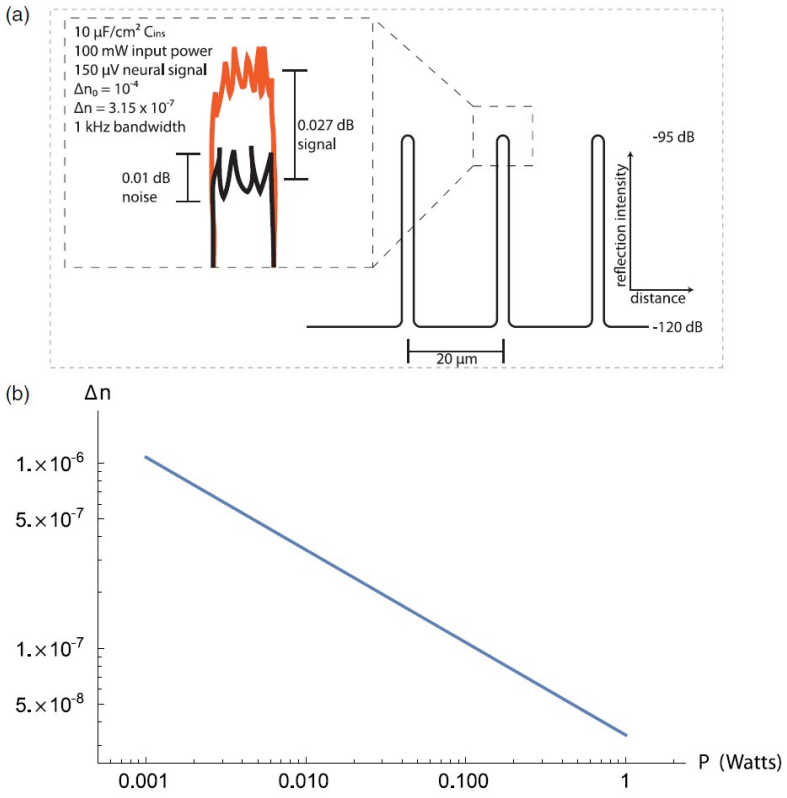

Figure 2-2: Fig. 2 Minimum resolvable value of Δn. **(a)** A schematic example of the expected output trace. Black color is the baseline reflection registered by the device; orange color is the reflection measured when the neuron fires. Spatial resolution is exaggerated for illustration. **(b)** The minimum change in the index of refraction of the optical fiber that can be sensed by an ideal, shot noise-limited reflectometer is shown as a function of the laser power P for a **1 kHz** bandwidth, index of refraction of **3.36**, quantum efficiency close to 1, and **1550 nm** wavelength.

So far, we have discussed the detection of changes in the refractive index via the modulation of reflectivity at each interface. An alternative strategy to detecting changes in the refractive index that accompany voltage signals is to measure the phase of the reflected light. This phase measurement can be performed with the identical Fourier-domain reflectometry scheme as for the amplitude-based measurement.

Spatial resolution. Other noise sources will also impact resolution in a realistic case, including laser power or phase noise and photodetector/amplifier/ADC noise. In particular, optical phase noise associated with the laser is limiting in current optical frequency-domain reflectometry (OFDR) systems (*92*); Littman–Metcalf external cavity tunable lasers, with narrow linewidths and low phase noise, can be swept at 1 kHz repetition rates over an optical

frequency range of several THz, leading to an OFDR "spatial" resolution of roughly 20 μm, which conveniently aligns with the average spacing between neurons in the cortex.

Repetition rate. Current commercial reflectometers achieve roughly 12 Hz repetition rates over 8.5 m. This corresponds to a measurement time of 1 ms for any given 10 cm segment of fiber, so using a similar device we anticipate that it would be possible to sense reflections along the length of a 10 cm fiber with a repetition rate of 1 kHz using frequency-domain reflectometers. In an OFDR system, the scan rate is limited by the frequency of laser wavelength scanning, the range of the scan determines the resolution, and the wavelength resolution of the scan and of the detector determines the scan range. Swept-source OCT constitutes demonstration of swept-source interferometry at a bandwidth of many kHz (*93*).

Electrooptic Modulation

Silicon electro-optic modulators are widely used in photonics to alter the propagation of light through a material in response to an applied voltage (*88*, *94*). Typical applications of electro-optic modulators take the form of electrically controlled optical switches: signals of roughly 5 V are used to drive optical phase shifts on the order of π. These devices are optimized for GHz bandwidths, with the goal of providing high speed, low power microchip interconnects (*87*), with bandwidths up to 30 GHz possible (*95*). Here, however, we are interested in the application of similar device physics to a very different problem: sensing extracellular neuronal voltages on the order of 100 μV at 1 kHz rates. Thus, our required switching rate is 1 millionfold slower, yet our required electrical sensitivity is on the order of 1 millionfold better. We are thus concerned with the design of electro-optic modulators optimized for sensitivity rather than bandwidth.

Free Carrier Dispersion Effect. The design shown in Fig. 1(b) consists of an extended multilayer semiconductor waveguide on a biased metal substrate, surrounded on three sides by insulation and on the fourth side by brain tissue or extracellular fluid. The "inner conductor" and "core" layers are weak, transparent conductors which function as resistive layers between the brain and the biased reference line. Throughout, we will assume that the core is made of n-doped InP, due to its large free-carrier dispersion effect (*96*), although other core materials are possible (see "Material Selection for the Capacitor Layer," below). Both above and below the core, there are ~5 nm thick layers [Figure 2-1(c)] in which the material alternates along the length of the fiber between the core material and a nonconductive material. The nonconductive material is chosen to have a refractive index that differs from that of the InP core by 0.01. At the boundaries between the alternating regions, there is an effective change in the index of refraction of $\Delta n_0 = 10^{-4}$, giving rise to a reflection to $R_0 = 2.16 \times 10^{-10}$ as per Eq. (2). This value of R_0 is chosen to avoid significant attenuation over the length of the fiber. Note that the sensitivity is independent of Δn_0 as long as we remain in the linear regime of Eq. (3). The alternating regions are 20 μm in length, with randomness introduced on the order of 1 μm to avoid the formation of strong peaks in the reflectivity with wavelength due to interference. The effective spatial resolution in this design is

then limited by the linear density of sensing sites, which are spaced at 40 μm from center to center, rather than by the underlying 20-μm spatial resolution of the reflectometer.

Above the core, there is an insulating layer that serves both as cladding, and as a capacitor over which most of the voltage will drop. The capacitor layer must be thick enough to serve as effective optical cladding, while also having a high capacitance. To satisfy these constraints, a material like barium titanate, strontium titanate, or calcium copper titanate may be preferred. We set this layer's thickness to ~1 μm. Clearly, the titanate layer must have lower refractive index than the core to act as a cladding. Although the optical properties of the titanate layer depend on its preparation, the band gap of a single crystal of barium titanate occurs at 3.2 eV (*97*), and the refractive index of barium titanate is ~2.4 for $\lambda = 600$ nm (*98*), so it is safe to assume $n < 2.4$ for $\lambda = 1.5\ \mu\text{m}$. Above the capacitor layer, there are alternating regions of metal and insulator, with the insulating regions coinciding with the alternating layers in the waveguide core. The metal regions provide the electrical interface to the brain and serve to define the sensing locations.

The InP core and inner conductor are doped and biased appropriately to allow most of the voltage to drop over the capacitor layer while maintaining low levels of optical attenuation, for example, $\sim 10^{17}\ \text{cm}^{-3}$. Other major materials requirements on the inner conductor are that it should ideally form an ohmic contact with both the InP core and the metal reference layer, and that its refractive index needs to be smaller than that of the n-doped InP, which is around 3.17 at $\lambda = 1.5$ μm. Potential materials candidates then include type III-V semiconductors with lower refractive indices, such as GaP, or II-VI semiconductors, such as ZnSe or CdS. These have lower refractive indices at 3.05, 2.45, and 2.30, respectively. These can be expitaxially grown on InP or vice versa due to the small lattice mismatch (*99*), and their conductivities can be tuned by doping. On the other hand, it would be important to prevent the formation of a rectifying junction at the semiconductor–semiconductor interface, the existence of which would depend on the band mismatch and doping levels. It might be possible to lower the junction barrier by, for example, minimizing the band gap difference between the two adjacent semiconductors. In the below analysis, we will assume that all junctions can be made ohmic. Note that the inner conductor is chosen to be thick enough to prevent optical attenuation due to the metal substrate (although there are other possible methods to reduce attenuation due to the metal, e.g., by removing the metal from the region directly under the waveguide, as in Ref. (*85*), and the metal substrate is chosen thick enough to provide a high-fidelity biased reference throughout the fiber.

The design relies on the free-carrier dispersion effect (also known as the plasma dispersion effect): the index of refraction within the InP core changes due to the accumulation of charge carriers in the InP when a voltage is applied across the capacitor layer (*85*, *96*, *100*). Many current integrated semiconductor electrooptic modulators are based on the free-carrier dispersion effect (*85*, *86*). In addition to the free-carrier effect, there exist other modalities of electro-optic modulation, such as the linear electro-optic (Pockels) effect (*101*), the quadratic electro-optic effect (Kerr) (*102*) and

the Stark effect (*103*). All of these effects would benefit from reducing the thickness d of the insulator layer to create a large electric field V/d (*104*). However, the free-carrier effect uniquely depends on the "charge," rather than the field, and can thus be amplified further by increasing the relative permittivity of the capacitor layer. In short, we need a large capacitor, which can be achieved by reducing the thickness and increasing the relative permittivity. For a material with a suitably large value of ϵ_r/d, the change in refractive index due to the free-carrier effect will be much larger than the changes that can be obtained via the other electro-optic effects. Although we focus on the free-carrier effect here, it should be noted that novel electro-optic materials, such as potassium tantalate niobate (*105*), with extremely high electrooptic coefficients compared to standard electrooptic materials like lithium niobate, could also potentially make possible designs based on the Pockels or Kerr effects.

An appropriate bias voltage will be applied through the reference conductor to ensure that the InP core layer operates in accumulation. This is necessary in order to avoid depletion (*106*), which would reduce the charge recruited to the surface of the capacitor for a given change in extracellular voltage, and thus reduce the sensitivity. Thus, we use the reference potential in the brain plus some fixed bias to achieve accumulation in the InP core along the waveguide. If needed, this bias could be achieved locally, but as long as the brain has no large voltage differences (e.g., >1 V), one global bias may be sufficient to allow the entire InP core to operate in accumulation.

Changes in the index of refraction in the free-carrier modulated region of the InP may be modeled as changes in the overall effective index of refraction of the fiber (*107*). The magnitude of this effective change is given by weighting the magnitude of the change in the free-carrier modulated layer by the percentage of power contained in that layer, i.e.,

$$\Delta n_{\text{eff}} = (1-\eta)\Delta n_{\text{active}} \tag{9}$$

where n_{active} is the index of refraction in the free-carrier modulated layer and $1-\eta$ is the fraction of the power in the beam contained in the active region. We will denote by d the thickness of the capacitor layer, by b the thickness of the layer of injected charge carriers in the InP, and by a the remaining thickness of the InP layer. An order-of magnitude approximation for η is then given by

$$\eta \cong \frac{a}{a+b} \tag{10}$$

and we have

$$\Delta n_{\text{eff}} = \left[1-\frac{a}{a+b}\right]\Delta n_{\text{active}} \tag{11}$$

For this reason, the InP waveguide is chosen to be thin to maximize the percentage of the optical wave contained in the layer containing the injected charges. Because of the deep subwavelength

thickness of the active layer, a precise calculation of Δn_{eff} could be done using a full-vectorial Maxwell simulation of the waveguide modes (*85*), but for our purposes, the approximation of Eq. (11) suffices to illustrate the basic scaling. Upon applying a voltage across the capacitor layer, the density of charge carriers injected into the active layer inside the InP core, denoted by ΔQ, is simply given by the equation for a parallel plate capacitor:

$$\Delta Q = \frac{C_{\text{ins}}}{eAb}\Delta V_{\text{ins}} \quad (12)$$

where C_{ins}/A is the capacitance per unit area of the insulator, e is the electron charge, b is the thickness of the layer of injected charge carriers in the InP, and ΔV_{ins} is the voltage dropped over the insulating region. Equation (12) may be recast in terms of the total voltage ΔV applied over the device by introducing an effective capacitance C_{eff}, such that

$$\Delta Q = \frac{C_{\text{ins}}}{eAb}\Delta V \quad (13)$$

In practice, C_{eff} will only deviate significantly from C_{ins} when the capacitance of the brain-fiber interface is significant (discussed below). The change in refractive index in the region with the injected charge is related to the change in the carrier concentration by a power law (*96*). When the injected carriers are electrons, the magnitude of the electro-optic effect in InP is greatest. The relation for the change in refractive index in the injected charge region is then

$$\Delta n_{\text{active,h}} = C_e \frac{C_{\text{ins}}}{eAb}\Delta V \quad (14)$$

where C_e is an empirically defined constant. For InP, the value of C_e is given for 1.55 μm light by (*108*)

$$C_e = -5.6 \times 10^{-21}\text{cm}^3 \quad (15)$$

This wavelength is chosen because the waveguide is made of InP, and InP is transparent at these telecom wavelengths. Telecom windows are around 1.3 and 1.5 μm due to local minima of the absorption of water, a hard-to-avoid contaminant in silica fibers. The exact choice of wavelength is not critical to the sensing mechanism itself; according to the Drude model of the free-carrier dispersion effect (*108*), the coefficient in Eq. (12) is quadratic in the wavelength.

Similar values are obtained for other semiconductors and other wavelengths (*96*, *108*). To find the effective refractive index within the InP waveguide, we multiply Eq. (14) by the volume factor $1 - \eta$ from Eq. (10). Assuming $b \ll a$ (i.e., that the injected charge layer is deeply subwavelength while the waveguide core thickness is on the same order as the wavelength), we find

$$\Delta n_{\text{eff}} \cong C_e \frac{1}{a+b}\left[\frac{C_{\text{eff}}}{eA}\Delta V\right] \quad (16)$$

Note that for a given waveguide thickness (i.e., $a + b$ constant), the result is independent of the thickness of the charged layer b. We will henceforth take $a + b \sim 500\,\text{nm}$. For a value of C_{eff}/A on the order of $10\,\frac{\mu\text{F}}{\text{cm}^2}$, justified below, we find $\Delta n_{\text{eff}} \sim 2 \times 10^{-7}$ for $\Delta V \sim 100\mu\text{V}$.

Effects of Other Capacitances. The brain–electrode interface also has a capacitance C_{ext} of $\sim 100\,\frac{\mu\text{F}}{\text{cm}^2}$ (*109*, *110*), which arises due to the presence of an electrical double-layer [see Figure 2-1(d)]. In addition, there is a capacitance C_Q due to the finite length scale of the charge distribution inside the semiconductor. This latter capacitance is given by $C_Q = \epsilon_{\text{core}}\epsilon_0 A/d_Q$, where ϵ_{core} is the relative permittivity of the core material, ϵ_0 is the permittivity of free space, A is the area of the sensing region, and d_Q is the charge centroid. The core can be one of many semiconductor materials (e.g., Si, InP), leading to similar fundamental electrostatics. We performed semiconductor simulations using the Sentaurus TCAD device simulator (version K-2015.06, June 2015) to evaluate C_Q. We used Si as the simulated core material, because of the availability of accurate and readily available tools for Si electrostatics simulation. To calculate the charge centroid d_Q, we simulated the electrostatics of an interface between silicon n-doped to a level 10^{17}cm^{-3} and a layer of oxide with relative permittivity of $\epsilon_r = 5$ and thickness of $d = 1\,\text{nm}$. When the silicon was in accumulation, the charge centroid was found to be $2.4\,\text{nm}$. The charge centroid is expected to be similar for our setup provided the value of $\frac{\epsilon_r}{d}$ is similar for the capacitor layer, which it would be for a layer of barium titanate with $\epsilon_r = 5000$ and $d = 1000\,\text{nm}$ (see below). It is more difficult to do these simulations for less common materials like InP, but we anticipate that the charge centroid for InP will be similar. Thus, in all following calculations, we will assume a capacitance of $4.5\,\frac{\mu\text{F}}{\text{cm}^{-2}}$ for the interface between the core and capacitor layer.

The electrostatics simulations also showed that, although the total charge ΔQ recruited to the capacitor layer surface upon application of a voltage is greater for higher doping levels, the relative change in charge ($\Delta Q/Q$) is greater for lower doping levels. However, for $\Delta Q \ll Q$, the sensitivity condition in Eq. (8) depends to a good approximation only on ΔQ, not on $\Delta Q/Q$, so the sensitivity of the device is increased for higher doping levels.

Figure 2-1(d) shows an equivalent circuit diagram of the device, which includes the interfacial capacitance and the capacitance associated with the charge distribution. At $< 1000\,\text{Hz}$, and subject to appropriate materials choices (see Sec. 2.2.3), the impedance of the circuit is dominated by these three capacitors rather than by purely resistive elements of the circuit. For this reason, we may ignore the purely resistive elements and treat the capacitances as though they were in series. To a good approximation, therefore, the charge that accumulates on the surface of the insulating region in response to a voltage ΔV across the entire device is given by

$$Q = C_{\text{eff}}\Delta V \tag{17}$$

where the effective capacitance C_{eff} of the surface and insulating region capacitors in series is

$$C_{\text{eff}} = \frac{1}{\frac{1}{C_{\text{ins}}} + \frac{1}{C_{\text{ext}}} + \frac{1}{C_Q}} \quad (18)$$

The capacitance of the insulating region is given by

$$C_{\text{ins}} = \epsilon_0 A \frac{\epsilon_r}{d} \quad (19)$$

where d is the thickness of the insulating region, ϵ_r is the relative permittivity, A is the area of the sensing region, and ϵ_0 is the permittivity of free space. Along with the laser power discussed above, the capacitance per unit area of the capacitor layer, C_{ins}/A, will be the primary figure of merit for determining the sensitivity and noise characteristics of the device. The effective capacitance C_{eff} is shown in Figure 2-3(a) as a function of the capacitance C_{ins} of the capacitor layer, assuming a surface capacitance per unit area (*109*, *110*) of $100 \frac{\mu\text{F}}{\text{cm}^2}$ and a sensing length of $20\ \mu\text{m}$. Note that the effective capacitance ceases to increase for values of $C_{\text{ins}}/A \gg 4.5\ \frac{\mu\text{F}}{\text{cm}^2}$, because for these values, the capacitance is dominated by the core-capacitor interface.

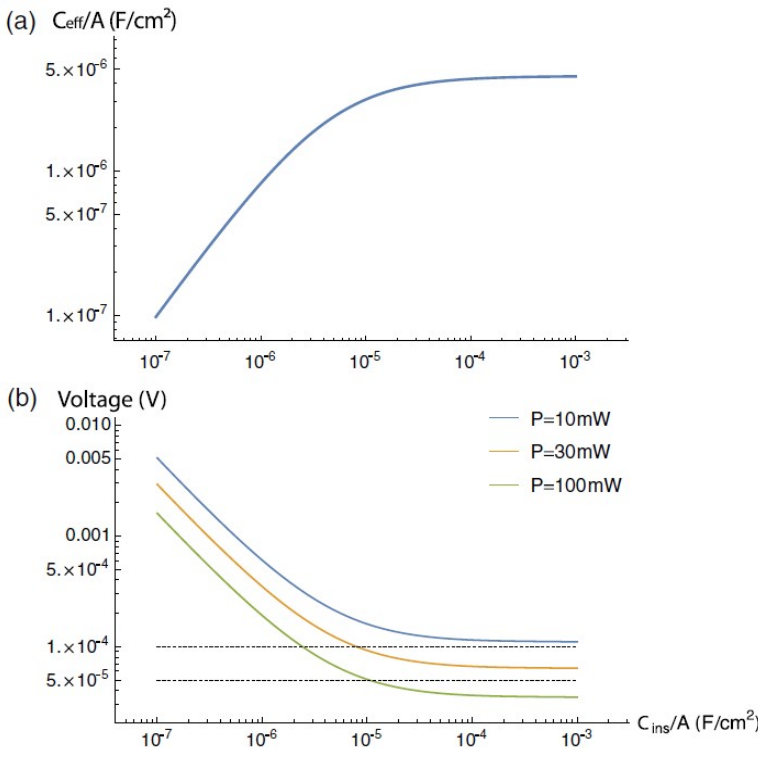

Figure 2-3: Properties of the design parametrized by C_{ins}/A. **(a)** The effective capacitance C_{eff}/A given in Eq. (18) is shown as a function of C_{ins}/A, assuming a capacitance of $100\ \mu\text{F/cm}^2$ at the surface of the device and a capacitance of $4.5\ \mu\text{F/cm}^2$ at the interface between the core and capacitor layers. **(b)** The minimum detectable change in voltage [obtained from Eqs. (8) and (16)] is shown as a function of C_{ins}/A for systems with (from top to bottom) a $100\ \text{mW}$ laser (blue), a $30\ \text{mW}$ laser (orange), and a $100\ \text{mW}$ laser (green). The black dashed lines correspond to 50 and $100\ \mu\text{V}$. To sense signals at the 50 μV level with a 100-mW laser, a capacitance on the order of $10\ \mu\text{F/cm}^2$ is necessary.

Noise Sources. A primary electrical constraint on the device is that the impedance at $1\ \text{kHz}$ must be dominated by the capacitor layer. If the effective capacitance per unit area of the capacitor is $C_{\text{eff}} = 5\,\frac{\mu\text{F}}{\text{cm}^2}$, corresponding to a value $C_{\text{ins}} = 10\,\frac{\mu\text{F}}{\text{cm}^2}$, then the capacitance of a region with width $4\ \mu\text{m}$ and length $20\ \mu\text{m}$ is $4\ \text{pF}$, corresponding to an impedance of $40\ \text{M}\Omega$ at $1\ \text{kHz}$.

Assuming that the metal layer has a resistivity no greater than $100\ \text{n}\Omega\text{m}$ (10× that of silver), if the metal layer is made at least $500\ \text{nm}$ thick, it will have a resistance of $10{,}000\ \Omega$ along the entire length of the fiber. The resistance of the inner conductor and core will be negligible compared to the huge capacitive impedance, provided they are chosen to be semiconductors. Finally, we must consider the voltage noise on the recording site itself, i.e., the metal contact pad interfacing directly with the brain. The recording site is often modeled as a constant phase element (*111*) and noise contributions come from the real part of its impedance (*75*), and are frequency dependent. We choose to write it in terms of parameters G and m, with an impedance of $Z_e = 1/G(j\omega)^{-m}$. The parameter G reflects the conductivity of the material, and the parameter m is often related to surface roughness and transport to the metal–electrolyte interface (*112*), with typical parameters ranging from 0.5 to 0.9. We will assume $m = 0.5$ for representing a rough surface, and a $1\ \text{kHz}$ impedance magnitude of $0.1\ \text{M}\Omega$. Thus, the resistance of the device is dominated by the recording site, as opposed to the ground lead or other elements, and the above parameters amount to a total RMS noise over a

$1\,\mathrm{kHz}$ band, found by integrating $4k_BT\mathrm{Re}(\mathrm{Z_e})df$ from $f = 0\,\mathrm{Hz}$ to $f = 1000\,\mathrm{Hz}$, of $\sim 1\,\mu\mathrm{V}$. This model agrees with what is found experimentally for similar sized electrode pads (*112*).

Efforts to reduce the recording site impedance are only needed for adjusting the noise influence of the recording site itself. Even an unplated gold surface will be sufficient here, because instead of $0.1\,M\Omega$ for an electroplated surface, we will have $Re(Z) = 1M\Omega$, with a resulting noise of $\sim 4\mu\mathrm{V}$ RMS instead of the $1\,\mu\mathrm{V}$ RMS calculated above. Only if C_{eff} were increased dramatically (e.g., to the equivalent impedance of $\sim 1\,\mathrm{M\Omega}$ at $1\,\mathrm{kHz}$), would efforts be needed to reduce the recording site impedance to prevent attenuation of the signal via the voltage divider. In any case, the voltage drops primarily over the capacitor layer and is not attenuated by resistors prior to the capacitor, and these electronic noise voltages are lower than the sensitivity of the device, which is limited by optical shot noise, and so can be neglected. Note that the impedances given here are also large enough for the input impedance of an implanted recording device (*113*).

Other forms of exogenous noise include mechanical bending of the fiber and thermo-optic effects, which may be particularly significant given the small width of the waveguide. However, these effects are expected to occur at a much lower frequency than the $\sim 1\,\mathrm{kHz}$ frequency content of spikes, and thus can be filtered out. Likewise, static or slowly changing bends (e.g., due to the heart beat) in the fiber can be subtracted off.

Dynamic Range. Local field potentials in the brain may vary by up to hundreds of millivolts, generating fields on the order of $1\ \mathrm{kV/cm}$ across a $1\,\mu\mathrm{m}$ capacitor layer. By contrast, the dielectric breakdown strength of barium titanate is roughly $10\ \mathrm{kV/cm}$, so dielectric breakdown is unlikely to be an issue (*114*). On the other hand, the dynamic range of the device may be limited by the density of states in the core, and thus it will be necessary to adjust the bias of the device (using the conductive reference layer) in order to ensure that the device can function in accumulation. If the device is allowed to function in depletion, C_Q will be much smaller than C_{ins}, thus reducing the sensitivity. Similarly, operation in inversion will suffer from deep depletion effects.

Tissue Heating. When we send light down the fiber, some light power may dissipate into the tissue. Depending on the level of round-trip light attenuation in the waveguide, each probe will dissipate a fraction f of the applied light power P. We next evaluate the acceptable level of such dissipation and how this constrains the device properties.

The human brain endogenously dissipates $25\,\mathrm{W}$ or $19\ \mathrm{mW/mL}$. The blood perfusion rate of human brain gray matter and white matter is roughly $r_{\mathrm{perfusion}} = 35000\,\mathrm{Wm^{-3}{}^{\circ}C^{-1}}$ (*115*). To avoid $> 2^{\circ}\mathrm{C}$ brain temperature rise, per the requirements laid out in Ref. (*30*) and elsewhere, we then require that each probe is surrounded by a perfusion volume of $V_{\mathrm{perfusion}} \approx \frac{fP}{r_{\mathrm{perfusion}} \cdot 2^{\circ}\mathrm{C}}$.

For a sense of scale, assuming that a $100\,\mathrm{mW}$ laser is used for the reflectometer, if $f \approx 50\%$ of this light power is dissipated into the tissue on a round-trip reflection, we then require a $250\,\mu\mathrm{L}$ perfusion volume, or a cylinder of radius $1.5\,\mathrm{mm}$ around each probe, assuming a $10\,\mathrm{cm}$ probe length. An attenuation of 50% over $20\,\mathrm{cm}$ corresponds to $\sim 3\,\mathrm{dB}$ over the length of the fiber, or $\sim 0.15\,\mathrm{dB/cm}$, on the order of the intrinsic optical attenuation of silicon (*116*) or indium phosphide (*117*). An additional potential source of tissue heating arises from transverse scattering of light at the interfaces between the successive waveguide segments of different refractive index. Using MEEP simulations to quantify the amount of light scattered out of the waveguide core, for adjacent segments with refractive indices of 3.409 and 3.41, we estimate that there will be a $3 \times 10^{-5}\%$ loss per boundary. With 500 boundaries per centimeter, this means a 0.015% loss per centimeter or 0.3% loss over a round trip in a $10\,\mathrm{cm}$ fiber. However, if Δn_0 is $\sim 10^{-4}$ instead, as discussed above, the amount of scattering generated this way is expected to be substantially reduced.

Attenuation due to bending is expected to be insignificant, with silicon-on-insulator waveguides reported to experience attenuation of only $\sim 0.1\,\mathrm{dB}$ per $90°$ turn at a radius of curvature of $1\,\mu\mathrm{m}$. Finally, to avoid transmitting any light into the brain tissue itself, a strong reflector can be placed at the end of the probe. Because the reflectometer has high spatial resolution, a large reflection from the end of the probe is not expected to interfere with the measurements.

Material Selection for the Capacitor Layer

The key figure of merit determining the properties of the device is the capacitance per unit area of the capacitor layer, C_{ins}/A. Along with the laser power, the figure of merit determines the sensitivity via Eq. (16). In Figure 2-3(b), the sensitivity of the device is shown as a function of C_{ins}/A. The vertical axis shows the minimum voltage signal that can be resolved using a shot noise-limited reflectometer, as calculated using Eqs. (8) and (16). The power law region (a straight line on the log–log plot) corresponds to the region in which $C_{\mathrm{eff}} \approx C_{\mathrm{ins}}$, so that the reflection coefficient $R \propto C_{\mathrm{ins}}\Delta V/\mathrm{A}$. For values of C_{ins}/A much greater than $4.5\,\frac{\mu\mathrm{F}}{\mathrm{cm}^2}$, we have $C_{\mathrm{eff}} \approx C_Q$, so the sensitivity does not improve with increasing C_{ins}/A.

Materials such as barium titanate, strontium titanate, and calcium copper titanate would likely be able to achieve a sufficiently large value of C_{ins}/A while also separating the core from the metal sensing pads. The chosen material must be able to maintain its high relative permittivity while film thickness is scaled down sufficiently to enable a high capacitance. Since dielectric properties often arise from grain boundaries within the material, the achievable grain size sets an approximate lower bound on the film thickness that can be utilized. Barium titanate films have been demonstrated with relative permittivities of roughly 5000 with grain sizes around $1\,\mu\mathrm{m}$ (*118*), r with relative permittivities of 2500 with grain sizes of $100\,\mathrm{nm}$ (*119*). Likewise, calcium copper titanate ceramics have been fabricated with relative permittivities between 1000 and 10,000 and

grain sizes from hundreds of nanometers to micrometers (*120*). Finally, relative permittivities on the order of 105 seem to be possible with larger grain sizes (*121*, *122*).

We are not aware, however, of direct measurements of the dielectric properties of high-dielectric ceramics in films of $< 1\,\mu\mathrm{m}$ thickness grown in InP substrates, thus verification of these properties should be a key question for early experimental studies of voltage probes like the one proposed here. A further potential concern with using dielectrics, such as barium titanate, is the presence of hysteresis in such materials (*123*). Since the neuronal signals involve potential changes on the order of $100\,\mu\mathrm{V}$, the hysteresis is expected to be small, but a detailed experimental characterization would be required.

We will assume that it is possible to fabricate a dielectric film with thickness $d \sim 1\,\mu\mathrm{m}$, d ~ 1μm, and with $\frac{\epsilon_r}{d} \sim 10^{10}$, for example, a $1\,\mu\mathrm{m}$-thick film of calcium copper titanate with $\epsilon_r \sim 10^4$, corresponding to a value of C_{ins}/A of $10\,\frac{\mu\mathrm{F}}{\mathrm{cm}^2}$. With such a capacitor, the device with a $30\,\mathrm{mW}$ laser would be capable of measuring signals at the $100\,\mu\mathrm{V}$ level and the device with a $100\,\mathrm{mW}$ laser would be capable of measuring signals at the $50\,\mu\mathrm{V}$ level.

Discussion

Ultra-large-scale neural recording is highly constrained both by physics and by the biology of the brain (*30*). Here, we have argued that an architecture for scalable neural recording could combine

1) the use of optical rather than electronic signal transmission to maximize bandwidth,
2) confined rather than free-space optics to reduce the effects of light scattering and absorption in the
3) spatial or wavelength multiplexing within each optical fiber in order to minimize total tissue volume displacement,
4) a thin form factor to enable potential deployment of fibers via the cerebral vasculature, and
5) direct electrical sensing to remove the need for exogenous dyes or for genetically encoded contrast agents.

Traditional electrode-based recording systems require a separate electrical connection for every recording site. They are limited in the depth they can access, because the magnitude of the thermal noise increases with the length of the probe. Furthermore, each connection must be accessed separately by the acquisition system (*124*). By contrast, the architecture proposed here offers several benefits, including the ability to read out neural activity over many centimeters with high sensitivity, the ability to multiplex tens of thousands of recordings into a single fiber with a simplified acquisition system, and the ability to scale the physical dimensions of the fiber without sacrificing performance.

In our proposed design, the $100\,\mu V$ scale extracellular voltage resulting from a neuronal spike is applied across a thin, high-dielectric capacitor. Charging of the capacitor results in modulating the accumulation layer in the neighboring InP waveguide core, altering the local refractive index of the InP and causing a detectable optical reflection. Reflectometry then enables multiplexed readout of these spike-induced reflections. Notably, the entire design fits into a package with a cross section that is in principle $< 5\,\mu m$ on a side (although additional material could of course be added for mechanical support if desired).

Every neuron in a mammalian brain is within a few tens of microns of the nearest capillary (*125*), well within the distance necessary for direct electrical sensing of the action potential (*30*), thus, in principle, the fine microvessels of the cerebral vasculature could serve as a delivery route for neural activity sensors, if the fibers could be made sufficiently thin (*79*), i.e, well below $10\,\mu m$ for the smallest capillaries. Thus, multiplexing thousands of neural signals into a single optical "wire" of $< 10\,\mu m$ thickness could potentially be enabling for novel endovascular approaches to neural interfacing.

It is worthwhile to contrast the proposed system to both microelectrode-based recording and optical imaging solutions. In our design, signals are captured electrically, similar to the recording mechanism of a microelectrode, and then are transduced to an optical communication channel for extracting the data from the brain. By contrast, in imaging approaches, the neuronal signal is transduced into the photochemical state of an indicator dye or protein inside the neuron itself, and then the signal is extracted by irradiating the brain and then capturing emitted fluorescent photons on a camera. Consequently, imaging approaches flood the brain tissue itself with light power and transduce signals via chemicals delivered to the neurons themselves. Our proposed method, in contrast, does not require flooding the brain tissue itself with light: the electrical pickup of the signal does not require power nor exogenous chemical probes, and the data collection is photon-efficient since, to the greatest extent possible, our design confines all light to the inside of the waveguide itself.

A key challenge in implementing such a design is to achieve a figure of merit C_{ins}/A for the capacitor sufficiently large to allow sensitivity to the neural signals of interest. We think that it would be possible using barium titanate or calcium copper titanate to achieve a figure of merit on the order of $10\,\frac{\mu F}{cm^2}$, which would allow the device presented here to sense signals of approximately $50\,\mu V$ with a $100\,mW$ laser.

In addition, supercapacitors with submicrometer thickness can be fabricated that achieve specific capacitances on the order of $1\ mF/cm^2$ (*126*), which would allow for the detection of $30\,\mu V$ signals with a $100\,\mu W$ laser, if they could be made compatible with our device. The sensitivity would also be improved substantially if a core material could be found with a smaller charge

centroid. Early experimental studies building on our theoretical estimates should seek to verify that a sufficiently high capacitance can be achieved in the desired form factor.

Several alternative strategies exist for improving the sensitivity of the device. The device senses the voltage in each sensing region twice, at the front and back ends of each sensing region, which could be factored into the analysis to improve SNR. If tissue-heating concerns can be overcome, the sensitivity of the device can be improved by increasing the strength of the laser. The sensitivity can also be increased by using a different core material with a stronger free-carrier dispersion effect. For example, at $\lambda = 1.3\,\mu\text{m}$, there is a maximum in the free-carrier dispersion effect of InP at a doping concentration around $3 \times 10^{17}\,\text{cm}^{-3}$ (*108*). By using a core material with a higher bandgap, such as GaP, it would be possible to perform reflectometry using visible light, for example, around $600\,\text{nm}$, which would increase the sensitivity of the device by increasing the overlap of the optical electric field with the charge-containing region of the core. Alternatively, silicon is also possible as a core material, for simplicity of fabrication. However, it would be necessary for the chosen core material also to have acceptable levels of field-induced birefringence and nonlinear response, effects which could cause frequency conversion or interfere with the reflectometry process. These processes should be evaluated empirically for a given power level, materials choice, and waveguide configuration. Finally, the sensitivity of the proposed device is dependent on the signal to noise ratio of the reflectometer. Although we have applied a conservative estimate of the shot-noise-limited resolution, other sources of noise will have to be minimized to achieve sufficient sensitivity for neural recording.

Finally, a major challenge will be the achievement of an attenuation level low enough to avoid excessive heating of the tissue. The heat dissipation can be reduced by reducing the laser power or using a core material with lower optical attenuation. For this reason, GaP is also an appealing option for the waveguide core, as it has been reported to have intrinsic optical attenuation much less than $0.1\ \text{dB/cm}$ at $600\,\text{nm}$ (*127*, *128*). Additionally, heat dissipation into the tissue could be reduced by the addition of an active heat transport system (such as a microchannel heat sink) to the device architecture (*129*, *130*).

The cost of the device will depend on the final choice of materials, the fabrication processes required, and the extent to which existing semiconductor fabrication pipelines are capable of meeting the requirements. Broadly, these devices can be fabricated with methods widely used in the nanofabrication field, but not all of these methods are industrialized at the scale of modern microchip manufacturing. Ultimately, such a device could be packaged together with optical or electrical stimulation channels for bidirectional neural interfacing.

If appropriate materials combinations can be fabricated, we have shown that the device could achieve the requisite sensitivity, noise level, and response time for recording both neural spikes

and local field potentials. More broadly, our results suggest that integrated photonics could enable highly multiplexed readout of neuronal electrical signals via purely optical channels.

Chapter 3
Implosion Fabrication

In late 2014, Expansion Microscopy became a major focus of the Boyden lab. Fei Chen and Paul Tillberg presented expansion microscopy to the lab at a lab meeting in November 2014, and I asked at the end whether the process was invertible (the answer was yes). Adam Marblestone realized that the idea had promise for the purpose of positioning DNA origami in 3D, and encouraged me to work on it. We met with Mark Skylar-Scott, who had done related work previously (*131–133*), and he helped us to get set up with the fluorescein patterning chemistry.

After several months of experiments, Dan Oran joined the project in early 2015. Dan has an undergraduate degree in photography, and brought with him extensive knowledge of old photographic chemistries. Dan succeeded in getting the basic patterning setup working, and I worked out the system for shrinking and dehydrating the gel in acid. Working with Ruixuan Gao, Dan, Rui and I validated the resolution of the patterning method using gold nanoparticles conjugated to the gel on the patterns.

A key breakthrough came in the spring of 2016, when Dan had the idea to leverage an old photographic chemistry to deposit silver onto the surface of gold nanoparticles anchored into the gel. This allowed us to create nearly solid silver nanostructures. Dan and I tried several different methods for sintering them, and I eventually discovered the laser sintering approach. Shortly after the silver methods were developed, I became more involved in other projects and we agreed that Dan would finish the remaining experiments, since I had done more work originally. Dan continued working on the project for over a year, making significant process improvements, preparing the samples for the final conductivity demonstrations in Figure 3 and the ring resonators in Figure 4, while I remained involved in a planning, mentoring, and analysis capacity. Once we completed the conductivity data in Figure 3, I wrote most of the manuscript, and we divided the work of making the figures. To Dan's credit, I advocated for submitting to Nature Nanotechnology rather than Science, but Dan prevailed on me that it was worth a shot at Science. The reviews were positive, and Dan finished revisions in 8 days.

Implosion Fabrication is unique among my projects for having been developed in a "working forwards" fashion, i.e., starting with the idea for the technology (reversing ExM) rather than an idea for the problem. The initial idea of using it to position DNA origami seemed unworkably complex, and we discovered in the process of the technology that high resolution is secondary in most nanofabrication applications to material control. Luckily, Implosion Fabrication gives both material control and resolution. But even at the time of the first paper, the application was unclear. We now believe that the key feature of ImpFab that other methods lack is the ability to

pattern materials with a density gradient. By patterning materials with high refractive-index contrast in a gradient pattern, we can make flat optical elements. It is an exciting success story that we ended up (unintentionally) with a technology capable in principle of achieving that goal. However, if we had started with that goal in mind, we might have taken a more direct route to the goal.

Summary

Lithographic nanofabrication is often limited to successive fabrication of two-dimensional layers. We present a strategy for the direct assembly of three-dimensional nanomaterials consisting of metals, semiconductors, and biomolecules arranged in virtually any three-dimensional geometry. We use hydrogels as scaffolds for volumetric deposition of materials at defined points in space. We then optically pattern these scaffolds in three dimensions, attach one or more functional materials, and then shrink and dehydrate them in a controlled way to achieve nanoscale feature sizes in a solid substrate. We demonstrate this process, Implosion Fabrication (ImpFab), by directly writing highly conductive, 3D silver nanostructures within an acrylic scaffold using a volumetric silver deposition process, achieving resolutions in the tens of nanometers and complex, non-self-supporting 3D geometries of interest for optical metamaterials.

Introduction

Most nanofabrication techniques currently rely on 2- and 2.5-dimensional patterning strategies. Although popular direct laser writing methods allow for the single-step fabrication of self-supporting, polymeric 3D nanostructures (*134–141*), arbitrary 3-D nanostructures (e.g., solid spheres of metal, or metallic wires arranged in discontinuous patterns) are not possible (*142*, *143*). This raises the question of whether a versatile 3D nanofabrication strategy could be developed that would allow independent control over the geometry, feature size, and chemical composition of the final material.

A hallmark of 2D nanofabrication strategies is that materials are deposited in a planar fashion onto a patterned surface. By analogy, we reasoned that a general 3D nanofabrication strategy could involve deposition of materials in a volumetric fashion into a patterned scaffold. However, such scaffolds face a fundamental tension: they should be porous and solvated, to allow for introduction of reagents to their interior, while also being dense, to allow material placement with nanoscale precision. To resolve this contradiction, we reasoned that an ideal scaffold could be patterned in a solvated state, and then collapsed and desiccated in a controlled way, densifying the patterned materials to obtain nanoscale feature sizes. Although several groups have previously experimented with shrinking materials, the shrinking process typically requires harsh conditions and chemical changes that may destroy functional materials (*144–146*). We use polyacrylate/polyacrylamide hydrogels for the scaffold material, as they have pore sizes in the range of 10nm to 100nm (*147*), are known for their ability to expand and shrink up to ~10-fold in linear dimension (*37*, *148–150*), and methods for optically patterning hydrogels are well-established (*131–133*, *151*, *152*).

Results

Our implementation takes place in three phases (see **Methods**). It has previously been found that two-photon excitation of fluorescein within acrylate hydrogels causes the fluorescein to react to

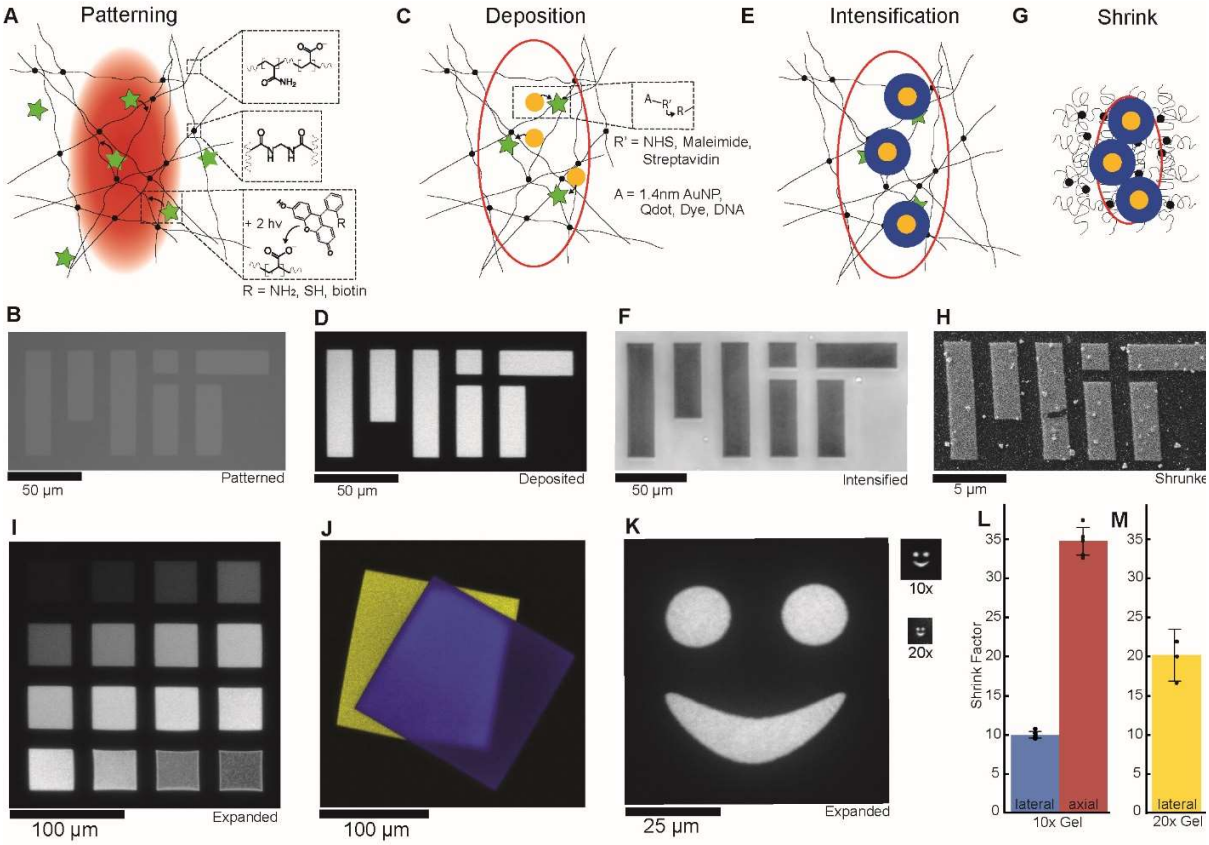

Figure 3-1: Implosion fabrication (ImpFab) process. (A) Schematic of the patterning process, showing the expanded polyelectrolyte gel (black lines and dots, top insets), and fluorescein (green star, bottom inset) binding covalently to the polymer matrix upon multi-photon excitation (red volume). Not to scale. Fluorescein bears a reactive group, R. **(B)** Residual fluorescence of patterned fluorescein immediately following patterning. **(C)** Schematic of functionalization of patterned gel by attaching small molecules, proteins, DNA or nanoparticles to reactive R groups from (A). Red outline indicates patterned volume in (A). **(D)** Image of fluorescent streptavidin nanogold conjugates attached to the pattern in (B). **(E)** Schematic of the volumetric deposition process, showing growth of silver (blue) on top of gold nanoparticles within the hydrogel matrix. **(F)** Image of silver deposited onto the pattern in (D) by transmission optical microscopy. Following silver growth, the pattern has high optical density. **(G)** Schematic of the shrinking and dehydration process. **(H)** SEM image of the silverized pattern from (F) following shrinking and dehydration. **(I)** Fluorescent patterns created with different laser powers (see **Methods**). **(J)** Image of a gel patterned with both metal nanoparticles (yellow) and CdTe quantum dots (blue) in different locations. **(K)** Images of fluorescent patterns before shrinking (left, 10x gel), after shrinking and dehydration in a 10x gel (top right), and after shrinking and dehydration in a 20x gel (bottom right). **(L)** The mean lateral (blue) and axial (red) shrink factors (initial size/final size) for 10x gels (n = 6), including dehydration. **(M)** The mean lateral shrink factor for 20x gels (yellow, n = 3). Error bars show s.d.

the hydrogel (*131–133*). We take advantage of this phenomenon to attach fluorescein molecules carrying reactive groups to the expanded gel in defined three-dimensional patterns (Figure

3-1A,B). In the second phase, following removal of the fluorescein patterning solution, the gel is functionalized by depositing materials onto the patterned reactive groups (Figure 3-1C,D), using

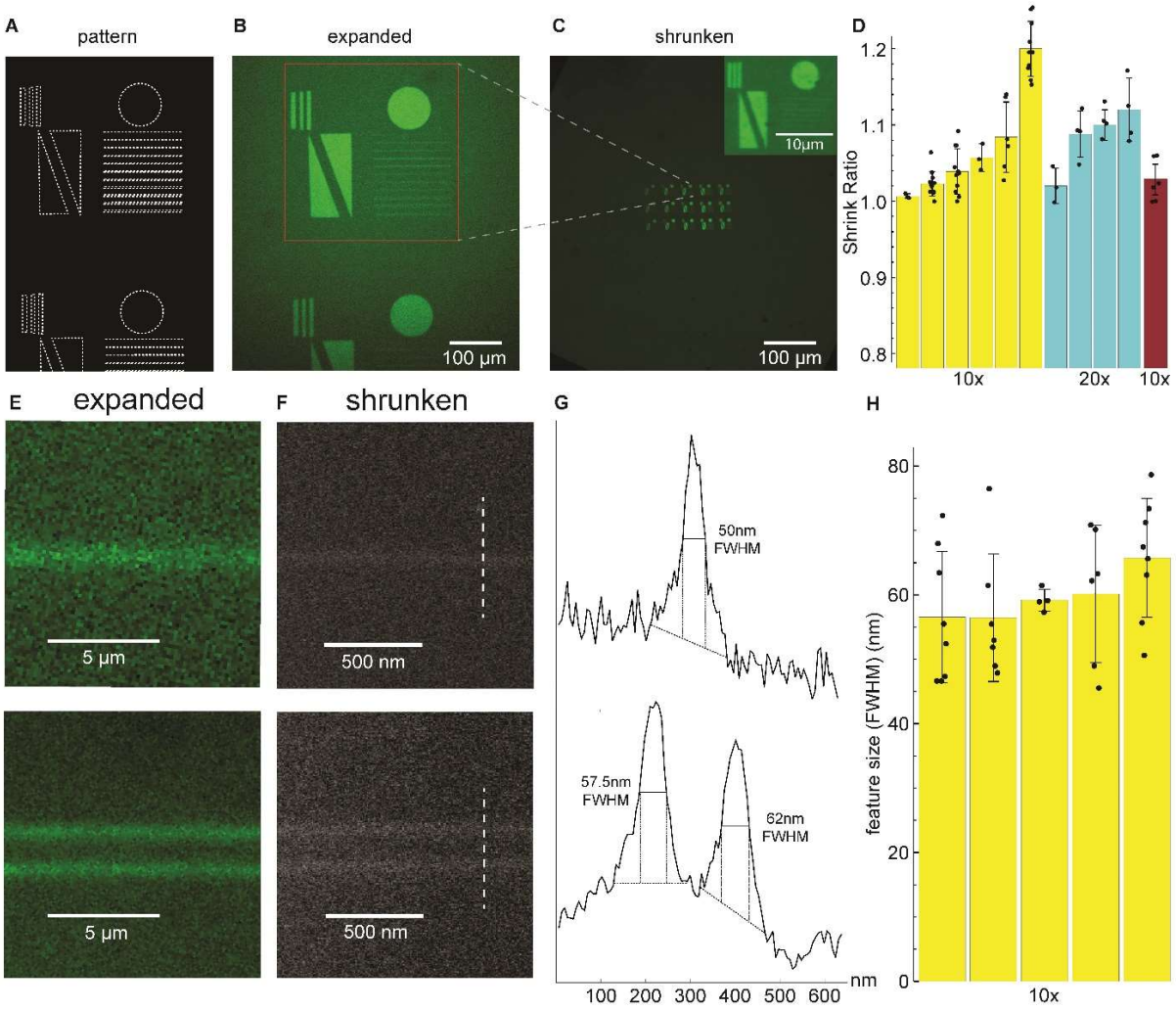

Figure 3-2: Resolution of implosion fabrication. (A) Design of the resolution test pattern including pairs of single-voxel-thick lines (bottom right). (B) Fluorescence image of the patterns from (A). (C) Fluorescence image of the pattern (from B) after shrinking. (D) Measures of isotropy in lateral and axial dimensions. Yellow and blue bars represent lateral isotropy for 10x gels and 20x gels, respectively, and the red bar represents axial isotropy for 10x gels. (E) Fluorescence images of single-voxel lines before shrinking. (F) Scanning electron microscopy (SEM) images of single-voxel lines after 10x shrinking. The gel was functionalized with gold nanoparticles for contrast. (G) Cross-sectional intensity profiles of the lines imaged by SEM (dashed lines in (F)), showing how full-width half-maxima (FWHM) of single voxel lines were measured. (H) Linewidths, measured in G, for five different gel samples. Dots are measurements for individual lines; bars indicate mean $\pm$ s.d. across individual lines within a single gel.

one of several available conjugation chemistries. This volumetric deposition step defines the composition of the material, and may be followed by additional deposition chemistries

("intensification") to boost the functionality of the deposited molecules or nanomaterials (Figure 3-1E,F). Importantly, the functional molecules or nanoparticles are not present during the patterning process, so the specific physical properties of the molecules or nanoparticles used will not affect the patterning. In the final phase, the patterned and functionalized scaffold is shrunken by a factor of 10 to 20 in each dimension with acid or divalent cations over a period of hours, and dehydrated to achieve the desired nanoscale resolution (Figure 3-1G,H). The scaffold is not removed, as it supports the nanofabricated material and allows for the creation of disconnected or high-aspect-ratio structures that would otherwise collapse outside of the gel.

We found the polyacrylate gel to be a suitable substrate for patterning and deposition. The gel readily accommodates a wide variety of hydrophilic reagents, including small molecules, biomolecules, semiconductor nanoparticles or metal nanoparticles (Figure 9-1A-C). For laser powers below a critical threshold, the density of the deposited functional material is controllable (Figure 3-1I, Figure 9-2). We estimated based on the maximum pattern fluorescence in Figure 9-2A that binding sites are patterned into the gel at concentrations of at least 79.2μM in the expanded state, leading to a final concentration in the shrunken state of greater than 272.0mM or 1.64×10^{20} sites per cubic centimeter for a 10x gel (see below). By repeating our patterning and deposition process, we were able to deposit multiple materials in different patterns in the same substrate, such as gold nanoparticles and cadmium telluride nanoparticles (Figure 3-1J). We observed by fluorescence that the deposition of the second material onto the first pattern was at most 18.5% of the deposition of the second material onto the second pattern, confirming that multiple materials may be independently patterned and deposited using this process (Figure 9-3).

The shrinking process is performed either by exposing the expanded gel to hydrochloric acid or to divalent cations (e.g. magnesium chloride, Figure 9-1A-C). The latter may be useful if the patterned materials are sensitive to acid, although we found that both streptavidin and DNA remain functionally intact following acid shrinking (Figure 9-1D). Gels that are shrunken in hydrochloric acid can subsequently be dehydrated, resulting in additional shrinking, and this process preserves the patterned geometry (Figure 3-1K). The final dehydrated gel is transparent (Figure 9-4A), and atomic force microscopy (AFM) characterization measured the surface roughness over a 1 x 1 μm^2 window to be ~0.19 nm (root-mean square; Figure 9-4B). Except where stated otherwise, all samples described as "shrunken" hereafter are shrunken and dehydrated. We tested two different gel formulations that differ only in cross-linker concentration: "10x" (0.075% cross-linker) and "20x" (0.0172% cross-linker) (see **Methods**). The 10x gels, and the patterns within, shrink consistently by a linear factor of 10.6 ± 0.8 in the lateral dimension (mean ± s.d., n=5 gels) and 34.8 ± 1.8 in the axial dimension (n=6 gels, Figure 3-1L), with the disproportionate axial shrink occurring during dehydration, possibly due to surface interactions between the shrinking polymer and the surface of the glass container. For the 20x gels, we observed 20.1 ± 2.9-fold shrink in the lateral dimension (n=4 gels, Figure 3-1M). The 20x gel

formulation is challenging to handle manually due to its delicacy, so the axial shrink factor was not measured, and they were not used further, except for distortion measurements.

To validate the minimum feature size of ImpFab, we designed a test pattern containing pairs of single-voxel-wide lines (Figure 3-2A-D). Since such post-shrink features are necessarily below the optical diffraction limit, we deposited gold nanoparticles and employed scanning electron microscopy (SEM) to assess the resolution after shrinking. We estimated the resolution by measuring the line width (full width at half maximum, FWHM) (Figure 3-2E-G), and obtained a value of 59.6 ± 3.8 nm (mean ± s.d. across samples; n = 5; Fig. 2H) for 10x gels. The mean within-sample standard deviation of the line width was 8.3 nm. We estimated the isotropy of the

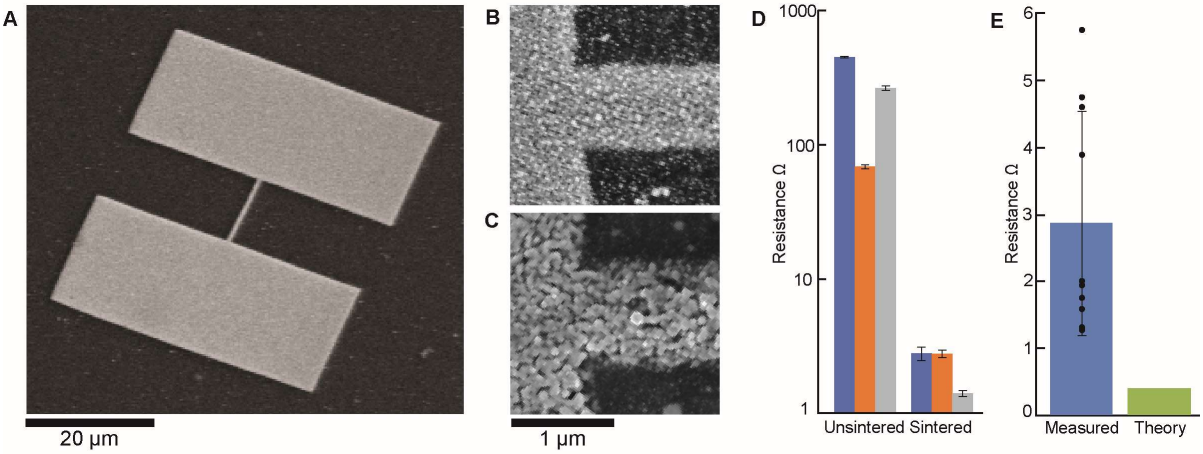

Figure 3-3: Characterization of silver conductivity. (A) SEM overview of a shrunken silver wire between two landing pads, prior to sintering. **(B)** SEM image of wires before and **(C)** after sintering. **(D)** Resistance of three separate conductive pads, of dimension 35x35 µm, measured before and after sintering. Each color represents a single conductive pad. Error bars show standard error in a four-point conductivity measurement. **(E)** Resistance of individual sintered wires (black dots), their mean (blue), and standard deviation, as compared to the theoretical conductivity of a similar structure made of bulk silver (green).

shrinking process by calculating the ratio of the longest diameter of the patterned circle to the orthogonal diameter (Figure 3-2C, D). The percent distortion thus calculated was 6.8 ± 6.9% for 10x gels (mean ± s.d., n=6 gels), and 8.2 ± 4.3% for 20x gels (n=4 gels). We found that the ratio of axial to lateral shrink was on average within 3.1 ± 2.5% of the mean of this ratio (n=6 10x gels), indicating that the disproportionate axial shrink is highly reproducible. Thus, it is possible to account for the disproportionate axial shrink in the design of the pattern. To illustrate this point with the fabrication of a cube, we patterned a rectangular prism and imaged it before and after dehydration (Figure 9-5). As expected, the rectangular prism contracts in the axial dimension during the dehydration step and turns into a cube.

Since nanoscale metal structures are broadly important in fields such as nanophotonics, metamaterials, and plasmonics, we applied ImpFab to create conductive silver structures. We anchored gold nanoparticles to patterned amines via a biotin-streptavidin linkage (see **Methods**). We were initially unable to deposit gold nanoparticles at high enough concentrations to form conductive structures. We thus developed an intensification process based on photographic intensification chemistries, in which silver is deposited onto the surface of gel-anchored gold nanoparticles in aqueous phase while the gel is in the expanded state (Figure 3-1E, F). Finally, the gel is treated with a chelating agent to remove any remaining dissolved silver, and is then shrunken via exposure to hydrochloric acid and subsequent dehydration.

Even with the silver intensification process, wire structures fabricated using the method above (Figure 3-3A) were not reliably conductive, or had resistances on the order of hundreds of ohms. We tested several different methods of sintering, including treatment with oxygen plasma, electrical discharge, and heating the sample to ~500 degrees in an oven. However, none of these methods resulted in well-preserved and evenly sintered silver structures. Instead, we found that the silver patterns could be sintered effectively using the same 2-photon setup used for the initial

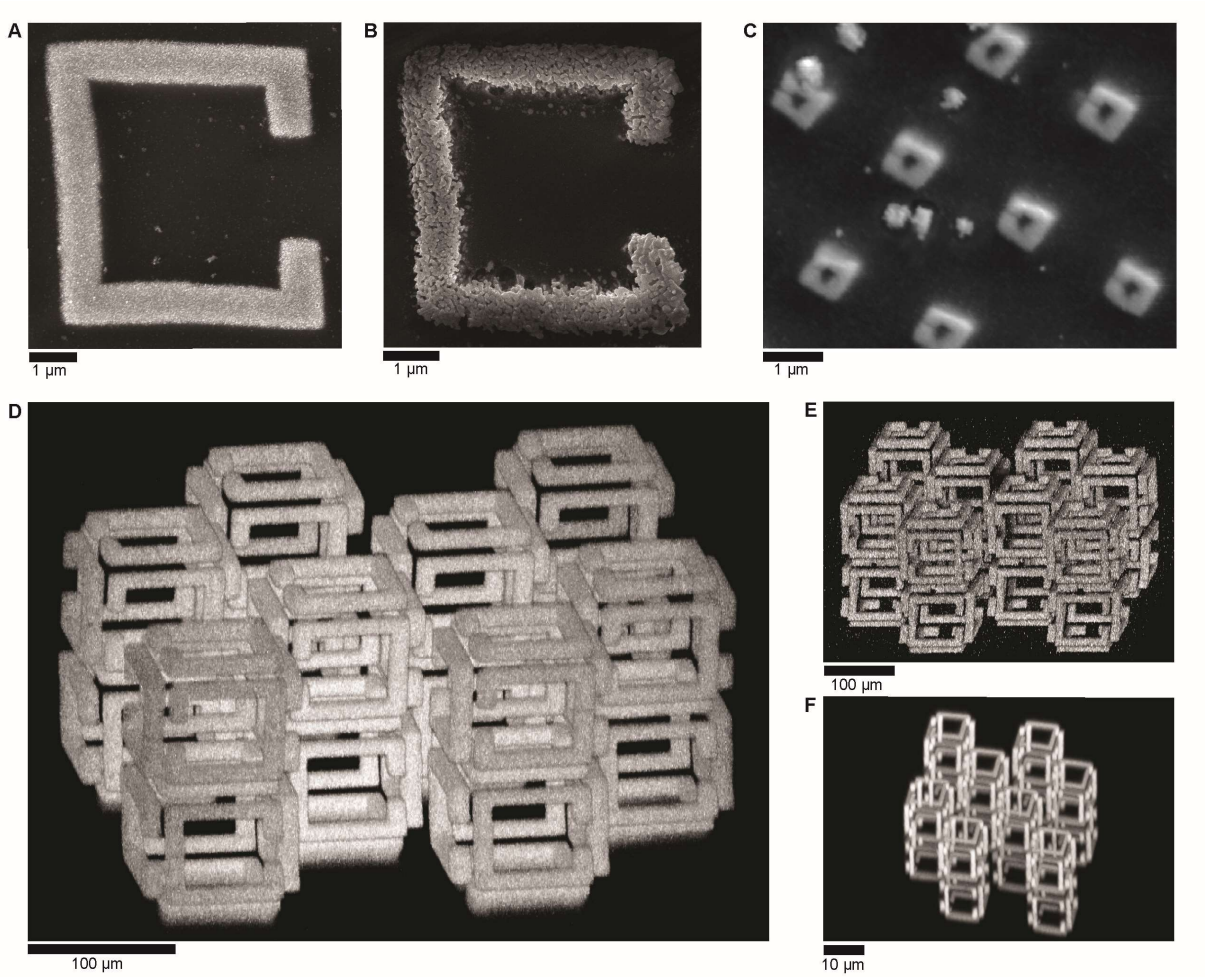

Figure 3-4: Fabrication of 3D metal nanostructures. (A) Two-dimensional structures fabricated with ImpFab with micron-scale resolution, before and **(B)** after sintering, visualized using SEM. **(C)** Similar structures fabricated with hundred-nanometer feature size, after shrinking and dehydration but before sintering. **(D)** Maximum intensity projection of fluorescence image of a 3D structure prior to shrinking (*135*, *351*). **(E)** Maximum intensity projection of a reflected light image from the same structure following volumetric silver deposition, prior to shrinking. **(F)** Maximum intensity projection of a fluorescence image of the same structure, shrunken but not dehydrated.

photopatterning step. We found that samples irradiated at relatively low power levels (see **Methods,** chapter 9) showed a distinct change in the morphology of the embedded silver

nanoparticles consistent with sintering (Figure 3-3B,C). We measured the conductivity of three patterned silver squares both before and after sintering, and found that the resistance of each square decreased by 20-200 fold (Figure 3-3D). Sintered wires were measured in a 4-point probe system, and were found to have linear IV curves (Figure 9-6). Wires sintered in this way had an average resistance of 2.85 ± 1.68Ω (mean ± standard deviation; n = 10), with the resistance depending on the density of the patterned silver (Figure 9-6B). By contrast, an ideal silver wire with the same geometry would have a resistance of 0.38Ω, suggesting that our sintered structures achieved a mean conductivity 13.3% that of bulk silver, with individual samples obtaining conductivities as high as 30% that of bulk silver (Figure 3-3E).

To verify that our method is compatible with a wide range of 3D geometries, we fabricated structures with dimensions ranging from hundreds of nanometers to several microns (Figure 3-4A-C). We found that these structures retain their morphology following sintering (Figure 3-4B). We fabricated a non-layered, non-connected three-dimensional structure comprised of many 2D substructures arranged at different angles relative to each other in space, which would not lend itself to fabrication by other means (Figure 3-4D). Whereas our previous experiments had only observed the fabrication of two-dimensional silver structures, we used confocal reflection microscopy to confirm that silver was deposited throughout the volume of the 3D pattern (Figure 3-4E). Finally, using confocal microscopy, we were able to validate that the structure retained its shape following shrinking (Figure 3-4F). Due to the modular nature of ImpFab, the extension of the ImpFab strategy to other kinds of materials, such as other semiconductors or metals, only requires the development of an aqueous deposition chemistry that is compatible with the gel substrate. Future iterations may use alternative chemistries, such as dendrimeric complexes for direct deposition of metals or semiconductors within the hydrogel (*153*, *154*), or DNA-addressed material deposition (*155*). Finally, we note that although we used a conventional microscope that is not optimized for patterning, and that was limited to a 4cm/s scan speed (in post-shrink dimensions), we were able to create objects spanning hundreds of microns to millimeters (Figure 9-7). Using faster patterning systems (*131*), ImpFab could ultimately enable the creation of centimeter-scale nanomaterials.

Chapter 4
Slide-seq

My understanding is that Slide-seq began as an idea that Evan Macosko proposed to Fei Chen, although Bob also claims to have thought of it before he met Evan. Fei presented it to me in spring of 2017, and I was drawn to it by the promise that we could use it for high-throughput projectomics, for example by combining it with methods for barcoding RNA (*16*). I am still interested in that application, although the application to create comprehensive 3D maps of gene expression is equally compelling.

At the beginning of the project, it was determined that Bob would work out the library preparation protocol, while I would work out how to make and sequence the pucks. Sequencing the pucks and developing the algorithms for reconstructing the images turned out to be relatively straightforward, but the library preparation was challenging. Bob came up with two innovations that each improved our RNA yield by an order of magnitude: firstly, he discovered in November 2017 that using liquid tape, rather than acrylamide, as a bind surface resulted in a ~5X-10X improvement in RNA yield. Then, although the hybridization step was initially performed dry, Bob had the idea to do the hybridization in 6X SSC, which gave a further ~5X improvement in yield. For whatever reason, the RT buffer itself was insufficient to enable RNA to bind to the beads.

As the project progressed, I specialized more in the data analysis, and Bob specialized more on the wet lab protocol and sample processing. Josh Welch had the idea to apply Liger (now in press) in order to determine the cell-type composition of each bead. After experimenting with Liger, I realized that Liger assigns beads to consensus cell types derived from both datasets together. This led to the cell types on each Slide-seq puck being slightly different, depending on the composition of the puck. However, what we really wanted was to map the cell types from the Slide-seq dataset onto fixed cell types derived from the single cell RNA sequencing data. Modifying the Liger algorithm to hold the cell types in the scRNAseq dataset constant resulted in a regression problem in the lower dimensional space provided by the NMF decomposition in Liger. I proposed this idea to Aleksandrina Goeva, and she developed and implemented it, resulting in the NMFReg algorithm that allows cell types from scRNAseq to be mapped onto the puck.

One key insight I had was that because Slide-seq data is intrinsically mixed, i.e. because many beads have RNA contributions from multiple cell types, Slide-seq is much more powerful when one uses *genes* as the primitive analysis object, rather than cell types. For example, Slide-seq is not useful for cell-type discovery or for differential expression between cell types: if one examines the differential expression between astrocytes in one location and astrocytes in another location, for

example, the genes one finds are typically contaminating genes from adjacent cell types. Moreover, because Slide-seq data is relatively sparse, most beads of a given cell type will fail to display most of the key markers for that cell type. However, Slide-seq has high statistical power for detecting spatial patterns of gene expression in a cell-agnostic way *de novo*. For example, if there are only 4 occurrences of a particular gene, but all 4 occurrences are immediately adjacent to each other on the puck, that is an extremely strong statistical signal, regardless of the cell-types of the beads on which they occur. This realization led me to develop the other two core analysis algorithms (besides NMFReg) that appear in the paper: the spatially significant gene calling algorithm, which detects spatially non-random distributions of gene expression, and the gene overlap algorithm, which determines when two genes are spatially correlated. These two algorithms served in turn as the basis for the analysis in Figures 3 and 4.

Summary

Spatial positions of cells in tissues strongly influence function, yet a high-throughput, genome-wide readout of gene expression with cellular resolution is lacking. We developed Slide-seq, a method for transferring RNA from tissue sections onto a surface covered in DNA-barcoded beads with known positions, allowing the locations of the RNA to be inferred by sequencing. Using Slide-seq, we localized cell types identified by scRNA-seq datasets within the cerebellum and hippocampus, characterized spatial gene expression patterns in the Purkinje layer of mouse cerebellum, and defined the temporal evolution of cell-type-specific responses in a mouse model of traumatic brain injury. These studies highlight how Slide-seq provides a scalable method for obtaining spatially resolved gene expression data at resolutions comparable to the sizes of individual cells.

Introduction

The functions of complex tissues are fundamentally tied to the organization of their resident cell types. Multiplexed *in situ* hybridization and sequencing-based approaches can measure gene expression with subcellular spatial resolution (*9*, *10*), but require specialized knowledge and equipment, as well as the upfront selection of gene sets for measurement. By contrast, technologies for spatially encoded RNA-sequencing with barcoded oligonucleotide capture arrays are limited to resolutions in the hundreds of microns (*156*), which is insufficient to detect important tissue features.

Results

To develop Slide-seq for high-resolution genome-wide expression analysis, we first packed uniquely DNA-barcoded 10 µm microparticles ('beads') —similar to those used in the Drop-seq approach to scRNA-seq (*25*)—onto a rubber-coated glass coverslip forming a monolayer we termed a "puck" (Figure 10-1). We found that each bead barcode sequence could be uniquely determined via SOLiD sequencing-by-ligation chemistry (Figure 4-1A, Figure 10-1) (*157*, *158*) (see **Methods**). We next developed a protocol wherein 10 µm fresh-frozen tissue sections were transferred onto the dried bead surface via cryosectioning (see **Methods**, Chapter 10). mRNA released from the tissue was captured onto the beads for preparation of 3'-end, barcoded RNA-seq libraries (*25*) (Figure 4-1B). Clustering of individual bead profiles from a coronal section of mouse hippocampus (see **Methods**, Chapter 10) yielded assignments reflecting known positions of cell types in the tissue (Figure 4-1C). Very fine spatial features were resolved, including the single-cell ependymal cell layer between the central ventricle and the habenula in the mouse brain (Figure 4-1C, inset). Moreover, Slide-seq could be applied to a range of tissues, including the cerebellum and olfactory bulb, where layered tissue architectures were immediately detectable (Figure 4-1D, Figure 10-2), as well as liver and kidney, where the identified clusters revealed hepatocyte zonation patterns (*159*) and the cellular constituents of the nephron, respectively. Slide-seq on postmortem human cerebellum was also successful in capturing the same architectural features observed in the cognate mouse tissue (Figure 10-3). Expression measurements by Slide-seq agreed with those from bulk

mRNA-seq and scRNA-seq, and average mRNA transcript capture per cell was consistent across tissues and experiments (Figure 10-4). Finally, we found no detectable difference in the dimensions of brain structures observed in Slide-seq and in FISH (Figure 10-5), implying that mRNA is transferred from the tissue to the beads with minimal lateral diffusion.

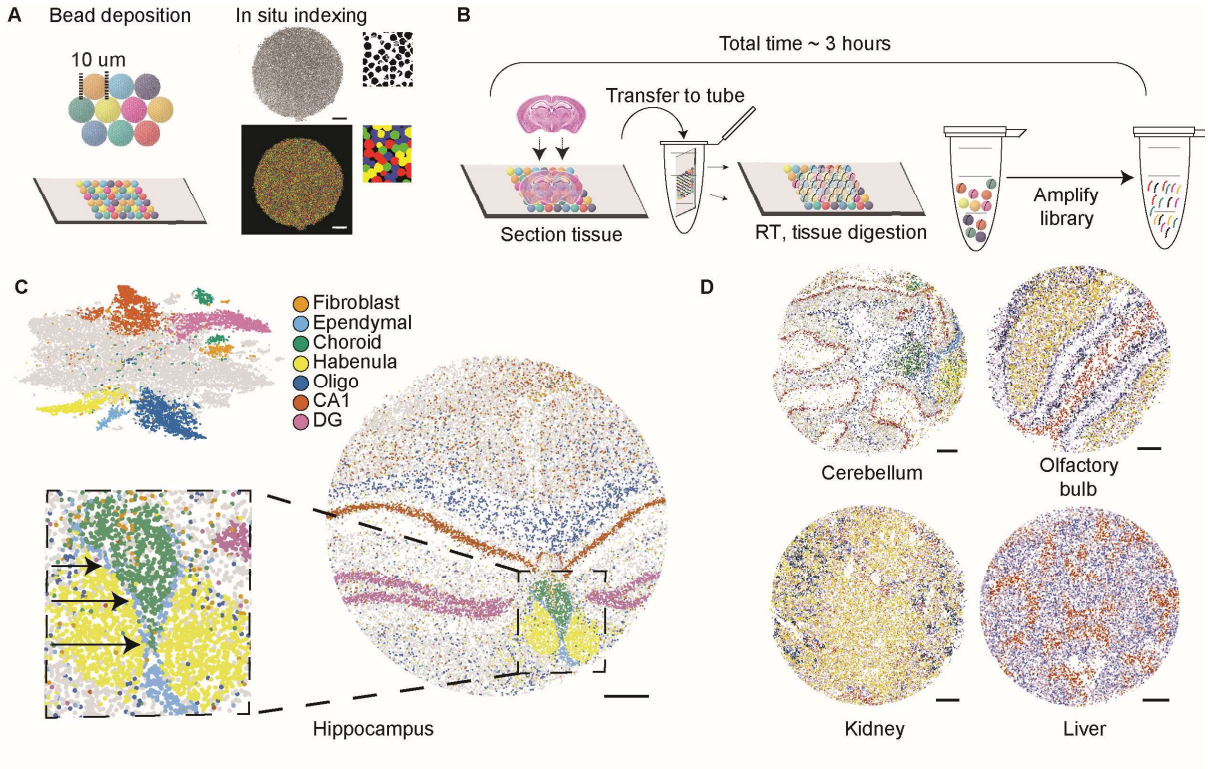

Figure 4-1: High-resolution RNA capture from tissue by Slide-seq. (A) Left: Schematic of array generation. A monolayer of randomly deposited, DNA barcoded beads (termed a "puck") is spatially indexed by SOLiD sequencing. Top Right: A representative puck with sequenced barcodes shown in black. Bottom Right: A composite image of the same puck colored by the base calls for a single base of SOLiD sequencing. **(B)** Schematic of the sample preparation procedure developed for Slide-seq. **(C)** Top left: tSNE representation of Slide-seq beads from a coronal mouse hippocampus slice with colors indicating clusters. Right: the spatial position of each bead is shown, colored by the cluster assignments shown in the tSNE. Bottom left: Inset indicating the position of a single-cell-thickness ependymal cell layer (black arrow). **(D)** As in (C), but for the indicated tissue type (see Figure 10-2 for clustering and cluster identities). All scale bars 500 µm.

To map scRNA-seq cell types onto Slide-seq data, we developed a computational approach called Non-negative Matrix Factorization Regression (NMFreg) that reconstructs expression of each Slide-seq bead as a weighted combination of cell-type signatures defined by scRNA-seq (Figure 4-2A). Application of NMFreg to a coronal mouse cerebellar puck recapitulated the spatial distributions of classical neuronal and non-neuronal cell types (*24*), such as granule cells, Golgi interneurons, unipolar brush cells, Purkinje cells, and oligodendrocytes (Figure 4-2B, Figure 10-6A). The mapping by NMFreg was found to be reliable across a range of factor numbers and

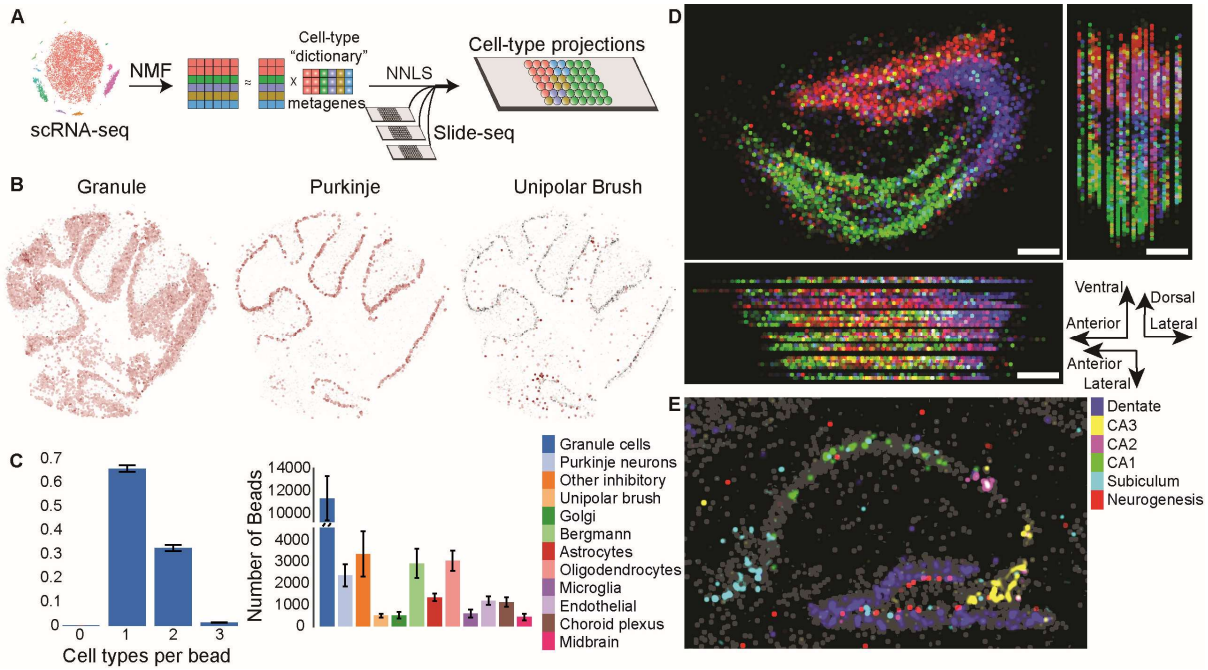

Figure 4-2: Localization of cell types in cerebellum and hippocampus using Slide-seq. (A) Schematic for assigning cell types from scRNA-seq datasets to Slide-seq beads using NMF and NNLS regression (NMFreg). **(B)** Loadings of individual cell types, defined by scRNA-seq cerebellum (*24*) on each bead of one 3 mm-diameter coronal cerebellar puck (red, cell type location, gray, Purkinje loadings plotted as a counterstain). Other cell types are in Figure 10-6. **(C)** Left: Number of cell types assigned per bead (Figure 10-7). Right: The number of beads called as each scRNAseq-defined cell type for cerebellar pucks (mean ± std. N=7 pucks). **(D)** Projections of hippocampal volume with NMFreg cell type calls for CA1 (green), CA2/3 (blue) and dentate gyrus (Red). Top left: Sagittal projection. Top right: Coronal projection. Bottom left: Horizontal projection. Bottom right: axis orientations for each of the projections. **(E)** Composite image of metagenes for six different cell types. All scale bars 250 µm. All metagenes are listed in Table 10-2.

random restarts (Figure 10-6B,C). We found that 65.8% +/- 1.4% of beads could be identified with a single cell type (see **Methods**, Chapter 10), whereas 32.6% +/- 1.2% showed mRNA from two cell types (mean ± std, N=7 cerebellar pucks) (Figs. 2C, S7). The high spatial resolution of Slide-seq was key to mapping cell types: when data were aggregated into larger feature sizes, cell types in heterogeneous regions of tissue could not be resolved (Fig. S8). Slide-seq collects a 2D spatial sample of 3D tissue volumes, thus caution should be taken when making absolute counting measurements throughout the 3D volume in the absence of proper stereological controls and sampling methods (*160*).

We first sequenced pucks capturing 66 sagittal tissue sections from a single dorsal mouse hippocampus (20 billion paired-end reads over 1.5 million barcoded beads), covering a volume of 39 cubic millimeters, with roughly 10 µm resolution in the dorsal-ventral and anterior-posterior axes, and ~20 µm resolution (alternate 10 µm sections) in medial-lateral axis (Figure 10-9A-D).

Using NMFreg, 770,000 beads in the volume could be associated with a single scRNA-seq-defined cell type. We computationally co-registered pucks along the medial-lateral axis, allowing for visualization of the cell types and gene expression in the hippocampus in three dimensions (Figure 4-2D, Figure 10-9E,F). We plotted metagenes comprised of previously defined markers (*24*) for the dentate gyrus, CA2, CA3, a subiculum subpopulation, an anteriorly localized CA1 subset (exemplified by the marker Tenm3) and cells undergoing mitosis and neurogenesis. The metagenes were highly expressed and specific for the expected regions (Figure 4-2E), confirming the ability of Slide-seq to localize both common cell-types as well as finer cellular subpopulations. The entire experimental processing of these 66 pucks (excluding puck generation) required ~40 person-hours (see **Methods**, Chapter 10), and only standard experimental apparatus.

We then developed a nonparametric, kernel-free algorithm to identify genes with spatially non-random distribution across the puck (Figure 10-10) (see **Methods**, Chapter 10). Application of this algorithm to coronally sliced cerebellum identified Ogfrl1, Prkcd and Atp2b1 as highly localized to a region just inferior to the cerebellum (Fig. S11A). We found Ogfrl1 in particular to be a specific and novel marker for PV interneurons in the molecular and fusiform layers of the dorsal cochlear nucleus (Figure 10-11B), likely the cartwheel cells of the dorsal cochlear nucleus (*161*, *162*). Our algorithm also identified *Rasgrf1* as expressed only in granule cells anterior to the primary fissure (Figure 10-11C, cyan, Figure 10-11D, left) (*38*), and further analysis revealed four previously uncharacterized genes expressed only posterior to the primary fissure (see **Methods**, Chapter 10) (Table 10-2, Figure 10-11C, yellow, Figure 10-11D, right).

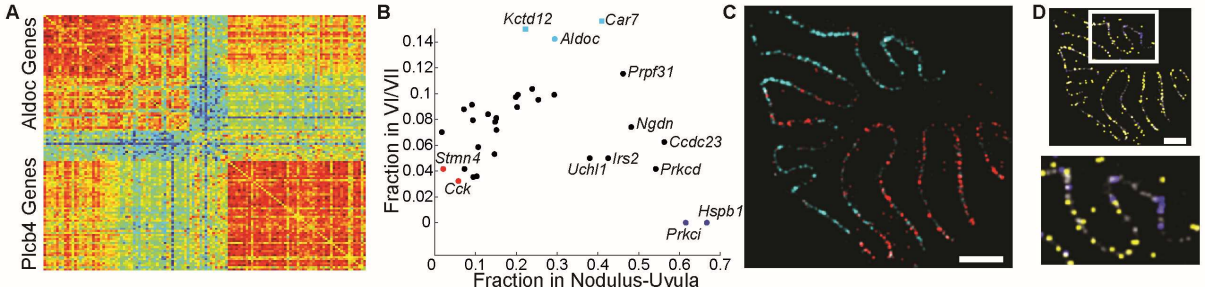

Figure 4-3: Identification of novel variation in cerebellar gene expression by Slide-seq. (A) Heatmap illustrating the separation of Purkinje-expressed genes into two clusters by spatial gene correlation. The *i*,*j*th entry is the number of genes found to overlap with both gene *i* and *j* in the Purkinje cluster (see **Methods**, Chapter 10). **(B)** For genes with significant expression ($p<0.001$, Fisher exact test) in the nodulus-uvula region (see **Methods**, Chapter 10), the fraction of reads localized to the nodulus/uvula and to the VI/VII boundary is shown. *Pcp4*, a ubiquitous marker for Purkinje cells, is in gray. **(C)** An *Aldoc* metagene in cyan. A *Cck* metagene in red. **(D)** A *H2-D1* metagene in yellow. A *Hspb1* metagene in blue. All scale bars show 250 µm. All metagenes are listed in Table 10-2.

The cerebellum is marked by parasagittal bands of gene expression in the Purkinje layer that correlate with heterogeneity in Purkinje cell physiology and projection targets (*163–166*). Several genes, including *Aldoc* (also known as the antigen of the Zebrin II antibody) show similar or

complementary parasagittal expression (*165*, *167*, *168*) but the extent of this form of expression variation is unknown, and these patterns have not previously been identified in single-cell sequencing studies. Using the spatial information afforded by Slide-seq, we identified 669 spatially non-random genes in the Purkinje layer (Table 10-2), of which 126 appeared either correlated or anticorrelated with the Zebrin pattern, using *Aldoc* and *Plcb4* as markers of Zebrin II(+) or Zebrin II(-) bands, respectively (Figure 4-3A). Among the anticorrelated genes were four ATPases and four potassium channels, including some which may explain differences in electrophysiology between Zebrin II(+) and Zebrin II(-) Purkinje neurons (Table 10-2). Moreover, we identified several other patterns of spatial gene expression, besides the Zebrin pattern. While most genes identified displayed a pattern consistent with Zebrin II staining (Figure 4-3B,C), several were exclusively expressed in or excluded from the vestibulocerebellar region (lobules IX and X) (*169*, *170*) (Figure 4-3D, Table 10-2), confirming that lobules IX and X have a distinct program of gene expression. Still other genes showed either exclusive expression in (e.g. *B3galt5* (*171*)) or exclusion from (e.g. *Gnai1*) lobules IX/X and VI/VII (Figure 10-11E,F), suggesting that these regions might share a pattern of gene expression, despite the disparate cognitive roles associated with them (*172*). Finally, although only Purkinje cells have previously been associated with the *Aldoc* pattern, we found that *Mybpc1*, a Bergmann cell marker previously only studied in the context of muscle, appears in both Slide-seq (Figure 10-11G) and *in situ* data (Figure 10-11H) to have a Zebrin pattern of expression. We thus conclude that the banded gene expression patterns divide many cerebellar cell types, including Purkinje cells, Bergmann glia, and granule cells, into spatially defined subpopulations, which was not indicated in previous single-cell sequencing studies (*24*, *173*).

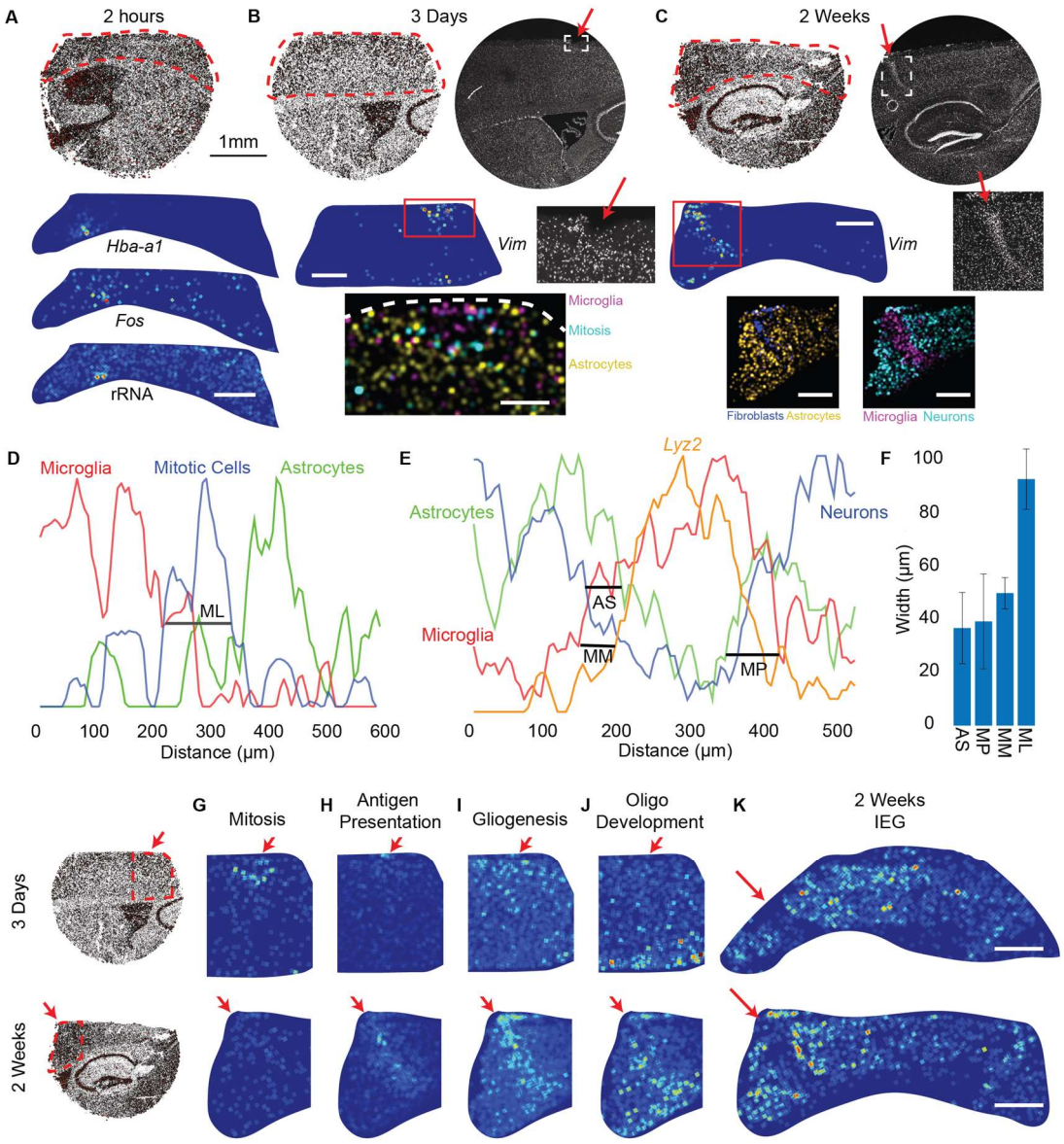

Figure 4-4: Slide-seq identifies local transcriptional responses to injury. (A) Top: All mapped beads for a coronal hippocampal slice from a mouse sacrificed 2 hours after injury, with circle radius proportional to transcripts. Bottom: genes marking the injury. **(B)** As in (A), for a mouse sacrificed 3 days after injury. Top and middle right: DAPI image of an adjacent slice. Panels with black backgrounds show NMFreg cell types as density plots. Scale bar: 250 μm (see **Methods**, Chapter 10). **(C)** As in (B), for a mouse sacrificed 2 weeks following injury. Bottom scale bar: 500 μm. **(D)** Spatial density profiles for the puck in (B) (see **Methods**, Chapter 10). **(E)** Spatial density profiles for the puck in (C). *Lyz2* is plotted as a marker of macrophages. The vertical axis in (D) and (E) represents cell-type density in arbitrary units (see **Methods**, Chapter 10). **(F)** The thickness of the features in (D) and (E) (mean ± std., N=6 for scar, N=6 for penetration, N=3 for mitosis layer). **(G-J)** Gene ontology-derived metagenes for the puck in (B) (top) or (C) (bottom). Warmer colors correspond to greater metagene counts. **(K)** The IEG metagene (Table S2) for two 2-week pucks. Circular images in (A-C) refer to the scale bar in (A). All scale bars for images with blue backgrounds 500 μm. Red arrows indicate the injury.

Cortical injuries were visualized in Slide-seq data by the presence of hemoglobin transcripts 2 hours after the injury (Figure 4-4A), or by transcripts of *Vim*, *Gfap*, and *Ctsd* at 3 days and 2 weeks after the injury (Figure 4-4B,C). *Vim*, *Gfap*, and *Ctsd* were chosen because they are known markers of the astrocytic (*Vim* and *Gfap*) or microglial (*Vim* and *Ctsd*) responses that were found to be highly upregulated at the injury in the Slide-seq data (Figure 10-13). We applied an algorithm to identify all genes that correlate spatially with those transcripts. At the 2-hour timepoint, only *Fos* and rRNA (*174*) were found to correlate spatially with the injury (Figure 4-4A, Figure 10-14). By contrast, at the 3-day timepoint, we found microglia/macrophages-assigned beads localized to the injury, bordered by a distinct layer of cells (thickness: 92.4 µm ± 11.3 µm, mean ± sterr, N=3) expressing mitosis-associated factors, followed by a layer of astrocyte-assigned beads (Figure 4-4D). Finally, at the 2-week timepoint, we observed microglia/macrophage-assigned beads filling the injury, surrounded by an astrocytic scar (thickness: 36.6 µm ± 13.4 µm, mean ± sterr, N=6), with evidence of microglia (but not macrophages) penetrating 39 µm ± 17.8 µm (mean ± sterr, N=6) through the astrocytic scar and into neuron-rich regions (Figure 4-4E,F). Macrophages were visualized using *Lyz2*, a specific marker for macrophages and granulocytes, however, we interpret this as a marker of macrophages, because other granulocyte-specific markers were not found to colocalize with *Gfap*, *Ctsd*, and *Vim*.

In order to investigate other changes in gene expression between the 3-day and 2-week timepoints, we identified genes that correlated spatially with *Vim*, *Gfap*, and *Ctsd* at the 3-day timepoint or the 2-week timepoint (see **Methods**, Chapter 10). Applying gene ontology analysis to these gene sets revealed enrichment of annotations relating to chromatid segregation, mitosis, and cell division at the 3-day timepoint (Figure 4-4G), and relating to the immune response (Figure 4-4H), gliogenesis (Figure 4-4I) and oligodendrocyte development (Figure 4-4J) at the 2-week timepoint. This suggests that cell proliferation occurs in the first few days following injury, and transitions to differentiation on the order of weeks. For example, although the degree to which oligodendrocyte progenitor cells (OPCs) differentiate into oligodendrocytes following a focal gray matter injury is controversial (*175*), we confirmed that both *Sox4* and *Sox10* localize to the region surrounding the injury at the 2 week timepoint, indicating the presence of immature oligodendrocytes (Figure 10-15). We also discovered evidence that several immediate early genes, including highly neuron-specific genes such as *Npas4* (Table 10-2), are upregulated in a region of width 0.72 mm ± 0.19 mm (mean ± sterr, N=4 measurements) around the injury at both the 3-day and the 2-week timepoints (*176–178*) (Figure 4-4K, Table 10-2), suggesting persistent effects of the injury on neural activity in a large area around the injury.

Here we demonstrate that Slide-seq enables the spatial analysis of gene expression in frozen tissue with high spatial resolution and scalability to large tissue volumes. Slide-seq is easily integrated with large-scale scRNA-seq datasets and enables discovery of spatially defined gene expression patterns in normal and diseased tissues. The primary cost of Slide-seq is the cost of

short read sequencing, which is ~$200-$500 for the pucks presented here. As the cost of sequencing drops further, we expect to be able to scale Slide-seq to entire organs or even entire organisms. We anticipate that Slide-seq will play important roles in positioning molecularly defined cell types in complex tissues, and defining new molecular pathways involved in neuropathological states.

Chapter 5
Protein Sequencing

In the winter of 2015, inspired by a series of meetings with Adam Marblestone, Ed Boyden, and others, I took up the question of how one could directly infer the sequence of proteins at the single molecule level. Although mass spectrometry could in principle be applied to single molecules (*179*), the most sensitive protein sequencing methods to date require tens or hundreds of thousands of copies. This approach, which I formulated with inspiration from Adam Marblestone, takes advantage of a set of N-terminal amino acid binders identified by Jim Havranek and Ben Borgo at WUSTL (*180*). My key realization was that although the binders were mostly not specific for any particular N terminal amino acid, their binding spectra were sufficiently different from each other that, by observing the kinetics of each binder for a given peptide, one could likely infer the identity of the N terminal amino acid.

This research was conducted entirely from January 2015 to March 2015. After the initial theoretical research, Andrew Payne, Dan Oran and I tried to implement our ideas experimentally. Our efforts failed, primarily due to lack of experience (I was in my first year, and Andrew and Dan were not even graduate students yet). We were afraid of being scooped on the experiments, so the theoretical work was not published until March 2019. As of the publication of this thesis, there has still been practically no movement in this field, despite numerous theoretical proposals (*181–183*), one major experimental report (*184*), and a company (Encodia Inc.) that appears to have been working in this space for more than 5 years. It is very challenging to distinguish the amino acids from each other on any chemical basis – it is remarkable that cells have evolved enzymes to do it –, and the standard challenges of single-molecule experiments (e.g. nonspecific binding and photobleaching) must likewise be overcome. Nonetheless, I believe that the method laid out here could be made to work, either using the published NAABs or using a dedicated, evolved or engineered set.

That this research was published is entirely due to Adam Marblestone, who unearthed the manuscript in early 2018 and proposed to me and Ed that we submit it. Without Adam's proposal, it would have remained buried.

Summary

We propose and theoretically study an approach to massively parallel single molecule peptide sequencing, based on single molecule measurement of the kinetics of probe binding to the N-termini of immobilized peptides (*180*). Unlike previous proposals, this method is robust to both weak and non-specific probe-target affinities, which we demonstrate by applying the method to a range of randomized affinity matrices consisting of relatively low-quality binders. This suggests a novel principle for proteomic measurement whereby highly non-optimized sets of low-affinity binders could be applicable for protein sequencing, thus shifting the burden of amino acid identification from biomolecular design to readout. Measurement of probe occupancy times, or of time-averaged fluorescence, should allow high-accuracy determination of N-terminal amino acid identity for realistic probe sets. The time-averaged fluorescence method scales well to weakly-binding probes with dissociation constants of tens or hundreds of micromolar, and bypasses photobleaching limitations associated with other fluorescence-based approaches to protein sequencing. We argue that this method could lead to an approach with single amino acid resolution and the ability to distinguish many canonical and modified amino acids, even using highly non-optimized probe sets. This readout method should expand the design space for single molecule peptide sequencing by removing constraints on the properties of the fluorescent binding probes.

Introduction:

Massively parallel DNA sequencing has revolutionized the biological sciences (*185*, *186*), but no comparable technology exists for massively parallel sequencing of proteins. The most widely used DNA sequencing methods rely critically on the ability to locally amplify (i.e., copy) single DNA molecules—whether on a surface (*187*), attached to a bead (*188*), or anchored inside a hydrogel matrix (*189*)—to create a localized population of copies of the parent single DNA molecule. The copies can be probed in unison to achieve a strong, yet localized, fluorescent signal for readout via simple optics and standard cameras. For protein sequencing, on the other hand, there is no protein 'copy machine' analogous to a DNA polymerase, which could perform such localized signal amplification. Thus, protein sequencing remains truly a single molecule problem. While true single molecule DNA sequencing approaches exist (*190–192*), these often also rely on polymerase-based DNA copying, although direct reading of single nucleic acid molecules is beginning to become possible with nanopore approaches (*193*) that may be extensible to protein readout (*194–196*) Thus, the development of a massively parallel protein sequencing technology may benefit from novel concepts for the readout of sequence information from single molecules.

Previously proposed approaches to massively parallel single molecule protein sequencing (*181*, *182*, *197*) utilize designs that rely on covalent chemical modification of specific amino acids along the chain. Such chain-internal tagging reactions are currently available only for a small subset of the

20 amino acids, and they have finite efficiency. Thus, such approaches would likely not be able to read the identity of every amino acid along the chain.

An alternative approach to protein sequencing (*180*, *198–200*) is to use successive rounds of probing with N-terminal-specific amino-acid binders (NAABs) (*180*). Recent studies have proposed that proteins derived from N-terminal-specific enzymes such as aminopeptidases (*201*), or from antibodies against the PITC-modified N-termini arising during Edman degradation (*202*), could be used as NAABs for protein sequencing. Yet designing or evolving highly specific, strong N-terminal binders to all 20 amino acids (and more if post-translational modifications, e.g., phosphorylation, are considered) is a challenge. Rather than attempting to improve the properties of the NAABs themselves, we will introduce a strategy—which we term "spectral sequencing"—to work around the limitations of existing NAABs and enable single molecule protein sequencing without the need to develop novel binding reagents.

Spectral sequencing measures the affinities of many low-affinity, relatively non-specific NAABs, collectively determining a "spectrum" or "profile" of affinity across binders, for each of the N-terminal amino acids. This profile is sufficient to determine the identity of the N-terminal amino acid. Thus, rather than requiring individual binders to be specific in and of themselves, we will infer a specific profile by *combining measurements of many non-specific interactions*. The spectral sequencing approach measures the single molecule binding kinetics in a massively parallel fashion, using a generalization of Points Accumulation for Imaging in Nanoscale Topography (PAINT) techniques (*203*, *204*) to N-terminal amino acid binders. A key advantage of this technique is that it overcomes photobleaching limitations previously observed with fluorescence-based single-molecule protein sequencing methods (*184*).

In what follows, we first derive the capabilities of single-molecule fluorescence based measurement of probe binding kinetics as a function of probe properties and noise sources. We then apply this analysis to the problem of sequencing proteins by measuring profiles of NAAB binding kinetics.

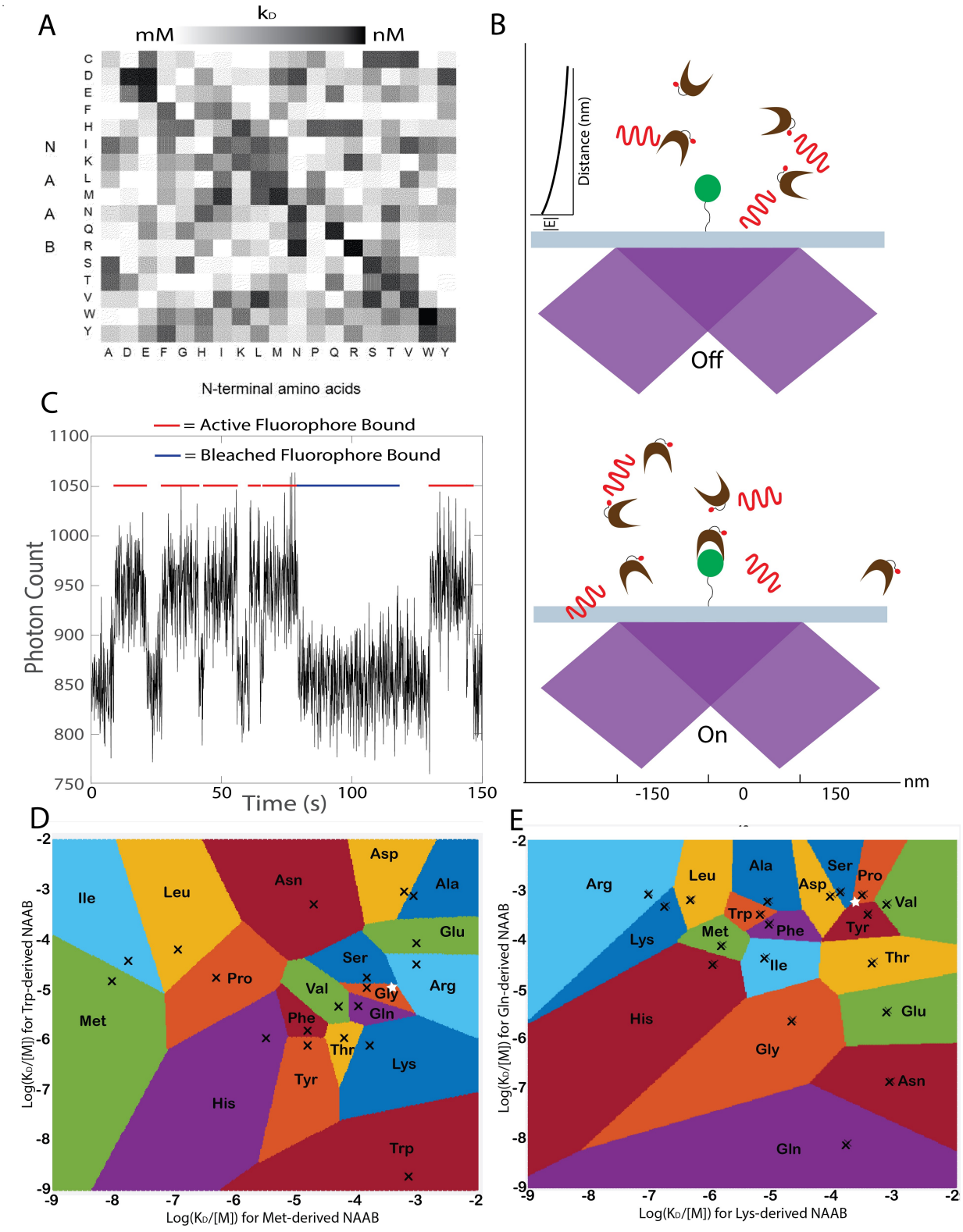

Figure 5-1: Identifying amino acids from kinetic measurements. Caption on next page.

from the existing measured NAAB kinetics (*180*), we estimate via simulation that the kinetic measurement scheme presented here could achieve 97.5% percent accuracy in amino acid identification over a total observation period of 35 minutes, even in the presence of errors arising from instrument calibration or variation in the underlying kinetics of the binders due to the effects of non-terminal amino acids.

Problem Overview

We consider the problem in which a set of peptides is immobilized on a surface and imaged using total internal reflection fluorescence (TIRF) microscopy. The surface must be appropriately passivated to minimize nonspecific binding (*183*, *200*, *205–210*). Moreover, an appropriate method must be selected for anchoring peptides to the surface. We assume that the reactive thiol group of cysteine is used to anchor peptides to the surface, but alternative methods, such as anchoring the C-terminal carboxylic acid to the surface, are also possible (*184*). In all that follows, we will assume that cysteine is used to anchor the peptides to the surface, in which case the sequencing ends at the anchored cysteine.

(A) Example affinity matrix for a set of NAABs. The affinities of each of the 17 NAABs are shown for all 19 amino acids excluding cysteine, which is used to anchor the peptides to the surface. Reproduced from (*180*). **(B)** In the proposed measurement scheme, the target (green disk) is attached to a glass slide and is observed using TIRF microscopy. NAAB binders (brown clefts) bearing fluorophores (red dots) are excited by a TIRF beam (purple) and generate fluorescent photon emissions (red waves). **(C)** When a fluorophore is bound, there is an increase in fluorescence in the spot containing the target. Photobleaching of the fluorophore is indistinguishable from unbinding events, so it is important to use a dye that is robust against photobleaching. Plot shows an illustrative stochastic kinetics simulation incorporating Poisson shot noise of photon emission. A relatively strong binder is shown solely for purposes of illustration. In practice, the method relies on many measurements performed on weak binders. **(D)** The plot shows the result of a proposed kinetic measurement on an N-terminal amino acid using only two NAABs. The affinity of each N-terminal amino acid (black Xs, excluding cysteine) for the methionine-targeting and tryptophan-targeting NAABs are shown as a scatterplot, with the affinity for the met-targeting NAAB on the x axis and the affinity for the Trp-targeting NAAB on the y axis. Upon measuring the affinities for these NAABs against an unknown target undergoing sequencing, the unknown target can be identified with the amino acid with expected vector of affinities closest in the two-dimensional Euclidean space (higher-dimensional in a full experiment) to the measured affinity. The colored regions correspond to the regions within which a measured multi-NAAB affinity vector would be assigned to a given amino acid. As an example, a pair of measurements yielding the white star in D would identify the target as glycine. **(E)** The affinities of the glutamine and lysine targeting NAABs are shown for each of the amino acids. Some amino acids that are practically indistinguishable using the Met and Trp NAABs are easily distinguished using the Gln and Lys NAABs. As an example, if the same target amino acid described in D were measured with only the Gln and Lys NAABs, yielding the white star, we would identify the target as proline. However, combining these measurements with those for the white star in D with Met and Trp NAABs, we see that the true identity of the target is serine. Thus, the higher dimensional measurement of the amino acid using many different NAABs allows disambiguation of the amino acid identity.

The limited vertical extent of the evanescent excitation field of the TIRF microscope allows differential sensitivity to fluorescent molecules which are near the microscope slide surface, which allows us to detect NAABs that have bound to peptides on the surface. Existing sets of NAABS (e.g. (*180*)), derived from aminopeptidases or tRNA synthetases with affinities biased towards specific amino acids, have low affinity or specificity (Figure 5-1A), so one cannot deduce the identity of an N-terminal amino acid from the binding of a single NAAB. Instead, we propose to deduce the identity of the N terminal amino acid of a particular peptide by measuring optically the kinetics of a set of NAABs against the peptide. After observing the binding of each NAAB against the peptide, we will carry out a cycle of Edman degradation (*211*, *212*), revealing the next amino acid along the chain as the new N-terminus, and then repeat the process. The process of observing binding kinetics with TIRF microscopy (Figure 5-1B,C) is similar to that used in Points Accumulation for Imaging of Nanoscale Topography (PAINT (*203*)), e.g., DNA PAINT (*204*). This process produces a high-dimensional vector of kinetically-measured affinities at each cycle (Figure 5-1D,E) that can be used to infer the N-terminal amino acid.

This method, while powerful and potentially applicable for current NAABs, ultimately breaks down for probes with off-rates faster than the imaging frame rate, or for which the bound time is so short that only a small number of photons (e.g. less than 100, corresponding to 10% shot noise) is released while the probe is bound. While fast camera frame rates can be used, the system ultimately becomes limited in the achievable fluorescent signal to noise ratio, unless the measurements are averaged over long experiment times. To extend these concepts into the ultra-weak binding regime, therefore, we propose not to measure the precise binding and unbinding kinetics but rather the time-averaged luminosity of each spot, which indicates the fraction of time a probe was bound. We find that this luminosity-based measurement scheme is highly robust and compatible with short run times.

Results

Our results are divided into three sections. We first consider the regimes of binder concentration and illumination intensity within which one would expect the proposed method to operate. We then discuss two possible methods for analyzing single molecule kinetic data. Finally, we perform simulations using the derived parameters and data analysis methods in order to estimate the sensitivity of the proposed sequencing method.

Distinguishability of amino acids based on their NAAB binding profiles

A set of binders (NAABs) is characterized by their affinities for their targets (e.g., the 20 amino acids), which can be expressed in the form of an affinity matrix. The affinity matrix $\boldsymbol{A}$ is defined such that the $\boldsymbol{i,j}$th entry of $\boldsymbol{A}$ is the negative log affinity of the $\boldsymbol{i}$th binder for the $\boldsymbol{j}$th target:

$$\boldsymbol{a_{i,j} = -\log(k_D)} \tag{20}$$

where k_D is the dissociation constant (we define τ_D as the dissociation time).

Throughout this paper, the values of the affinities encoded in the affinity matrix will be referred to as the *reference* values, to distinguish them from the *measured* values obtained in the experiment and from the *true* values, which may depend on environmental conditions but which are not known by the experimenter; the reference values are known and will be used in our computational process of identifying amino acids. As shown in Appendix A (Chapter 11), we estimate that it would be possible to determine the identities of the N terminal amino acids from affinity measurements with 99% accuracy, provided that the affinity measurements occur according to a distribution centered on the reference value with standard deviation no greater than 64% of the mean.

Model Parameters

In order to evaluate the feasibility of the kinetic measurement strategy, we designed a model to simulate the observation of NAAB binding and unbinding from a peptide target, using TIRF microscopy. In evaluating the kinetic measurement strategy, we must make assumptions about the relevant photophysical parameters.

1. The rate R of photons from a single fluorophore captured by the detector, per second. This is a product of numerous parameters specific to the experimental implementation, including the collection efficiency of the optical setup, the illumination intensity, the quantum efficiency of the fluorophore, and the quantum efficiency of the detector. We use realistic values in the range of 10,000 photons per second (*204*, *213–215*).
2. The mean number N_q of photons that a fluorophore can emit before it photobleaches. Realistic numbers on the order of $N_q \sim 10^7$ have been reported for Atto647N (*204*).
3. The pixel size. We will assume that peptides are anchored to the surface sparsely enough so that there is at most one peptide per pixel. We will further assume that each pixel collects light from a cylindrical region $300\ \text{nm}$ in diameter and $100\ \text{nm}$ in depth, corresponding to visible TIRF illumination. It is useful to bear in mind that a free fluorophore occupation number of $n_{\text{free}} \approx 1$ in every cylinder with diameter $300\ \text{nm}$ and height $100\ \text{nm}$ corresponds to a molar density of $235\ \text{nM}$.
4. The background level. Each pixel collects some amount of background light. We draw a distinction between transient emission sources (such as diffusing fluorophores) and constant sources of background photons, such as autofluorescence and excitation of fluorophores in the bulk by first- and higher-order beams. Transient emission sources are modeled, but we decline to model autofluorescence and bulk excitation, because previous studies have shown that the contribution of those sources are small compared to the fluorescence of fluorophores excited by the zeroth order beam (*204*).

5. The free NAAB concentration, n_{free}. The choice of n_{free} is up to the experimenter and may be chosen differently for different NAABs. It will need to be optimized to maximize the dynamic range of the k_D readout experiment.

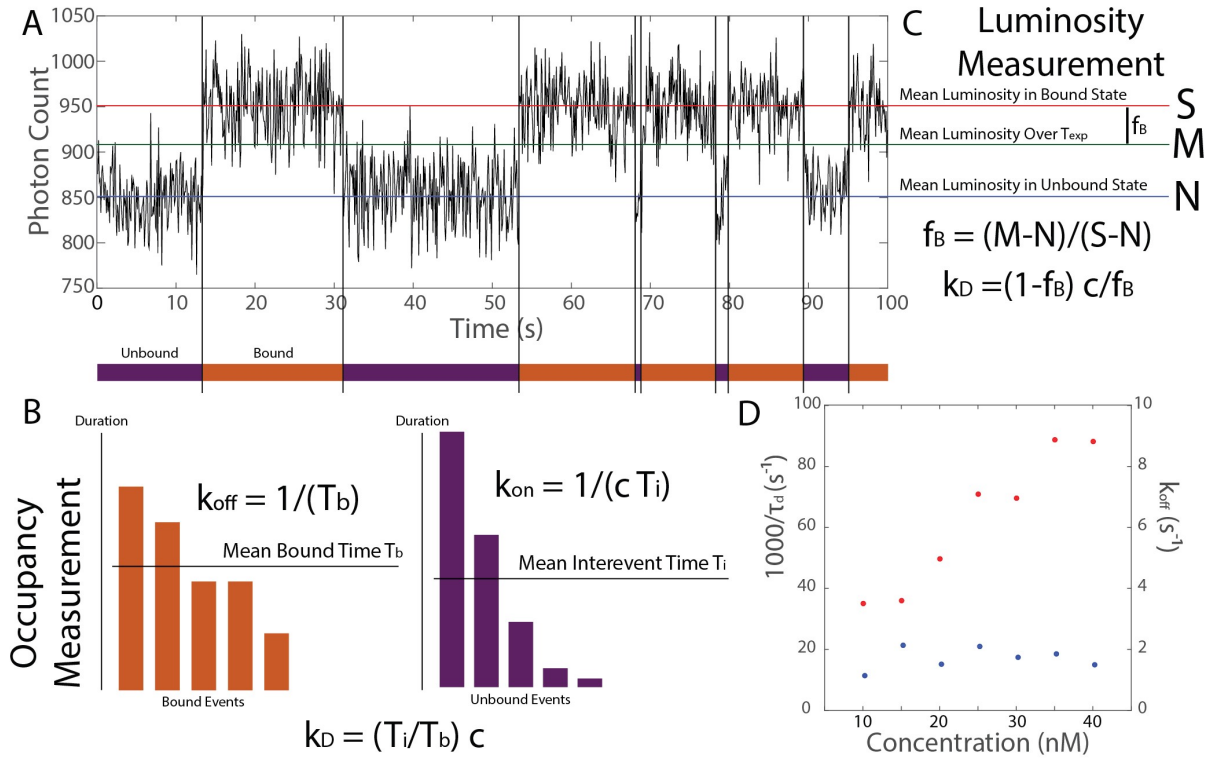

Figure 5-2 Two types of affinity measurements using TIRF microscopy. **(A)** A measurement performed using the proposed scheme yields a fluorescence intensity trace where periods of high intensity correspond to the target being bound and periods of low intensity correspond to the target being free. The affinity of a binder against the target may then be determined in two ways, either via occupancy measurements or via luminosity measurements. **(B)** An occupancy measurement is performed "along the time axis," by calculating $\boldsymbol{k_{on}}$ from the average time between binding events, and $\boldsymbol{k_{off}}$ from the average length of binding events. **(C)** On the other hand, a luminosity measurement is performed "along the brightness axis," by calculating $\boldsymbol{k_D}$ directly from the average luminosity of the target over the whole observation period. **(D)** We validated our simulation by applying occupancy measurements to determine $\boldsymbol{k_{on}}$ and $\boldsymbol{k_{off}}$ from simulated data. The parameters used here were identical to those used in the production of Fig 2a in (*204*). See text for symbol definitions.

Methods of Data Analysis

A single-molecule experiment using TIRF yields a time series such as that shown in Figure 5-2A. We now discuss the two primary options for extracting the kinetics from this data and the experimental conditions that are optimal for each scheme, given the constraints discussed above.

Occupancy Measurements

The first measurement, used commonly in the field of single molecule kinetics (*204*, *216*), relies on detecting changes in the occupancy state of the target. The measurement scheme is depicted schematically in Figure 5-2B. This measurement is performed "along the time axis," in the sense that it relies on temporal information—when probes bind and unbind—and is relatively insensitive to analog luminosity information beyond that needed to make these digital determinations.

In this method, the parameters of interest are the free NAAB concentration n_{free} and the frame rate, $f = 1/\tau_{\mathrm{obs}}$. The upper limit on the dynamic range of this method is set by the frame rate, i.e.,

$$\boldsymbol{\tau_{\mathrm{obs}} \ll 1/k_{\mathrm{off}}} \tag{21}$$

On the other hand, the lower bound on the dynamic range is set by the duration of the experiment T_{exp}, via the requirement that

$$\boldsymbol{T_{\mathrm{exp}} \gg 1/k_{\mathrm{off}}} \tag{22}$$

so that unbinding events can also be observed, and also that

$$\boldsymbol{T_{\mathrm{exp}} \gg 1/(k_{\mathrm{on}} c)} \tag{23}$$

so binding events can be observed. For a value of k_{on} between $10^5 \mathrm{M}^{-1}\mathrm{s}^{-1}$ and $10^6 \mathrm{M}^{-1}\mathrm{s}^{-1}$ (e.g. (*204*, *217*)) and a concentration on the order of $100\ \mathrm{nM}$, this requirement implies that an experiment time of at least 100 seconds is necessary in order to see several binding events with high probability. In addition, we will choose $f = 100\ \mathrm{Hz}$ for this measurement modality, which then implies a dynamic range of roughly 5 orders of magnitude in k_{off}. The values of k_{off} that can be discerned are also constrained by photobleaching and by the background. Specifically, if R is the rate of photon detection, N_q is the mean number of detected photons emitted by the fluorophore before bleaching, and B is the mean rate of background photon detection (due to camera noise, autofluorescence, etc.), then we also have

$$\boldsymbol{\frac{R}{B} \gg k_{\mathrm{off}} \gg \frac{R}{N_q}} \tag{24}$$

The value of k_D is determined in this modality as follows. If the binding and unbinding events may be identified, then one may determine the average binding time T_b and the average time between binding events T_i, which we will refer to as the inter-event time. If photobleaching may be neglected, then we have

$$\boldsymbol{k_{\mathrm{off}} = \frac{1}{T_b}} \tag{25}$$

and

$$k_{\mathrm{off}} = \frac{1}{T_i c} \tag{26}$$

where c is the free binder concentration. Thus,

$$k_D = \frac{T_i}{T_b}\mathrm{c} \tag{27}$$

Additionally, if the on-rate k_{on} is known, then it is possible to determine k_{off} even in the presence of photobleaching. (See Appendix C, Chapter 11, for details.)

Luminosity Measurements

An alternative to the occupancy-time measurements described above involves deducing k_D directly from the fraction f_B of time that the target is bound by a probe. This quantity may in turn be deduced from the average luminosity of the spot containing the free binder over the period of observation, as depicted in Figure 5-2C. Whereas occupancy measurements are performed "along the time axis," neglecting luminosity information, luminosity measurements are performed "along the luminosity axis," neglecting temporal information about the series of binding and unbinding events. Because it does not attempt to track individual binding and unbinding events, this method is particularly suited to measurements of weak binders performed at high background concentrations, where binding and unbinding events may occur faster than the camera frame rate. Moreover, this method relies on each NAAB of a given type having approximately the same brightness, which could be achieved using a high-efficiency method for monovalently labeling the NAAB N- or C-terminus (*218*, *219*).

If the target is bound a fraction f_B of the time, then the dissociation constant is given by

$$k_D = \frac{1 - f_B}{f_B} c \tag{28}$$

where c is the background binder concentration. We denote by S the average brightness of the spot when a fluorescent binder is attached to the target, and by N the average brightness of the spot when the target is free. Neglecting photobleaching, the average brightness of the spot over the whole experiment is given by

$$M = f_B S + (1 - f_B) N \tag{29}$$

If S and N are known, then f_B may thus be deduced directly from the measured photon rate M averaged over the entire experiment, via

$$f_B = \frac{M - N}{S - N} \tag{30}$$

S and N can be measured directly for example by anchoring NAABs sparsely to a surface and measuring the brightness of the resulting puncta (to deduce S), or puncta-free regions (to measure N).

One significant advantage of this method is that the observation period τ_{obs} can be chosen to be arbitrarily long by averaging the photon counts of many successive frames (i.e., we have $\tau_{\text{obs}} = T_{\text{exp}}$. In practice, we will use $\tau_{\text{obs}} = 100\ \text{s}$. With this value, we can use a relatively high concentration of $2\ \mu\text{M}$ (corresponding to $n_{\text{free}} \gg 1$) and a relatively low emission rate of $R = 10^3\ \text{s}^{-1}$. The choice of a high NAAB concentration and low illumination intensity increases the dynamic range of the measurement scheme, by increasing the sensitivity both to small values of k_D, where photobleaching might be an issue, and to high values of k_D, where observation of binding events may be an issue. However, unlike in the case of occupancy measurements, there is no way to account for photobleaching, if it occurs. Nonetheless, we do not expect photobleaching to have a significant impact on our results, since most of the NAABs have fairly high off-rates (*180*, *201*).

Simulations

In order to determine whether the TIRF measurement scheme described above can be used to identify single amino acids on the N-termini of surface-anchored peptides, we simulated N terminal amino acid identification experiments.

We first used a specific NAAB affinity matrix given in (*180*). Importantly, random affinity matrices (see Appendix E, Chapter 11) generated by permuting the values of the NAAB affinity matrix perform similarly well in residue-calling simulations. To generate the random affinity matrices with statistics matching the statistics of the NAAB affinity matrix, each matrix element was chosen by randomly sampling values from the NAAB affinity matrix of (*180*), without replacement. The simulations described here can therefore be assumed to apply to general ensembles of N-terminal binders with affinity value statistics similar to those displayed by these existing NAABs.

In the simulations, there is assumed to be one free target in the volume analyzed, which is a cylinder of diameter $300\ \text{nm}$ and height $100\ \text{nm}$ as discussed above. Thus, we assume that peptides are arrayed sparsely enough on the surface that there is at most one peptide per diffraction-limited spot. The simulation considers each frame of the camera in succession, and models the number of photons registered at the camera. At the start of the simulation, or as soon as the target becomes free, a time T_{free} is drawn from an exponential distribution with mean $1/k_{\text{on}}c$, where c is the concentration of binders. Once a time equal to T_{free} has passed, the binder is considered occupied, and a time T_{bound} is drawn from an exponential distribution with mean $1/k_{\text{off}}$. In addition, upon binding, a time $T_{\text{photobleach}}$ is drawn from an exponential distribution with mean N_q/R, where N_q is the number of photons seen by the detector on average before the fluorophore bleaches and R is the number of photons seen at the detector by a single fluorophore per second. For a dye like Atto 647N, we use $N_q = 1.2 \times 10^7$ (*204*). If the time $T_{\text{photobleach}}$ is less than the time Tbound, the fluorophore ceases to emit photons after time $T_{\text{photobleach}}$. Within a given frame, the simulation tracks binding, unbinding, and photobleaching events, and computes

the number of signal photons detected by the camera by drawing from a Poisson distribution with mean RT_{on}, where R is the single fluorophore photon rate and T_{on} is the amount of time during the frame in which an unbleached fluorophore was attached to the target.

In addition to background photons, the dominant contribution to noise in the simulation is expected to come from fluorophores attached to free binders that enter and leave the observation field (*216*). At the end of each frame, the simulation draws the number of free binders that enter the observation field during the frame from a Poisson distribution with mean n_{free}/f, where f is the frame rate and n_{free} is the free binder occupation number of the frame. For each binder that enters the observation field, we draw a dwell time t from an exponential distribution with mean τ_{dwell} as calculated in Eq. (46) from diffusion theory (see Appendix B, Chapter 11), and a total photon contribution from a Poisson distribution with mean Rt. Finally, we calculate the detector shot noise from a Gaussian distribution with mean p and standard deviation equal to $0.1p$.

Validation of the Simulation Pipeline

To validate the simulations, we reproduced the DNA PAINT kinetics data collected by (*204*) using the parameters reported in that paper. There, values of $k_{\mathrm{on}} \sim 2.2 \times 10^6\ \mathrm{M^{-1}s^{-1}}$ and $k_{\mathrm{off}} \sim 1.8\ \mathrm{s^{-1}}$ were reported. Although the photon rate was not directly reported in that paper, other papers using similar laser intensities and fluorophores reported photon rates on the order of $R \sim 10000\ \mathrm{s^{-1}}$ (*213–215*), so we used this value. From our simulated data, we were able to reproduce the measured off- and on-rates, as shown in Figure 5-2D.

Measurements of k_D

We next compared the ability of occupancy and luminosity measurements to determine the dissociation constant k_D of binders for the target.

Occupancy Measurements

We performed 100 simulations of occupancy measurements for each of five different values of k_{on} between $10^4\ \mathrm{M^{-1}s^{-1}}$ and $10^6\ \mathrm{M^{-1}s^{-1}}$, which is consistent with standard values observed for antibodies (*217*), and for each of five different values of k_D between $100\ \mu\mathrm{M}$ and $10\ \mathrm{nM}$. We assumed a frame-rate of $100\ \mathrm{Hz}$, detector read noise of $1\ \mathrm{e^-}$, and a single-fluorophore detection rate of $10^5\ \mathrm{s^{-1}}$. The NAAB concentration was $300\ \mathrm{nM}$, and the observation time was $T_{\mathrm{exp}} = 100\ \mathrm{s}$.

In order to analyze the data, we ran a control simulation in which k_{on} was set to 0, so that no NAABs bound to the target. In practice, this calibration could be performed by observing a spot that does not have a target. From this, we calculated the mean and standard deviation of the noise on a per-frame basis. We then identified binding and unbinding events as follows. First, we identified all frames in which the photon count was more than 2 standard deviations above the noise mean. These frames will be referred to as "on" frames, whereas all other frames will be referred to as "off" frames. If three such "on" frames occurred in a row, the event was identified as

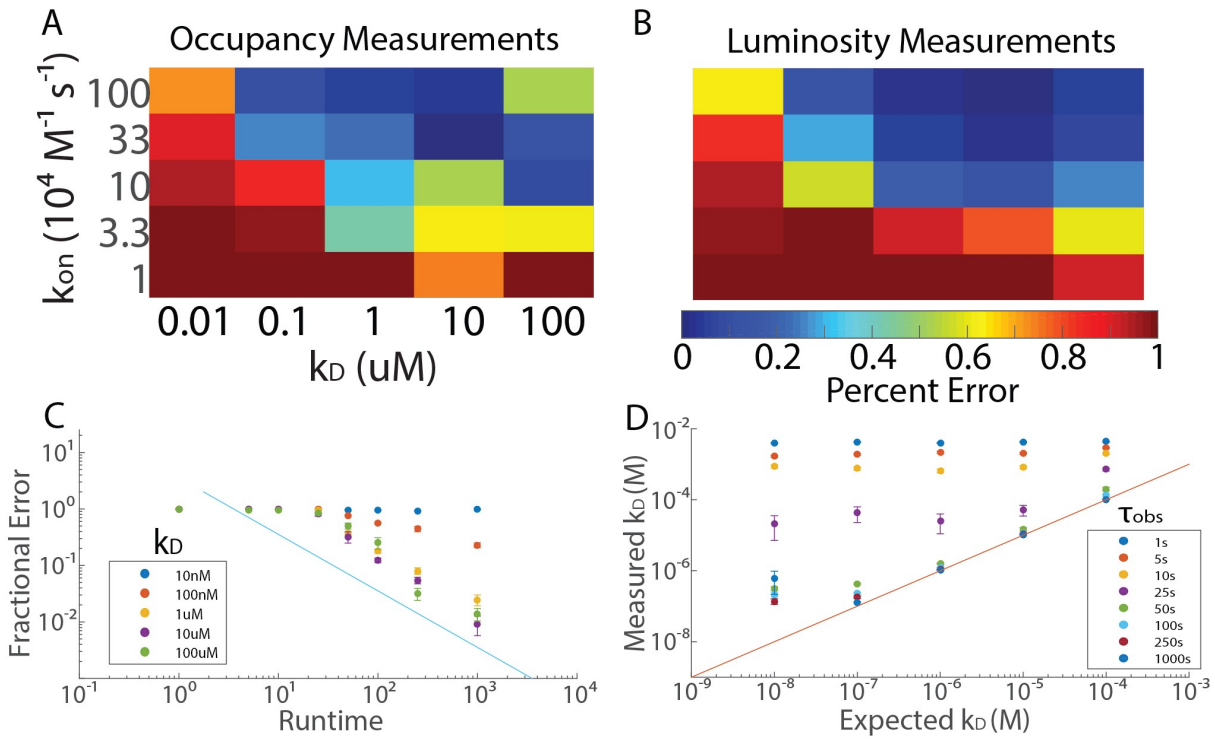

Figure 5-3: Two types of affinity measurements using TIRF microscopy. **(A)** The accuracies of occupation measurements of k_D are shown as a function of k_D and k_{on} for the simulation described in the text, with $T_{exp} = 100\ s$. These measurements achieve high accuracy for $k_{on} \geq 10^4\ M^{-1}s^{-1}$ and $k_{off} \ll 100\ s^{-1}$. For values of k_{off} on the order of $100\ s^{-1}$ (upper right-hand corner), the accuracy deteriorates significantly. **(B)** The accuracies of luminosity measurements of k_D are shown as a function of k_D and k_{on}. These measurements achieve high accuracy for $k_{on} \geq 10^5\ M^{-1}s^{-1}$ and $k_D \geq 100\ nM$. The heat map shown gives the fractional errors as a function of k_D and k_{on} for the simulation described in the text, with $T_{exp} = 100\ s$. In contrast to occupation measurements, the accuracy of luminosity measurements does not deteriorate for very high values of k_{off}. **(C)** For luminosity measurements only, the mean fractional error in the measured value of k_D is plotted as a function of the observation time for five different values of k_D. The line $y = 1/x$ is plotted as a guide to the eye. For $k_D = 10\ nM$ and $k_D = 100\ nM$, the effects of photobleaching are evident at longer runtimes. **(D)** Also, for luminosity measurements only, the measured value of k_D is plotted as a function of the actual value of k_D for 8 different values of the runtime. The performance of the algorithm improves dramatically for $T_{obs} > 25\ s$. The line $y = x$ is plotted as a guide to the eye. Error bars in C, D denote standard error over 100 trials.

a binding event. The binding event was considered to continue until at least two "off"-frames in a row were observed. Once all the binding and unbinding events were identified, the average inter-event time and the average binding time were calculated, and from these the kinetics were deduced (Fig 2A).

The accuracy of the k_D measurements was found to improve with increasing k_{on}, and to improve with increasing k_D for values of k_{off} below $10\ s^{-1}$ (Figure 5-3A). For values of k_{off} significantly above $10\ s^{-1}$, it was no longer possible to distinguish individual binding and unbinding events

from noise (Figure 5-3A, upper right-hand corner). Moreover, for values of k_{on} below $10^5\ \mathrm{M^{-1}s^{-1}}$, the condition $T_{\mathrm{exp}} \gg 1/k_{\mathrm{on}}c$ was no longer satisfied. Finally, for very small values of k_D, photobleaching limited the accuracy of the analysis. For $k_{\mathrm{on}} > 10^5\ \mathrm{M^{-1}s^{-1}}$ and $k_{\mathrm{off}} \sim 10\ \mathrm{s^{-1}}$, it was possible to obtain the correct value of k_D to within approximately $5 - 10\%$. However, the accuracy deteriorated sharply for combinations of k_{on} and k_{off} deviating from these ideal conditions.

Luminosity Measurements

We then simulated luminosity measurements of k_D using comparable parameters. Because these measurements depend only on the average luminosity over the entire experiment, the entire experiment was lumped into a single camera frame. In practice, however, the same results can be obtained by averaging over the photon counts of multiple frames. The photon detection rate was set to $R = 1000\ \mathrm{s^{-1}}$, and the free binder concentration was set to $2\ \mu\mathrm{M}$. The photon rate of the off-state was determined first by running the simulation with the value of k_{on} set to 0. The photon rate in the on-state was then determined by running the simulation with the value of k_{on} set to $10^{10}\ \mathrm{M^{-1}s^{-1}}$, and the value of kD set to $10^{-20}\ \mathrm{M}$. Because the exposure time used in this experiment is very long compared to the dwell time of free binders in the observation field, it was assumed that all free binders that enter the observation field emit a number of photons equal to $R\,\tau_{\mathrm{dwell}}$ (i.e., the noise was taken to be approximately Poissonian), which substantially reduces the computational complexity of the algorithm. Once the average luminosity over the experiment was determined, the value of f_B was deduced.

For observation times shorter than $50\ \mathrm{s}$, the analysis sometimes returns values of f_B arbitrarily close to or greater than 1 or arbitrarily close to or less than 0. This can happen as a consequence of statistical error in the luminosity measurements, even in the absence of systematic error. For this reason, in order to avoid negative or outlandishly large values of k_D from compromising the analysis, we chose the maximum value of f_B to be equal to the value expected when $k_D = 1\ \mathrm{nM}$, and we chose the minimum value of f_B to be equal to the value obtained when $k_D = 10\ \mathrm{mM}$. Any values of f_B outside of this range were adjusted to the maximum or minimum value, appropriately.

In order to enable comparison to the occupancy measurements, the simulation was run 100 times for each of five values of k_{on} between $10^4\ \mathrm{M^{-1}s^{-1}}$ and $10^6\ \mathrm{M^{-1}s^{-1}}$ and for each of five values of k_D between $100\ \mu\mathrm{M}$ and $10\ \mathrm{nM}$. The accuracy was found to be comparable to that obtained in the occupancy experiments (Figure 5-3A), except that the accuracy did not deteriorate for very high values of k_{off} (Figure 5-3B, upper right-hand corner). For values of k_{on} on the order of (or greater than) $10^5\ \mathrm{M^{-1}s^{-1}}$ and values of k_D greater than $1\ \mu\mathrm{M}$, k_D could easily be determined to within the accuracy condition required by Eq (37).

To ascertain the effect of τ_{obs} on the accuracy, the simulation was run 100 times for each of the same 25 combinations of k_{on} and k_{off}, with 8 different values of τ_{obs} between $1\ \text{s}$ and $1000\ \text{s}$ and a free binder population of $2\ \mu\text{M}$ (Figure 5-3C). As expected, the accuracy was found to undergo a sharp transition when τ_{obs} was on the order of $25\ \text{s}$, corresponding to $1/k_{\text{on}}c \ll \tau_{\text{obs}}$. For values of $\tau_{\text{obs}} > 25\ \text{s}$ and values of k_D greater than $1\ \mu\text{M}$, the error in the measurement of k_D decreased like $1/\tau_{\text{obs}}$ (Figure 5-3C). For observation times greater than $25\ \text{s}$, the value of k_D could be calculated with standard deviation less than 64% of the mean for values of k_D on the order of or greater than $1\ \mu\text{M}$, although photobleaching leads to saturation and significant losses of accuracy for smaller values of k_D (Figure 5-3D).

Separately, to ascertain the effect of the free binder concentration on the accuracy, the simulation was run 1000 times on each of the same 25 combinations of k_{on} and k_D, with $\tau_{\text{obs}} = 50\ \text{s}$ at seven different values of the concentration between $10\ \text{nM}$ and $5\ \mu\text{M}$. For values of k_{on} such that $\tau_{\text{obs}} \gg 1/(k_{\text{on}}c)$, the effect of increasing k_{on} was found to be similar to the effect of increasing τ_obs (data not shown).

Identifying Amino Acids

Because standard deviations in k_D below 64% of the mean could consistently be achieved in the luminosity measurements across a broad range of values of k_{on} and k_D, it is reasonable to expect that luminosity measurements of NAAB binding kinetics with the affinity matrix in Figure 5-1A could allow for the identification of amino acids at the single molecule level. We thus simulated an experiment, using the luminosity measurement paradigm, in which a peptide with an unknown amino acid is attached to a surface, and is observed successively in multiple baths, each containing a single kind of fluorescent NAAB.

Simulation of systematic errors.

Two kinds of systematic error may confound identification of amino acids. The first kind of error, which we refer to as kinetic error, refers to the case in which the actual dissociation constant for a particular NAAB-amino acid pair is different from the expected value. This may arise due to issues such as the secondary structure or the identities of non-terminal amino acids. To simulate this, for each NAAB, the effective dissociation constant $\tilde{k}_D$ for the NAAB-amino acid pair was drawn from a normal distribution centered on the reference value k_D, with standard deviation equal to $\sigma_K k_D$, where σ_K parametrizes the effect of non-terminal amino acids and other environmental factors on the dissociation constant.

In addition, luminosity measurements are also sensitive to error in the calibration of the measurement apparatus, for example if the brightness of the bright and dark states is not known exactly. We refer to this kind of error as calibration error. The bright and dark states S and N could likely be calibrated by doping in labeled reference peptides to the sample to be sequenced. Still, there may be some error in the measurements of S and N. To simulate this kind of error, the

true calibration levels S and N were first determined as the luminosity of the bound and unbound states. The measured calibration levels $\tilde{S}$ and $\tilde{N}$ were then determined by drawing from a normal distribution with mean equal to S and N and with standard deviation equal to $\sigma_C S$ and $\sigma_C N$, respectively. The values of σ_K and σ_C will be given below in percentages. For a discussion of computational strategies for coping with calibration error, see Appendix D, Chapter 11.

Amino acid identification.
In this simulation, amino acids were randomly chosen from a uniform distribution. Binders were added to the solution at a concentration of $1\ \mu\mathrm{M}$ and the photon detection rate was set to $1000\ \mathrm{s}^{-1}$. For each NAAB, effective values of the dissociation constant $\tilde{k}_D$, the on-rate $\tilde{k}_{\mathrm{on}}$, the effective brightness $\tilde{R}$, and the calibration levels $\tilde{S}$ and $\tilde{N}$ were determined for the NAAB-amino acid pair. The spot containing the NAAB was then observed over a period of time τ_{obs}, which ranged from 50 to 500 seconds, and the total number of photons observed was stored. This process was repeated for each NAAB, generating a vector $\vec{M}$ of observed photon counts.

Analysis was performed by comparing the measured photon counts to the photon counts that would have been expected for each amino acid, as described above. For each NAAB-amino acid pair, the expected photon count was calculated from the NAAB concentration c, the reference value of k_D and the measured calibration level $\tilde{S}$ and $\tilde{N}$, via

$$\vec{E} = \frac{c}{c + k_D}\tilde{S} + \left(1 - \frac{c}{c + k_D}\right)\tilde{N} \tag{31}$$

The resulting expected photon counts were then assembled into a matrix W, such that the (i,j)th element of W is the photon count that one would have expected on the measurement of the ith NAAB if the target were the jth amino acid, given the calibration levels $\tilde{S}$ and $\tilde{N}$. Finally, the amino acid identity I_{aa} was determined by minimizing the norm between the vector of observed photon counts $\vec{M}$ and the columns of W, i.e.,

$$I_{\mathrm{aa}} = \mathrm{argmin}_k \left\| \vec{M} - \vec{w}_k \right\| \tag{32}$$

where $\vec{w}_k$ is the kth column of W. In Figure 5-4A-C, the accuracy with which amino acids can be identified is shown as a function of the observation time and the systematic error, for a $1\ \mu\mathrm{M}$ free binder concentration. In the absence of systematic error, amino acids could be identified with greater than 99% accuracy after a $50\ \mathrm{s}$ observation. Moreover, the experiment also showed robustness against kinetic error up to the 25% level, with progressive deterioration in the measurement accuracy observed for values of σ_K above 25%. Calibration error was found to have the most substantial effect on the accuracy, with calibration errors on the order of 10% reducing the achievable accuracy below 90% even for an observation time of $250\ \mathrm{s}$. The effects of calibration error on the accuracy could be substantially reduced by reducing the concentration of free binders (Figure 5-4D), which has the effect of increasing the gap between the S and N. However, in order to preserve the requirement that $T_{\mathrm{exp}} \gg 1/(k_{\mathrm{on}} c)$, it is necessary to increase

the experiment length by a similar factor. (For this reason, a free NAAB concentration of **1 μM** was used, rather than **2 μM** as used above.) Moreover, this improvement comes at the cost of

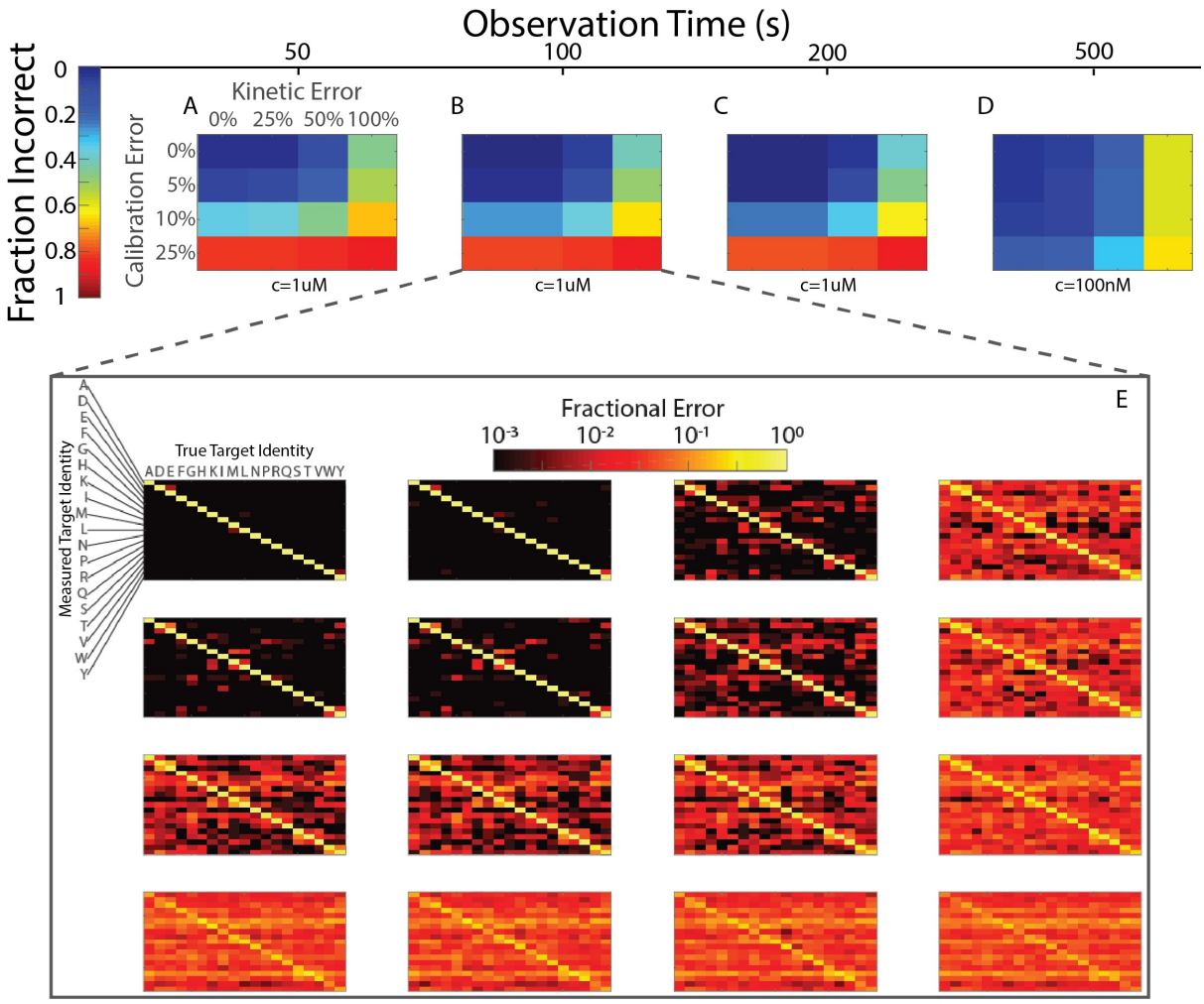

Figure 5-4: Identification of amino acids is robust against systematic error. The fraction of amino acids incorrectly identified is plotted as a function of $\boldsymbol{T_{obs}}$ for four different values of the systematic calibration error $\boldsymbol{\sigma_C}$ and four different values of the systematic kinetic error $\boldsymbol{\sigma_K}$ (as described in the text). (A) In the absence of systematic error, measurements with $\boldsymbol{T_{obs} = 50\ s}$ result in correct amino acid identification more than 98% of the time. For 25% error in $\boldsymbol{k_D}$, the accuracy drops to 97.5%, and if 5% calibration error is added, it drops further to 92%. More than 5% systematic error in the calibration leads to very significant numbers of mistakes in amino acid identification. (B) With $\boldsymbol{T_{obs} = 100\ s}$, an accuracy of 97.5% was obtained for 25% error in $\boldsymbol{k_D}$ and 5% error in the calibration. Axes for B, C, and D are the same as in A. (C) Increasing $\boldsymbol{T_{obs}}$ beyond 100 s at the same binder concentration leads to diminishing improvements in the accuracy. (D) The sensitivity to calibration error could be substantially reduced by decreasing the concentration of free binders to **100 nM**. However, this decreased concentration necessitates a longer runtime. E For $\boldsymbol{T_{obs} = 100\ s}$, plots are shown for each value of $\boldsymbol{\sigma_C}$ and $\boldsymbol{\sigma_K}$, depicting the probability that a given target amino acid (on the horizontal axis) was assigned a particular identity (on the vertical axis). Off-diagonal elements correspond to errors.

increased sensitivity to systematic error in k_D. If the calibration error can be kept below 5%, and if the systematic error in the kinetics can be kept below 25%, then our simulations indicate that it would be possible to identify amino acids with greater than 97.5% accuracy over an observation window of 100 s.

Application to Randomized Affinity Matrices

In order to determine whether the protein sequencing method proposed here is limited to the specific affinity matrix given in (*180*), we generated affinity matrices with comparable binding statistics by randomly shuffling the k_D values in the NAAB affinity matrix. For 100 such random affinity matrices, we then performed identical simulations as in Figure 5-4E, assuming 5% calibration error and 25% kinetic error. To calculate the overall error rate for a given matrix, we summed the frequencies of incorrect residue calls (the off-diagonal elements of the matrices in Figure 5-4E). The overall error rate for the NAAB affinity matrix, calculated in this way, is 0.0124, and the distribution of error rates across the random matrices is shown in Figure 5-5. Only one randomly generated affinity matrix had an error rate lower than the NAAB error rate. Nonetheless, it is clear that most affinity matrices with affinity statistics similar to the NAABs (*180*) would yield errors in the range of 1%-4%, and thus the sequencing method described here is generalizable to a range of similar N-terminal amino acid binders.

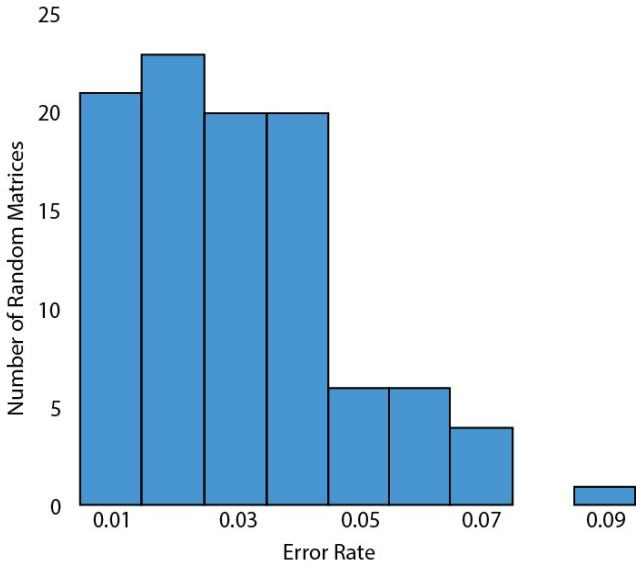

Figure 5-5: Overall error rates for 100 random affinity matrices. A histogram of the overall error rate, calculated as the sum of incorrect residue calls divided by the total number of residue calls over 10000 trials, is plotted for 100 random affinity matrices.

Discussion

The calculations and simulations discussed above indicate that if the measurement apparatus can be calibrated with an accuracy of 5%, and if the reference values of k_D can be kept within 25% of the true values, it is theoretically possible to determine the identity of an N-terminal amino acid with greater than 97.5% accuracy by measuring the kinetics of the NAABs against the target amino acid. Crucially, k_D can be inferred just from the time-averaged local concentration of NAABs within the observation field, and thus the measurement can be performed at relatively high background binder concentrations, because it does not rely on being able to distinguish individual binding and unbinding events.

Primary Uncertainties

Three primary uncertainties exist regarding the validity of the simulations performed here. Firstly, our simulation did not incorporate the effects of non-specific binding of NAABs to the surface. However, non-specific binding will simply increase the level of background fluorescence, and numerous recent single-molecule imaging studies have demonstrated surface passivation techniques that minimize nonspecific background (*204*, *220*).

Secondly, the sequencing will take place in non-denaturing buffers, as is necessary for the NAABs. We anticipate that small, surface-anchored peptides derived by cleaving proteins will be accessible for NAAB binding, as has been shown previously, for example in the case of biolayer interferometry (*180*). However, some peptides may not be sequenceable in this method due to secondary structures or other idiosyncrasies. In addition, some uncertainty exists surrounding the value of N_q for the organic dyes of interest to us, with values between 10^5 and 10^7 being reported (*204*, *221*). However, we expect our method to be relatively robust to photobleaching due to the relatively low affinity and high off-rates of most of the NAABs. Moreover, it is possible that more photostable indicators such as quantum dots could be used in place of organic dyes. Note that with any labeling scheme, there will be some concentration of "dark NAABs" that are not labeled. We do not expect this to be a major issue for the detection scheme provided the total NAAB concentration is less than the dissociation constant (i.e., as long as the target is free most of the time). However, if this is an issue, several other strategies are available to ensure high-efficiency labeling of NAABs, for example expressing the NAABs as fusions to a fluorescent protein, or to a peptide tag or protein (e.g. the SNAP tag) that can be used to link the NAABs to small molecule fluorophores with high efficiency. Moreover, a high concentration of dark NAABs can always be compensated for by reducing the total NAAB concentration and increasing the measurement duration. Nonetheless, the concentrations reported for the simulations above should be regarded as the concentrations of successfully labeled "bright NAABs."

Calibration Error

The luminosity measurement scheme is particularly sensitive to calibration error. This is because the brightness of puncta in the luminosity measurement scheme is used to infer the fraction f_B of the time that the NAAB is bound, and when that fraction is close to 1 or close to 0, then small systematic errors in estimating f_B can contribute to large errors in estimating k_D. A more robust scheme might be to use the relative luminosity of different NAABs, which would then account for effects due to the structure of the peptide (e.g. aromatic residues such as tryptophan might contribute to quenching) and due to local variations in surface passivation. One straightforward way to do this would be to normalize the luminosities to the luminosity of a particular high-affinity NAAB.

Parallelization

We anticipate that the approaches discussed here could be parallelized in a way reminiscent of next-generation nucleic acid sequencing technologies, allowing for massively parallel protein sequencing with single-molecule resolution. In the ideal case, if a 64 megapixel camera were used with one target per pixel, we would have the ability to observe the binding kinetics of NAABs against approximately 10^7 protein fragments simultaneously. With an observation time of 100 seconds per amino acid-NAAB pair, this corresponds to approximately 35 minutes of observation time per amino acid, or 5 days to identify a protein fragment of 200 amino acids in length. As the method is scaled up, the imaging time will come to dominate over the time needed for the fluidic and chemical steps. For example, one flow cell could be imaged while Edman degradation proceeds on a different flowcell. More generally, because imaging requires photon collection through a magnification system, and data transfer to a computer, it is likely to be largely serial, or parallel only up to the number of parallel cameras, whereas fluidic wash and reaction steps can occur in parallel over an entire large surface. Thus in the limit of acquiring data from large flow cells the chemical cycle time of the Edman degradation steps is negligible compared to the imaging time. On average, therefore, the sequencing method as a whole would have a throughput of approximately 20 proteins per second per 64-megapixel camera and its associated imaging setup.

However, the throughput of the device could be improved dramatically if the readout mechanism were electrical, rather than optical. CMOS-compatible field-effect transistors have been developed as sensors for biological molecules (*222*–*225*). Moreover, electrical sequencing of DNA has been accomplished using ion semiconductor sequencing (*226*). Most recently, CMOS-compatible carbon nanotube FETs have been shown to detect DNA hybridization kinetics with better than $10\,\mathrm{ms}$ time resolution (*227*, *228*). Similar CMOS-compatible devices have been adapted to the detection of protein concentrations via immunodetection (*229*). These systems have the added benefit that they sense from a much smaller volume than TIRF does (sometimes as small as ~ 10 cubic nanometers (*228*)), substantially reducing the impact of noise on the measurement. A single 5-inch silicon wafer covered in transistor sensors at a density of 16 transistors per square micron would be capable of sequencing 10^{12} proteins simultaneously, corresponding to an average throughput of 2,000,000 proteins per second on a single wafer, or one mammalian cell every 7 minutes. Such an approach could make use of dedicated integration circuitry to compute the average NAAB occupancy at the hardware level, greatly simplifying data acquisition and processing. Moreover, if the devices were made CMOS-compatible, they could be produced in bulk, greatly improving scalability. If the intrinsic contrast provided by the NAABs is insufficient for measurements with FETs, the NAABs can be further engineered to have greater electrical contrast, for example by conjugating them on the C-terminus to an electrically salient protein such as ferritin. A combination of electrical and optical readouts may also be desirable. Recently, CMOS-compatible single-photon avalanche diode imaging systems have been developed that are capable of detecting the presence of fluorophores on a surface without optics (*230*).

Finally, although the use of TIRF microscopy in the case studied here restricts the proposed approach to operate close to a reflecting surface, the use of thin sections or alternative microscopies could potentially allow such protein sequencing methods to operate in-situ inside intact cells or tissues.

Conclusion

We have shown that single molecule protein sequencing is possible using low-affinity, low-specificity binding reagents and single molecule fluorescent detection. Achieving a high-quality single molecule surface chemistry and TIRF measurement setup will be a challenge, but if this can be achieved, our results show that a wide range of binding reagent families should be adaptable to single molecule protein sequencing.

Chapter 6
Tickertape

As observed most prominently by Adam Marblestone (*30*), there are numerous physical limitations that need to be overcome to enable a brain activity recording approach to scale to the whole brain. In 2011, Konrad Kording proposed that neural activity recording could be scaled to the whole brain level by engineering neurons to *record their own activity*, for example into some kind of a molecular recording device that he termed a "tickertape" (*231*). In the original scheme, it was thought that such a tickertape would operate using an error-prone DNA polymerase with an error rate that was modulated by calcium, allowing the history of calcium in a cell to be inferred by DNA sequencing. However, experiments by Bradley Zamft and Adam Marblestone (*232*), as well as Keith Tyo's lab (e.g. (*233*)), have indicated that it is challenging to endow polymerases with exquisite sensitivity to calcium ions. Moreover, all systems for DNA-based recording of cellular activity operate on timescales of days, which are too slow for recording any activity that would be relevant to neuroscience (*232*, *234–243*).

This project began with an idea by Fei Chen and Asmamaw Wassie, inspired by a paper showing the possibility of polyuridinylating RNAs to record protein-protein interactions (*244*). Fei and Oz had the idea to record neural activity into the poly(A) tail of mRNAs using poly(U) and poly(A) polymerases, which constitutively add Us and As to the 3' ends of RNAs. The idea was that the poly(U) activity would be made calcium-dependent by modulating binding of the poly(U) protein to the RNA, while the poly(A) activity would be constitutive. I began work on the project, but we found it challenging to validate that the proteins were acting as desired. Moreover, we never detected RNAs with more than ~15 Us on their 3' side, despite the fact that poly(U) proteins were known to add hundreds of Us to the poly(A) tail *in vitro* (*245–251*). We hypothesized that this was because polyuridinylation is a marker for RNA degradation in mammalian cells (*252*).

In September 2017, Fei and I sat down to try to figure out how to accelerate our path to a tickertape paper. Fei was convinced that the way forward was to use the ADAR base editing protein, which had worked well for Jonathan Gootenberg and Omar Abudayyeh (*253*), but I insisted it would not be enough simply to make an integrator, and we didn't know how to make a tickertape from the single base editor on its own. Together, we realized that it would be possible to create a tickertape by using the fact that there are many RNAs in each cell, and (counterintuitively) having the integrator integrate the constant function, i.e. time. If each RNA integrated the time since its own creation, we would be able to infer the timing of promoter activity from the ensemble of RNA integrators. Through work with Linlin Chen, this concept became the tickertape described below, which I expect will appear in print before the end of 2019.

Summary:

Time varying transcriptional programs and cellular dynamics are often transient, and are difficult to monitor in their native context. Synthetic cellular memory devices which record biological signals in nucleic acid substrates would allow longitudinal study of cellular dynamics to be derived from a single endpoint measurement. Several recently published methods have succeeded in recording cell-state information into the sequence of DNA in living cells, but all such methods operate on timescales greater than the generation time of the cell (days to years, for mammalian cells), and are thus insufficient for recording transcriptional responses to perturbations, which typically place over hours. Here, we describe a molecular recorder (an "RNA Tickertape") that encodes the absolute timings of transcriptional events in mammalian cells into the sequences of reporter RNA molecules. Whereas DNA recorders rely on relatively slow DNA repair mechanisms, our reporter relies on the fast RNA editing reaction of Adenosine Deaminase Acting on RNA (ADAR), and thus enables the timings and amplitudes of transcriptional events in single cells to be inferred from endpoint measurements with single-hour accuracy. We demonstrate the ability to decode arbitrary temporal patterns of transcriptional activity reaching up to 12 hours prior to cell lysis. Finally, by coupling the tickertape to immediate early genes in neurons, we achieve the first sequencing-based readout of neural activity, which may ultimately enable the study of deep and otherwise inaccessible populations of neurons in the brain. RNA tickertapes thus open up possibilities for the high-throughput, multiplexed interrogation of the temporal dimension of cellular behavior.

Introduction

The introduction of green fluorescent protein to the biological toolkit was transformative for many areas of biology. In particular, the application of fluorescent proteins as reporter genes has allowed longitudinal temporal dynamics in individual cells to be inferred by imaging, with applications ranging from neural activity to gene expression (*254*, *255*). Today, several new sequencing technologies have attempted to add a temporal dimension to RNA-seq, for example by metabolically labeling newly synthesized RNAs (*256–258*), or by inferring the instantaneous change in cell state from the abundance of unspliced transcripts (*259*). However, all existing tools for inferring temporal information from RNA sequencing only provide an instantaneous snapshot of the temporal activity of a cell: metabolic labeling only identifies RNA made within one specific

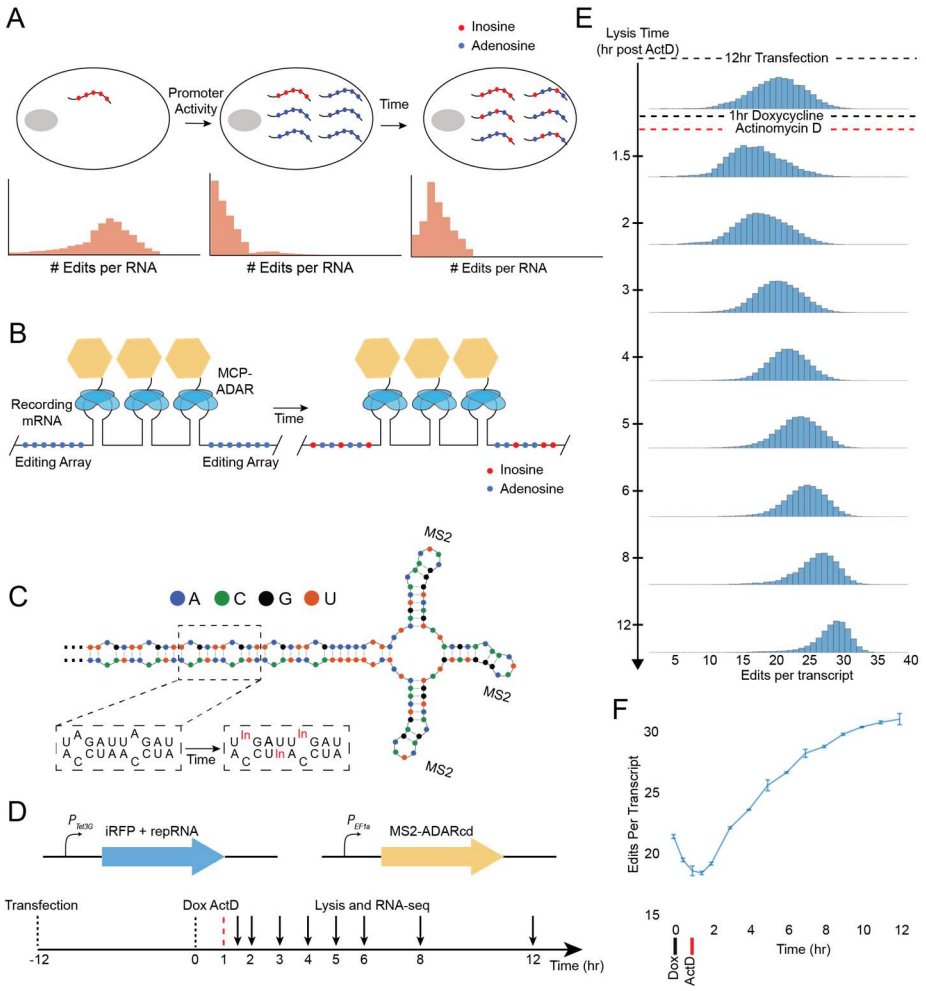

Figure 6-1: Encoding of temporal information through RNA edits. (A) Schematic of the RNA tickertape concept. The temporal history of promoter activity is determined by examination of the distribution of the number of A to I edits per RNA. Prior to promoter activity, the distribution is at steady state (left). A burst of promoter activity generates a population of new, unedited RNAs (distribution of edits per RNA shifts to lower values (center)). These RNAs are then gradually edited (distribution of edits per RNA shifts to higher values over time (right)). **(B)** Reporter RNAs (repRNAs) consist of editing arrays of adenosines (blue dots) and several MS2 step loops in the 3' UTR of an mRNA. In the presence of an MCP-ADAR fusion (MCP, blue ellipses, ADAR, yellow hexagon), repRNAs are edited over time by catalytic conversion of adenosine to inosine (red dots). **(C)** The structure of a portion of one repRNA, showing MS2 stem loops and the repetitive, double-stranded RNA motif that serves as the editing substrate. **(D)** Schematic representation of the Tet-responsive tickertape system and experimental timeline. **(E)** Transcription by the TRE promoter was induced by doxycycline, and was stopped by actinomycin D one hour later and sequenced as in the schematic of (D). Doxycycline induction shifts the editing distribution towards lower values as new RNAs are generated. After promoter activity ceases, the repRNAs accumulate edits and the distribution moves to higher values. All histograms are normalized so the sum of all values is 1. **(F)** Mean edits per transcript for TRE induction as a function of time for the TRE tickertape system. Error bars show standard deviation (s.d), N=3 biological replicates.

state. We asked whether it would be possible to design a reporter gene that would report the longitudinal temporal dynamics of gene expression in an RNA sequencing assay.

The ability to record temporal information about cell state into the sequence of nucleic acids would enable the interrogation of gene expression and cellular activity in cell populations or over timescales that do not lend themselves to imaging. For example, it has previously been proposed that a nucleic acid reporter for neural activity would allow for the interrogation of deep and otherwise-inaccessible populations of neurons, which would be transformation for neuroscience (*231*). In pursuit of similar goals, several labs have recently demonstrated the ability to record temporal information about cell state into the sequence of DNA (*232*, *234–243*). However, DNA is intrinsically a low-temporal-resolution recording device: DNA repair processes operate on timescales comparable to the generation time of the organism, whereas transcriptional programs are much faster, typically operating over timescales of hours in mammalian cells. Unlike DNA, RNA is regularly used by cells to store dynamic information about cell state with high temporal resolution over relatively short times, for example during progression through the cell cycle, or in the circadian rhythm (*260*, *261*). However, there is currently no known mechanism by which temporal information can be directly encoded, without a DNA intermediate, into the sequence of RNA. We here demonstrate, using RNA editing enzymes, the ability to encode temporal information about cell state into the sequence of RNA for subsequent inference via RNA sequencing.

Our goal was to design a system capable of estimating the magnitude of gene expression in one hour intervals stretching back for at least 12 hours. To build an RNA recorder, we reasoned that the history of the activity of a given promoter could be inferred from the distribution of ages of the RNAs generated by that promoter (Figure 6-1A). Conceptually, if reporters accumulate 1 edit per hour, then a population of 50 RNAs with 10 edits each corresponds to an event 10 hours ago, and a population of 10 RNAs with 5 edits each corresponds to an event 5 hours ago, with one fifth the magnitude. We designed reporter RNAs (repRNAs) that are capable of reporting their age via the gradual accumulation of A to I edits caused by an engineered version of the human Adenosine Deaminase Acting on RNA 2 catalytic domain (ADAR2cd, Figure 6-1B). The repRNAs consist of adenosine-rich editing arrays, in the 3' UTR of a mRNA encoding a fluorescent protein(*262*), that are designed to be favored substrates of the ADAR enzyme(*263–265*) (Figure 6-1C). Edits in this region can subsequently be identified as A to G mutations in high throughput sequencing of the repRNAs. ADAR2cd is specifically targeted to MS2 binding sites in the editing region of the repRNA through a fusion with the MS2 Capsid Protein (MCP)(*266*). We screened multiple repRNA and ADAR variants, and settled on a pair for which the editing in HEK239T cells occurs over hours, a timescale relevant for most endogenous transcriptional activity (Figure 12-1). We confirmed that the majority of edits observed is due to the MCP-ADAR fusion, rather than endogenous ADAR (Figure 12-2A). Furthermore, repRNAs do not degrade over the 12 hour observation time (Figure 12-2B), so information encoded into the repRNAs is not lost due to RNA degradation. We refer to the combination of a repRNA with the MCP-ADAR E488QT490A

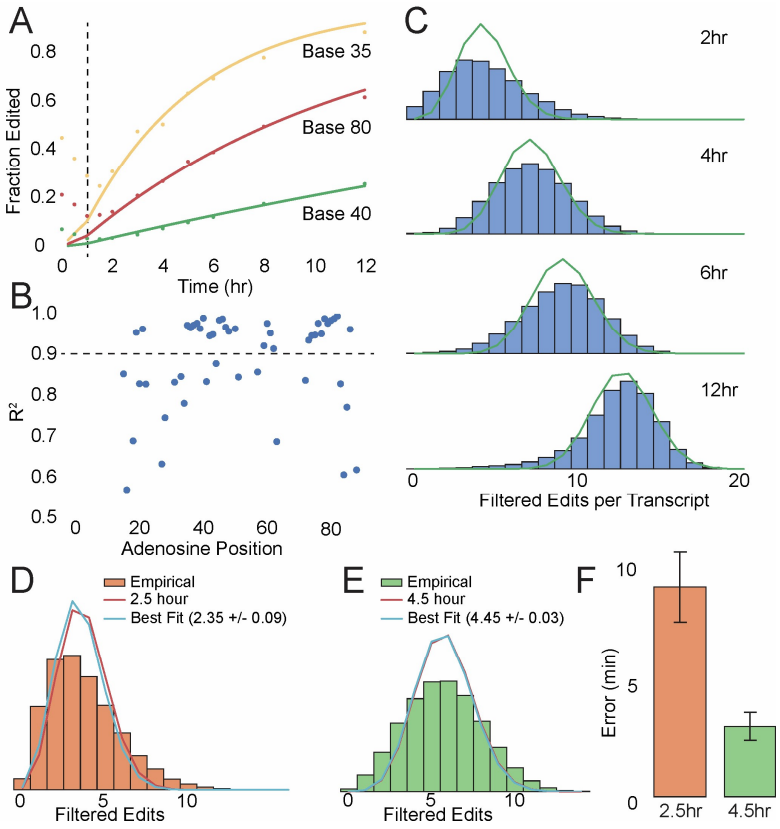

Figure 6-2: Inference of the timing of promoter activity using RNA tickertape. All editing histograms are normalized to sum to 1. **(A)** The fraction of A>I edits as a function of time is shown for three different bases on the repRNA, data from one replicate of 1E. Best exponential fits are shown. The black dotted line indicates the addition of actinomycin D. **(B)** For the same replicate as in (A), the R^2 value of the exponential fit is shown for each base on the transcript. The black dotted line indicates the $R^2 > 0.9$ cutoff used for the exponential model. **(C)** The masked editing histograms for four timepoints from the same replicate are shown (only the bases with $R^2 > 0.9$ are included). In green, the Poisson binomial distribution for each timepoint including all the bases with $R^2 > 0.9$ (see Methods). **(D)** In orange, the masked ($R^2 > 0.9$ in all 3 replicates from 1E, see Methods) editing histogram for a single 2.5 hour replicate along with Poisson binomial distribution for 2.5 hours (red line), and the Poisson binomial distribution with least KL divergence from the empirical distribution (blue line). The time estimate is mean ± s.d. (N=3 technical replicates). **(E)** As in (D), but for the 4.5 hour timepoint. **(F)** The mean absolute error is shown for the (D) and (E). Error bars show standard deviation.

protein, a variant with catalytic activity comparable to wild type ADAR but with reduced adenosine base-flipping activity (*264*), as the RNA tickertape system.

Results:

RNA Tickertape Infers the Timing of Isolated Transcriptional Events with High Resolution

To test the response of the RNA tickertape to a pulse of transcription, we incubated HEK293T cells expressing the RNA tickertape system under the control of the tetracycline response element (TRE) in medium containing doxycycline for one hour. We subsequently added actinomycin D, which inhibits RNA transcription(*267*), and then lysed the cells after a variable amount of time for repRNA sequencing (Figure 12-1D). As anticipated, we observed that a population of unedited RNAs was generated following doxycycline induction, and that these RNAs became gradually more edited over time (Figure 12-1E,F). The low variance (std=12 min +/- 8.25

min, N=15 timepoints) observed in the mean of the editing distribution between biological replicates suggested that the tickertape could be used for temporal inference.

In order to determine whether the system is in principle capable of inferring the timing of transcriptional events, we designed a statistical model to predict the RNA age distribution associated with a single transcriptional pulse as a function of time since doxycycline induction. If the adenosines on the repRNA template are edited independently and uniformly in time, then for each adenosine on the repRNA, the fraction of RNAs with adenosines at that site should decrease exponentially with the time since transcription, with a site-specific rate constant that depends on the local sequence context. For each adenosine on the repRNA, we fitted an exponential cumulative distribution function (CDF) to the editing fraction over time at that base (Figure 6-2A). We found 24 bases which fit well to the model (i.e., for which the value of R^2 was greater than 0.9 across all replicates) (Figure 6-2B). Analyzing only those bases, the distribution of edits per RNAs was well-approximated by a Poisson binomial distribution with a single parameter, t, which represents time since doxycycline was added to the medium (see Methods, Chapter 12), with the weights in the Poisson binomial distribution given by the exponential CDFs (Figure 6-2C). We used this Poisson binomial distribution to infer the times of cells induced at 2.5 and 4.5 hours prior to lysis, timepoints that had not been included in the dataset used to fit the exponential CDFs (Figure 6-2D,E). By minimizing the Kullback-Leibler divergence (which is equivalent to maximizing the likelihood) between the test distributions and the Poisson binomial distribution over t, we inferred that timing of those events to be 2.35hr ± 0.09hr and 4.45hr ± 0.03hr (mean ± s.d., N=3 technical replicates), respectively, implying that tickertape can localize individual transcriptional events with resolution less than 1 hour, as required.

The Poisson binomial approach is the preferred approach for estimation because it accounts for the exponential nonlinearity inherent in Poisson processes. However, we also found that a simple linear interpolation of the mean yields accurate estimations in many cases. In the case of the TRE tickertape, the mean interpolation estimated the 2.5hr and 4.5hr timepoints as 2.53hr ± 0.08hr and 4.38hr ± 0.02hr (mean ± s.d., N=3 replicates), with errors of 5min ± 0.3min and 7.5min ± 1.1min (mean ± s.d., N=3 replicates), respectively. To confirm that this accuracy is not limited to the TRE tickertape or to HEK cells, we performed similar experiments in 3T3 cells using repRNAs expressed under a light-inducible Vivid promoter, induced with blue light for one hour (*268*, *269*). We estimated the timing of light induction by interpolation of the mean number of edits per RNA, and yielded a temporal resolution of 17.7 ± 7.5 minutes (Figure 12-3, mean ± s.d., N=9 samples total across three timepoints). The fact that tickertape works with multiple promoters raises the possibility of recording the activity of multiple promoters simultaneously in a single cell population, and we validated that this is possible using barcoded repRNAs responsive to the Tet and Vivid promoters (Figure 12-4).

RNA Tickertape Infers the Timing and Magnitude of Complex Transcriptional Programs

Although the Poisson-Binomial model above achieves high accuracy for isolated transcriptional pulses, it is not applicable to more complicated transcriptional programs. In

]

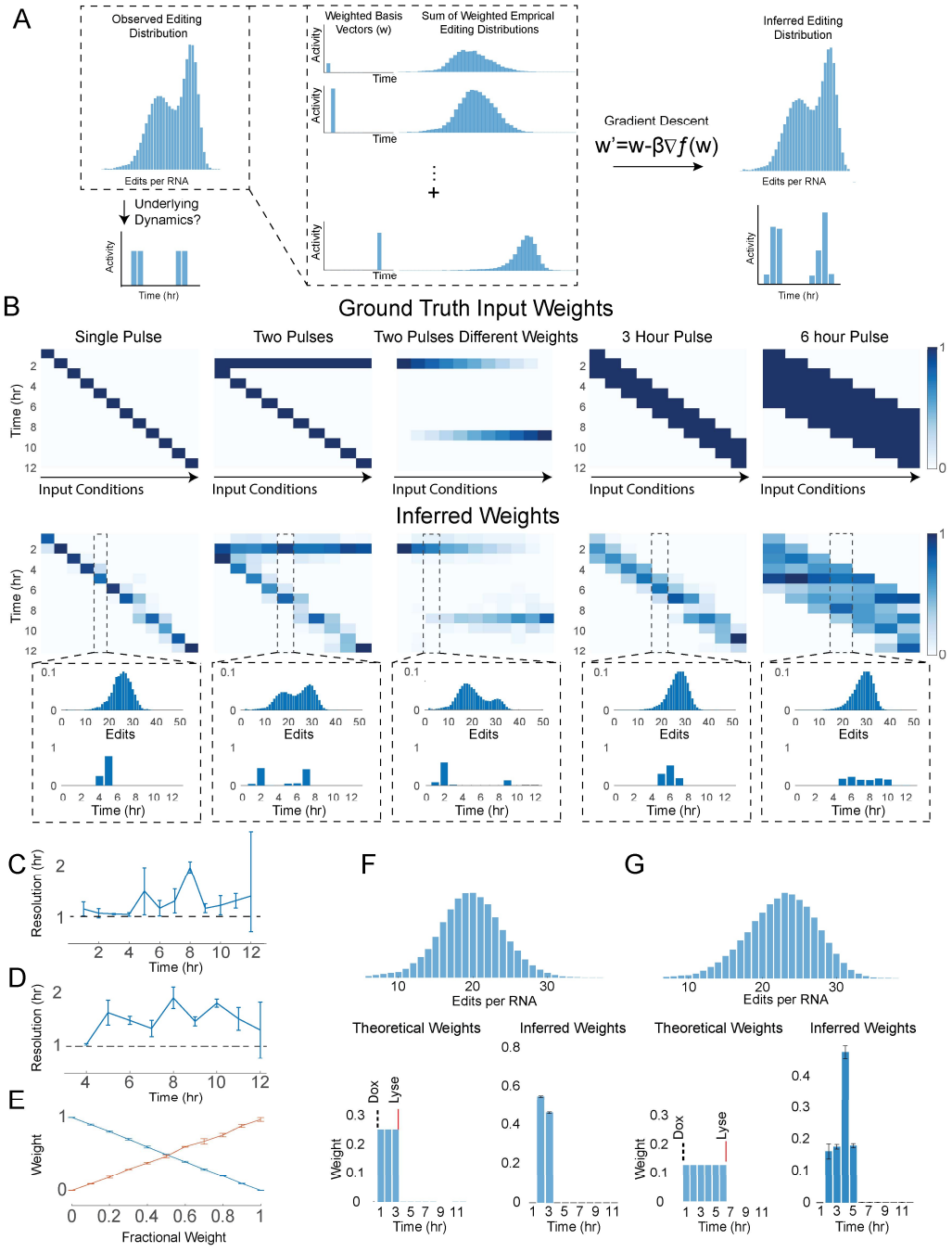

Figure 6-3: Tickertape is capable of decoding complex transcriptional programs. All editing histograms are normalized to sum to 1. **(A)** Arbitrary temporal patterns of transcriptional activity (top left) can be recorded into histograms of the number of edits per repRNA (bottom left). Arbitrary histograms can be modeled as convex sums of the one-hour distributions observed in the TRE tickertape experiments (middle). An approximation of the true history of transcriptional activity is recovered using gradient descent (top right) to minimize the difference between the observed editing distribution and the convex sum (bottom right). Caption continues on next page.

transcriptional events in the distribution of the number of edits per reporter RNA in the cell. In order to recover the timecourse of activity from the distribution, we built a general purpose decoder that estimates the transcriptional activity as a function of time in one-hour intervals stretching 12 hours into the past. Because the RNAs are edited independently, we reasoned that arbitrary transcriptional programs could be represented as convex weighted sums of the single-hour editing distributions (i.e., our one-hour "basis distributions") as measured with the TRE tickertape. Thus, for example, the editing distribution associated with two single-hour pulses could be represented as a weighted sum of the editing distributions for each of the single-induction pulses individually. We built a gradient descent algorithm to minimize the L2 norm between observed editing distributions and convex sums of these basis distribution (Figure 6-3A). As a first test, we applied the algorithm to the single-induction editing distributions themselves (Figure 6-3B, left). In all cases, to avoid overfitting, we averaged the editing distributions from two of the biological replicates in Figure 6-1E, and used the resulting averages as the basis functions for decoding the third replicate. The resulting estimates closely matched the expected single-hour profiles, and corresponded to a temporal resolution of 1.27h +/- 0.33h (mean +/- s.d. over all replicates and timepoints, N=36, see Methods, Chapter 12. Note that a temporal resolution of 1h would correspond to perfect estimation). Remarkably, the temporal resolution did not appear to depend on the length of time elapsed between induction and lysis (Figure 6-3C).

We next asked whether the tickertape is capable of identifying the presence of multiple transcriptional pulses. Since most complex transcriptional programs cannot easily be generated in cells, we generated simulated editing distributions as convex weighted sums of the single-induction editing distributions measured in our TRE experiments, and compared the weights obtained using the decoder to the ground truth weights used to simulate the data. To avoid overfitting, the data used as 'basis' functions in the decoder were distinct biological replicates from the data used to generate the simulated distributions. We simulated distributions consisting of a pulse at 2 hours and a second pulse of equal magnitude at subsequent times, which we refer to as a double induction condition (Figure 6-3B, center left). The decoder successfully identified the presence of two transcriptional pulses in every case, and estimated the timing of the second pulse with 1.5h

(B) Various transcriptional programs can be decoded using a general purpose tickertape decoder. Each panel shows the ground truth transcriptional program (top); the inferred programs (middle, predictions of 3 biological replicates), and the editing histogram and inferred weights for a randomly chosen example (bottom). From left: single-induction conditions; double-induction conditions, with two 1 hr pulses separated by a gap in time; double induction conditions with unequal amplitudes on each induction; three-hour continuous inductions; six-hour continuous inductions. **(C)** The temporal resolution of the predictions on the single-induction conditions as a function of time since induction. **(D)** The temporal resolution of predictions on the second induction in a double-induction condition as a function of time since induction. **(E)** The amplitude assigned by the decoder to timepoints ≥8 hours (orange) or ≤3 hours (blue) for each condition in the variable amplitude double induction tests. **(F)** The decoder applied to experimental three-hour induction conditions, with empirical editing histogram (top), putative ground-truth weight distribution (bottom left), and inferred weight distribution (bottom right) shown. **(G)** Same as F, for experimental six-hour inductions.

+/- 0.32h time resolution (N=27 timepoints), again independent of the delay between the lysis and the first pulse (Figure 6-3D). To determine whether the decoder is sensitive to the relative magnitudes of different transcriptional events, we mixed the 2 hour timepoint with the 9 hour timepoint with various coefficients of mixing (Figure 6-3B, center). We then calculated the total weight assigned to timepoints above 8 hours or below 3 hours (inclusive), respectively, and found that the decoder is sensitive to the amplitude of transcriptional events (Figure 6-3E). The decoder estimated the amplitude above 8 hours to within 5.3% +/- 3.8% of the true value; and estimated the weight below 3 hours to within 2% ± 2.5% of the true value. Thus, we conclude that the decoder is sensitive to both the relative timing and relative magnitudes of transcriptional pulses.

To determine whether this sensitivity extends to more complex transcriptional programs, we first applied the decoder to extended temporal square waves (i.e. pulses longer than 1 hour). For this case, the temporal resolution of the decoder is not well-defined, so we instead measure the percentage of the weight assigned by the decoder to the correct timepoints (see Chapter 12). For comparison, we note that the decoder correctly assigned 77.9% +/- 25.2% of the weight in the case of the single-induction estimates, corresponding to a time resolution of 1.27h. Applying the decoder to simulated pulses of 3 hours (Figure 6-3B, center right), the decoder correctly assigned 77.7% ± 12.2% of weight (mean ± s.d., N=3 replicates for each of 10 conditions, see Chapter 12). Applied to simulated pulses of 6 hours (Figure 6-3B, right), the decoder correctly assigned 83.3% ± 5.7% of the weight (N=3 replicates for each of 7 conditions). Thus, the accuracy of the decoder as applied to extended transcriptional programs is similar to the accuracy obtained for the single-induction timepoints. In order to evaluate the decoder on the most challenging case, in which the transcription rate may increase or decrease rapidly, we simulated random transcriptional functions by sampling the basis function weights from a 12-dimensional Dirichlet distribution. In this case, the decoder correctly assigned 71.9% ± 9.2% of the weight (mean+/-std, N=1000) (Figure 12-5). Although this represents a reduction in accuracy compared to the single-hour timepoints, it implies that the tickertape is still able to estimate arbitrary transcriptional programs with high accuracy. To evaluate the reasonableness of this accuracy on experimental data, we generated 3-hour and 6-hour pulses using the doxycycline-actinomycin setup described above (Figure 6-3F,G). In line with the expected accuracies from the simulations, the algorithm successfully assigned 66.7% of the weight for the 3 hour square wave (N=3 biological replicates. The value was exactly 66.7% for all three replicates.), or 64% ± 1.6% for the 6 hour square wave (N=3 biological replicates). Thus, the accuracy of the tickertape on experimental data is expected to be similar to the accuracy obtained on our simulated datasets.

RNA Tickertape Operates in Single Mammalian Cells

The accuracy of RNA tickertape depends on observing enough repRNAs that the empirical distribution of edits per repRNA accurately approximates the true distribution. One particularly interesting application of the RNA tickertape involves determining the relative timing of events in a population of single cells. We transfected HEK cells with barcoded TRE tickertapes, induced with doxycycline, and then followed one of three protocols: cells in condition 1 were left in doxycycline for 3 hours prior to lysis; cells in condition 2 were silenced with actinomycin D after 1

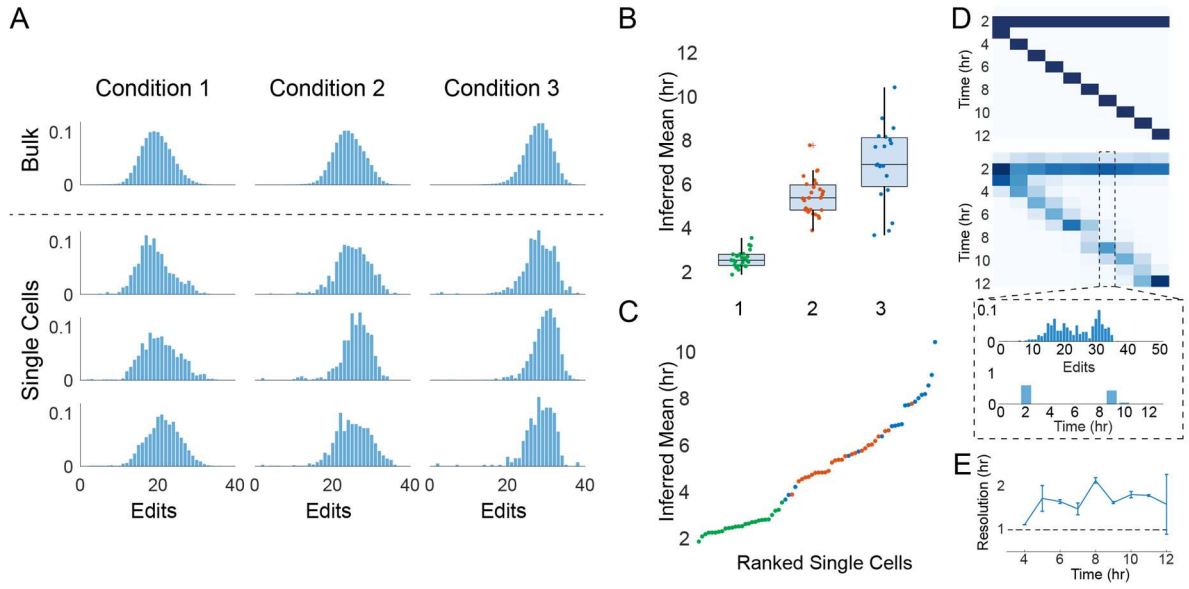

Figure 6-4: Tickertape can decode transcriptional programs in single cells. All editing histograms are normalized to sum to 1. (A) Editing histograms for bulk conditions 1 through 3 (top) and randomly chosen single cells (bottom). **(B)** Predicted induction times for all single cells in the experiment, calculated as the center of mass of the inferred weight distributions (N=24 for condition 1, N=27 for condition 2, N=19 for condition 3). **(C)** The predicted induction time for each single cell, ranked from least to greatest. Green dots correspond to condition 1; red dots correspond to condition 2; blue dots correspond to condition 3. **(D)** Predictions performed on double-induction editing distributions after subsampling to 300 reads. Top: ground truth transcriptional programs. Middle: mean prediction over 100 such samples. Bottom: One randomly chosen sample from the indicated condition, and its inferred editing distribution. **(E)** The resolution with which the second timepoint in the double-induction conditions can be inferred is shown as a function of time. All error bars show s.d., N=3 biological replicates at each of 10 timepoints.

hour, and then left for 3 hours, and cells in condition 3 were silenced with actinomycin D after 1 hour and then left for 7 hours. Individual cells were then sorted into wells of a 96 well plate, and we subsequently performed single-cell repRNA sequencing (Figure 6-4A). Applying our decoder to the resulting editing histograms yielded faithful estimates of the induction time: the absolute deviation between the temporal estimate for the single cells in condition 2 and a bulk of 100,000 cells in condition 2 was 1.2hr +/- 0.8hr (mean +/- std, N=27), while for condition 3 it was 1.5hr +/- 1.0hr (mean +/- std, N=19), which is similar to the temporal resolutions obtained for the bulk single-induction conditions above (Figure 6-4B, see Chapter 12). Thus, the tickertape is capable of recording the transcriptional activity of single cells.

The ability to order single cells according to the timing of transcriptional events would have great utility for studying the diversity of responses to cellular perturbations (*270*, *271*). To that end, we asked whether the tickertape can be used to order the individual cells from our single cell experiment, according to when the perturbation arrived. Ordering the cells according to their estimated times, we found that there were a total of 5 transpositions (i.e., 10 cells out of order)

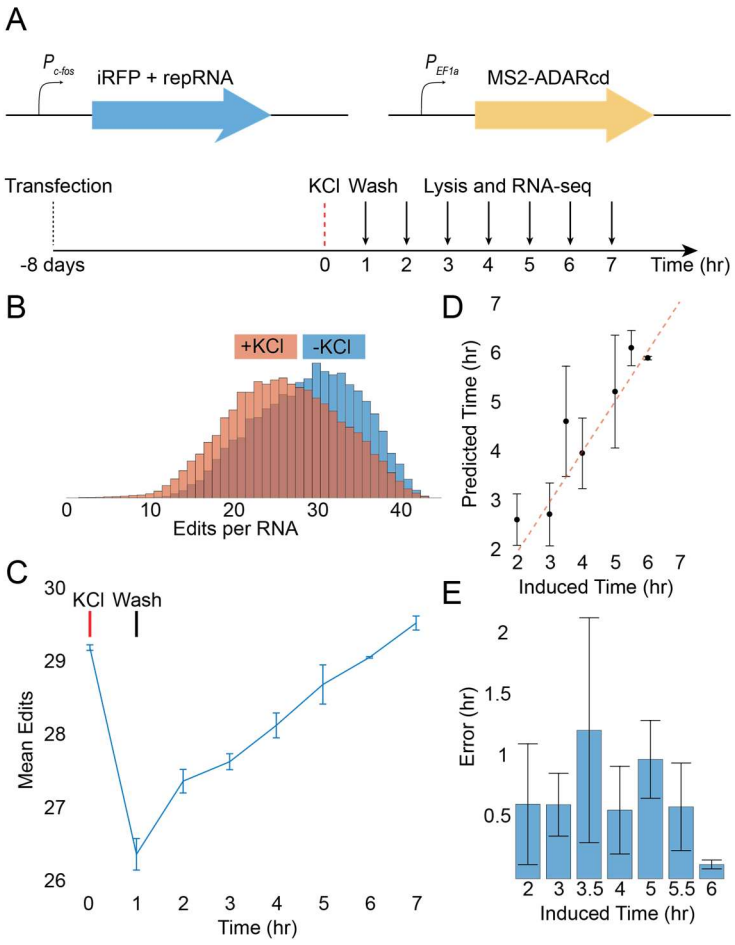

Figure 6-5: Sequencing-Based Activity Measurement in Neurons using c-Fos Tickertape. (A) Schematic of tickertape constructs and experimental timeline for neuronal recording. **(B)** Editing histograms are shown for neurons prior (blue) and one hour following (orange) KCl induction. The lower overall editing rate for the +KCl case indicates the generation of new repRNAs by the c-fos promoter. Editing histograms are normalized so the sum of all values is 1. **(C)** The mean editing rate is shown as a function of time following KCl induction. **(D)** The predicted and actual time estimates are shown for all timepoints. Dotted line is a guide for Y=X. There are no estimates for the 1 hour and 7 hour timepoints due to mean interpolation (see Supp. Fig. 3). **(E)** The mean absolute error in the predictions from (C) is shown as a function of time since induction. All error bars show s.d. (N varies, see Chapter 12).

out of 72 cells, an accuracy rate of 86% (Figure 6-4C). Finally, in order to determine whether the tickertape can encode the presence of multiple transcriptional pulses in single cells, we simulated distributions consisting of only 100 RNAs drawn at random from the double-induction distributions (Figure 6-4D). As in the case of the bulk double-induction, the decoder found two transcriptional pulses, and estimated the timing of the first pulse with an accuracy of 1.92h ± 0.35hr resolution (Figure 6-4E), thus demonstrating that the tickertape is capable of detecting multiple transcriptional events even with an extremely limited number of RNAs.

RNA Tickertape can be used to infer the timing of neural activity

All systems for temporally resolved detection of neural activity in single cells today rely on optical detection, or on the detection of electric or magnetic fields, and, as such, it is challenging to record from many neurons simultaneously, or from deep neural populations. We hypothesized that tickertape could be used to perform a sequencing-based readout of the transcriptional history of immediate early genes, which are often used for detection of neurons recently active in a neural network, but are mostly used to perform such measurements at single time points(*272*). We placed the repRNA expression under the control of a c-fos promoter, and

transfected the tickertape system into primary mouse hippocampal neuron culture at 6 days in vitro (DIV), which is used as a model for the study of coupling between excitation and transcription in neurons (*273*, *274*). At 14-15 DIV, we subsequently induced neural activity by adding a potassium-based depolarization medium to the culture (see Chapter 12) (Figure 6-5A). There was a clear shift in the repRNA editing histogram towards lower values following one hour of induction (Figure 6-5B), indicating that new repRNAs were being produced in a depolarization-dependent manner.

In order to estimate the temporal history of neural activity, we generated standards by inducing neurons for one hour with the depolarization medium, washing them back into normal (non-depolarizating) medium, and then lysing them at one hour intervals. For up to 7 hours after induction, a population of new repRNAs could be seen to gradually accumulate edits. Even in the presence of a large population of background repRNAs generated by constitutively *fos*+ neurons, the mean number of edits per RNA increased linearly over time (Figure 6-5C), at a rate of approximately 0.5 edits per hour. The linearity of the editing mean suggests that the editing mean should be a good predictor of the time since depolarization. We estimated the times of each replicate for the 2hr, 3hr, 4hr, 5hr, and 6hr timepoints by linear interpolation (see Chapter 12). We found that these replicates could be predicted from the standards with an average accuracy of 37 ± 23 minutes (Figure 6-5D,E, mean ± s.d.), which is comparable to the ~1hr temporal resolution intrinsic to immediate early gene transcription. Then, in addition, we stimulated neurons at 3.5 and 5.5 hour timepoints, and found that these could be predicted with an average accuracy of 72 ± 55 and 35 ± 22 minutes, respectively. Thus, RNA tickertape accurately reports the timing of immediate early gene transcription in neurons.

Discussion:

RNA tickertape is a novel molecular recording device that enables the recording of the temporal history of transcription into the sequence of RNA. It is likely that, using the same concept, alternative systems could be designed that record other kinds of signals, besides transcription. For example, by using alternative dimerization systems (*275*, *276*) to link ADAR to constitutively expressed repRNAs in a stimulus-specific manner, it may be possible to construct tickertapes that report on the timing of other kinds of cellular events, such as calcium or other signaling molecules. Together, these observations suggest that RNA tickertape is a scalable and extensible approach for recording the temporal activity of cells.

Chapter 7
Molecular Barcoding for Connectomics

Throughout my graduate school career, I thought extensively about strategies for molecular barcoding of neurons, with an eye towards connectomics. Fundamentally, optical barcoding approaches involve labeling neurons in some way that allows for the identity of the neuron to be inferred by imaging them in some number of color channels with an optical microscope. In the simplest approach, each channel is either present or absent in a cell, providing 2^N combinations, where N is the number of channels. As discussed below, if it is possible to label individual molecules with multiple channels, and if it is possible to distinguish those molecules optically, so that each cell can contain *multiple combinations of channels*, the combinatorial diversity could be much higher. This is the case for RNA barcoding approaches, in which the RNAs are typically relatively sparse in the cell, or for protein-based approaches with sufficiently strong superresolution microscopy to enable single-molecule imaging.

From February to October 2015, I worked with Noah Jakimo on a strategy called Brainbar, inspired by a strategy conceived by Adam Marblestone (*277*), for delivering barcoded RNAs to the processes of neurons. These barcodes were designed to be read out using multiplexed FISH techniques, rather than direct in-situ sequencing techniques as originally proposed, because I was convinced that in-situ sequencing protocols were too challenging and took too long to achieve widespread adoption. Leveraging the ability to image RNAs at the single molecule level, the Brainbar barcodes were designed in a way so that sufficiently high combinatorial diversity could be obtained in a single round of imaging, rather than in many successive rounds of imaging as is necessary for in-situ sequencing. However, the method failed at the first hurdle: Noah and I found that RNA barcodes expressed off of the U6 Pol III promoter never left the nucleus in neurons, and RNAs expressed off of a Pol II promoter (such as CAG) are not produced at high enough concentrations, and do not traffic in the processes, despite our best efforts to improve the trafficking, for example using RNA degradation resistance elements (*278*, *279*). In October 2015, I became convinced that protein-based barcodes, rather than RNA barcodes, were the correct path forward, because they avoided the RNA trafficking, stability, and expression issues. (GFP expressed in a neuron will fill the cell without any engineering.) The remainder of this chapter proposes a similar barcoding approach, in which proteins would be labeled with many different epitopes. The proteins would be imaged at the single molecule level, allowing each protein to be imaged in multiple optical channels, thus constructing the barcode.

Nonetheless, I became discouraged by the difficulty of the 20x expansion protocols that would be necessary to resolve individual protein molecules, and did not begin working on these ideas experimentally until April 2017. At that point, I realized in a conversation with Nick Barry that

the simple 2^N scaling obtainable with non-single-molecule protein barcoding would be sufficient for connectomics if we could image in ~20 to 30 channels. However, imaging a cell in 20 to 30 channels would require a system for multiplexed imaging, since one can typically only image 4-5 channels at a single time in a standard optical microscope. Nick proposed that we could use MIBI to read out a many-color Brainbow, but MIBI and related hyperspectral systems are generally point-scanning systems, and are thus too slow for most applications (*42*, *280*). At the time, high-quality protocols for multiplexed antibody staining in expansion microscopy (which we assumed would be essential to achieve high enough resolution to visualize spines) did not exist. We considered antibody-oligo conjugates as a way to attach oligonucleotides to the barcode proteins, but good antibody-oligo conjugate protocols also did not exist.

We also considered whether we could conjugate antibodies to DNA binding proteins, like TAL effectors, as a way to attach oligonucleotides to the antibodies *after staining*. In May 2017, I wondered in a conversation with Adam Marblestone whether we might be able to simply to express the TAL effectors in the cells as a way of generating barcode proteins that could be stained with oligonucleotides. The key question was whether the TALEs would retain their ability to bind DNA after fixation. In a one-day experiment, I expressed two TALEs in HEK cells, and showed that they retained their binding activity and specificity when the cells were fixed in methanol, but not PFA. Nick and I then performed a number of experiments over the following year. Remarkably, both TALEs and zinc fingers have this property, and zinc fingers are even somewhat robust to formaldehyde fixation. The method worked extremely well in culture, but staining in vivo was weak.

However, the landscape has changed dramatically since we first began this project, and we reevaluated the landscape in late 2018 and determined that the time was ripe for direct barcoding via direct antibody staining. Several factors informed this decision: firstly the publication of spaghetti-monster fluorescent proteins provided us with a method for delivering epitopes on a fluorescent protein scaffold (*281*); recombinases had been shown to be ineffective for the application described in this chapter (*282*), but a new family of blood-brain-barrier-crossing viruses provided an alternative way to generate combinatorial diversity (*283*–*285*). The MARC1 mouse demonstrated that it is in principle possible to generate a mouse line with 30 or more transgenes, so a 30-color protein barcode would be compatible with a transgenic ultimately (*286*). Finally, a number of new ExM-compatible antibody-multiplexing approaches had been published (*14*, *15*), as well as new oligo-conjugated antibody methods (*287*). Bobae An joined our team at that time, and has been instrumental in spearheading the new protein barcoding method. As of the publication of this dissertation, the work is still ongoing.

Summary:
We suggest a simple protein-based multicolor optical strategy for uniquely barcoding large numbers of neurons, based on tractable genetic methods and enabled by Expansion Microscopy (ExM) (*37*). The diversity generated scales super-exponentially with the number of available colors. We discuss the application of this strategy to barcoding entire Drosophila or larval Zebrafish brains using only off-the-shelf recombinase-based cassettes and 6-color microscopes, as well as its extension to mammals.

Introduction

Scalable arbitrary-color optical super-resolution

We recently developed Expansion Microscopy (ExM) (*37*): rather than using lenses to create optical magnification in a microscope, we recently found that physical magnification of the specimen itself is possible. Polymerizing electrolyte monomers directly within a sample to form an electrically charged polymer network, followed by solvent exchange, results in specimen expansion. By covalently anchoring specific molecules within the specimen to this polymer network and proteolytically digesting away unwanted endogenous biological structure, we found that samples could be expanded isotropically 4.5-fold in linear dimension. We discovered that this isotropic expansion applies to nanoscale structures, and thus this method can effectively separate molecules located within a diffraction limited volume, to distances great enough to be resolved with conventional microscopes. As a side effect, this process renders the sample transparent.

In the first paper [2], we expanded tissue by 4:5linearly ($>$ 100volumetrically). Crucially, in recent work using novel expansion polymer strategies (*288*), we can expand tissue by up to roughly 20 linearly, implying that a diffraction-limited microscope with $300\,\mathrm{nm}$ optical resolution can achieve an effective resolution post-expansion of $15\,\mathrm{nm}$. This works with an arbitrary color-palette of fluorescent dyes.

The power of multiple colors:

Our main question here is: given this new capability in scalable, fast, multicolor, fully 3D, super-resolution optical imaging, can we devise a multicolor optical labeling strategy which would robustly enable connectome extraction? Ideally, the use of multiple colors would allow us to extract connectomes from optical imaging data without the need for complex machine learning or human annotation. Specifically, the use of multiple colors could serve to disambiguate neuronal identities in cases where the pure membrane geometry appears ambiguous. Furthermore, multiple colors could be used to error-correct one another: where one color or label fails or exhibits an ambiguity at a given position in the neural geometry, another color could potentially "fill in the gap." Here we propose to take this approach to the extreme, endowing each neuron in a brain with a unique "color code" that identifies it on the basis of its color contents, over and above the information that can be gleaned by tracing its morphology.

We will show that multicolor labeling could enable unique neuronal barcoding with only moderate requirements on fluorophores, epitopes, recombination sites, and other biotechnological primitives. Moreover, we explain how the available barcode diversity can be made to scale super-exponentially with the number of available colors. In what follows, we first remind the reader of prior approaches for unique cell barcoding, then describe schemes for increasing the label diversity super-exponentially, and then illustrate examples for labeling animal brains of various sizes.

Prior approaches for optically barcoding neurons

Nucleic acid-based barcoding

Recently, researchers have proposed to endow neurons with unique genetically-encoded molecular barcodes in the form of RNA strings, which can be read out through bulk sequencing (*289*), or through in-situ RNA sequencing in an optical microscope (*277*, *290*), or through multiplexed in-situ hybridization (*9*, *10*). Such unique RNA labels can be read out from any point in a cell, regardless of distance from the parent soma, removing the need for complete image-based morphological tracing of the cell's geometry. These methods, however, as currently conceived, require a large amount of sequence diversity to be encoded into a single RNA strand and then to be read out over multiple cycles of chemical interrogation of the RNA. Thus, an in-situ barcoding technique that achieves high label diversity with more facile genetic and readout techniques would be desirable. We will show below that this can be achieved by splitting the barcode information over multiple molecules.

Brainbow Barcoding

Researchers have endowed neurons with random cell-specific combinations of fluorescent protein expression levels, giving neurons distinct fluorescent colors under the confocal microscope, a so-called “Brainbow” method (*291*, *292*). The cell can manufacture enough of the Brainbow label proteins to enable almost complete coverage of the cell membrane. It is anticipated that such protein-based labeling will have an advantage over RNA-based barcoding methods which, due to the sparser number of RNAs per cell, may not be able to completely tile a neuron with cell-identifying barcodes. Unfortunately, the number of distinguishable colors generated by Brainbow has been limited to a few hundred, which is not sufficient to disambiguate densely-labeled neural circuitry over long distances. Thus, a protein-based optical barcoding technique which achieves greater effective color diversity would be desirable. We will show that this can be achieved by splitting the barcode information over multiple proteins or subcellular structures each of which can be interrogated individually rather than as mixtures.

Concepts:

Criteria for an optimal barcoding strategy

With the emergence of scalable, arbitrary-color super-resolution optical microscopies comes an opportunity to invent new labeling strategies. We will rely on a generalized Brainbow-type

approach, in which recombinases stochastically diversify genomic cassettes, leading to a stochastically chosen set of proteins that serves as a unique barcode for a given neuron. These proteins will then be imaged via ExM using antibodies conjugated to fluorescent dyes. Our strategy aims to combine several attributes:

1) **Combinatorial genetic diversity:** labels sufficient to uniquely label every cell in a brain must be encoded into the genome. The diversity of labels that can be generated is limited at the genotypic level by the number of orthogonal recombinase sites that can be used simultaneously in a single organism: 10 in current practice. For example, in Drosophila we are aware of 4 orthogonal recombinases (*293*), one of which (Flp) is already known to possess 3 orthogonal sites (*292*), for a total of a least 6 orthogonal sites. If we add three known orthogonal Cre sites and a ϕC31 site (although ϕC31 is irreversible, restricting the number of values that a single cassette can generate), we expect to be able achieve at least 10 sites in Drosophila. It may be possible to increase this number further, e.g., via generating additional orthogonal LoxP sites for Cre. In addition, up to 7 putatively orthogonal sites have been reported elsewhere for Flp (*294*), and up to 6 orthogonal pairs of attP/attB sites have been reported for ϕC31 (*295*). More speculatively, Appendix 2 (Chapter 13) discusses potential means to engineer greater genotypic diversity including temporal multiplexing of recombination sites by using inactivated Cas9 to block specific sites at specific times. The maximum genetic diversity is also limited by the number of possible values that can be generated by each cassette, which has been limited to 4 in prior BrainBow systems, although below we will discuss extensions to > 7 values per cassette.
2) **Protein-based barcodes:** to enable high copy-numbers of labels and hence complete filling of the cell. Our proposed readout scheme is based on antibody staining of epitopes attached to scaffolding proteins (*281*). The primary limitation here is the number of epitopes/antibodies that can be probed simultaneously via immuno-histochemistry[1]: the number of primary and secondary antibodies that must be used simultaneously will be assumed ≤ 10 initially, with the possibility of adding a few more via additional efforts in some cases.
3) **Spatial resolution:** we will assume next-generation 20x ExM enabling roughly $15\ \text{nm}$ spatial resolution, i.e., that we can resolve multiple discrete molecules or distinct

[1] When a large number of orthogonal epitopes are required to be labeled in a single wash, and secondary antibodies are to be used (rather than just labeled primaries), we will need a sufficiently large set of orthogonal secondary antibodies, e.g., to achieve a set of size 12 we may need to target distinct primary antibody isotypes in addition to the host species of the primary antibody. It may also be possible to use pre-adsorbed secondaries. Spectral multiplexing of up to 10 simultaneously applied fluorescent antibodies has been demonstrated, and more for mass-spec-based readout of primary antibody identity (*350*).

sub-cellular structures (actin filaments, microtubules, membrane, mitochondria) even inside a single thin axon.

4) **Protein expression level:** supposing that we can generate at most 10^7 genetically encoded label protein molecules per cell, then a $1\,\text{mm}$ long axon of $300\,\text{nm}$ diameter would exhibit an expressed protein density of at most 1 protein per $20\,\text{nm}$ cube if proteins were randomly distributed in 3D the cytosol, or 1 protein per $10\,\text{nm}$ square if proteins were randomly distributed in 2D on the membrane. In other words, operating at $15\,\text{nm}$ resolution voxel size, we have roughly one protein per resolution voxel. Thus, imaging expressed proteins at this level becomes a single-molecule problem.
5) **Digital rather than analog encoding:** for noise-robustness with respect to color intensities. Specifically, we prefer encoding of information in the discrete subunit composition of individual molecules, rather than in molecular densities, for robustness to expression-level noise, although we will also consider digital control of protein expression level as a useful primitive in certain cases. With digital encoding, the primary limitation is the number of orthogonal fluorescent channels available, ≤ 10in standard practice, although it should be possible to extend beyond 10 channels with additional efforts.
6) **Ability to read out the labels in a single cycle of imaging** , without the need for many successive fluidic wash steps as in FISSEQ-BOINC (*277*). We will assume that only one round of imaging and labeling is possible, although multi-wash labeling and imaging has been demonstrated in array tomography (*296*), serial FISH (*297*), FISSEQ (*290*) and similar methods.
7) **No need for novel genetic diversification methods:** adaptations on existing recombination-based methods should suffice.

Genotype/Phenotype Diversification

Consider a system using C orthogonal recombinase sites, generating v possible stochastic values per cassette. C genome-integrated cassettes are distinguished by their use of orthogonal recombination sites, which prevents inter-cassette recombination. For example, in a system using 3 orthogonal Cre/Lox sites and 3 orthogonal Flp/Frt sites we would have $C = 6$. Cre recombinase is capable both of excising the region between similarly-oriented LoxP sites, and of reversing the region between oppositely-oriented (inverted) LoxP sites. Between every pair of inverted LoxP sites, a region can be inserted which codes either for a transcript A in the forward state, or a different transcript B in the reverse state. We will call such a region an "inversion unit." With two such inversion units in tandem, it is possible to encode $v = 4$ distinct transcripts onto a single cassette, each chosen stochastically upon recombination with roughly ¼ probability, as was used in the original BrainBow system (*291*).

With more than two inversion units in tandem the distribution over the possible recombination states of the cassette may not be uniform (see Appendix 1, Chapter 13), by virtue of the

possibility of internal excisions, which bias the system away from the "middle" states. To evaluate this possibility, we performed simulations of the recombination process. In simulations of a 4-value cassette (i.e., two inversion units in tandem), it was found that all 4 values are achieved with equal probability, which is to be expected since there are an equal number of recombination pathways to each value. In the simulations of an 8-value cassette (i.e., four inversion units in tandem), the Shannon entropy of the resulting distributions over recombination sites was approximately 2.9 bits (as compared with $\log_2 8 = 3$ bits for a perfectly uniform distribution) if we allow a sufficient amount of recombination for each cassette to approach equilibrium. This corresponds to $2^{2.9} = 7.46$ effective values per cassette. When only a few recombinations occur per cassette, the entropy can drop below 2.8 bits. This behavior is shown in Appendix 1 (Chapter 13). The above scheme can generate v^C distinct genotypes. Using cassettes with two inversion units, hence $v = 4$, and 9 orthogonal recombinases, hence $C = 9$, we have $v^C = (2^2)^9 = 262144$, or 18 bits, sufficient to barcode the Drosophila brain or larval Zebrafish brain. More generally, in order to be able to assign a unique genotype to each of N neurons, we must have

$$C \log_2 v \gg \log_2 N$$

The term $\log_2 N$ is the total number of bits needed to assign a unique genomic ID to every neuron in the system in question. In Figure 7-1, the necessary number of bits needed is shown for some model organisms. It is then up to the experiment designer to design a system in which all of those genotypes translate into optically-distinguishable cellular phenotypes that can be read out from small regions of interest at arbitrary positions along the length of the axon.

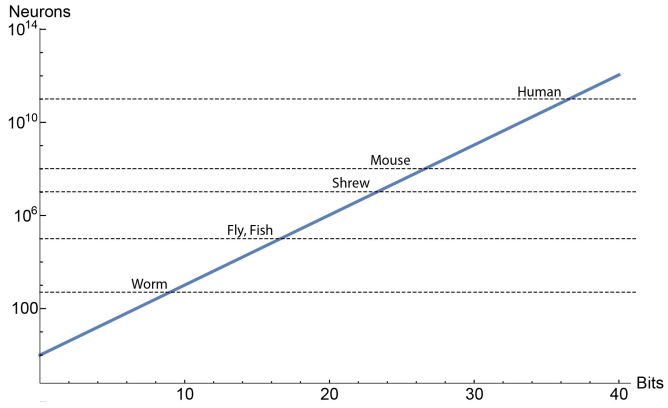

Figure 7-1: Number of Bits Needed for Uniquely Barcoding Animal Brains. The number of neurons that can be orthogonally labeled by a genotyping strategy is shown as a function of the number of bits encoded by the genotyping strategy.

Readout

Once neurons have been assigned unique genotypes, the corresponding phenotype must be read out optically. The key requirement on the readout strategy is that it must have a bit capacity at least as large as the bit capacity of the genotyping method.

Peptide Epitopes

Our phenotyping strategy will rely on the expression of peptide epitopes displayed on scaffolding proteins (*281*) (see Appendix 5, Chapter 13, for a discussion of RNA labels). These epitopes can then be detected either using fluorescently labeled primary antibodies, or using primary antibodies

followed by fluorescently labeled secondary antibodies for signal amplification. In the absence of an additional mechanism of multiplexing, a strategy with E epitopes can encode only E bits, i.e., 2^E possible phenotypic states, where we assume that an epitope is either present or absent but that we cannot readily distinguish analog levels of epitope. We seek to design a system that requires as few orthogonal epitopes as possible, especially if signal amplification using secondary antibodies is required[2].

Number of Colors

We will denote the number of spectrally orthogonal fluorophores by F. In preliminary experiments, we have been able to perform single-molecule imaging of 6 orthogonal fluorophores on a microscope with 4 lasers. Illumination systems are available from Coherent with up to 8 lasers. Using such a system, we expect that we could easily perform amplified single-molecule or bulk imaging of 9 orthogonal fluorophores. In Appendix 3 (Chapter 13), we illustrate a possible strategy for achieving up to 12 orthogonal fluorescence channels, without the need for spectral demixing.

Readout Multiplexing

In general, the genotypes and phenotypes assigned to cells in the barcoding strategies we propose will contain more information than $F \approx 12$ bits of information, i.e., a pure digital Brainbow strategy with $2^{12} = 4096$ color combinations would have insufficient readout diversity. Even with 3 distinguishable intensity levels and 12 fluorescence channels, we have only $3^{12} = 531441$ possible readouts which is insufficient for barcoding whole mammal brains. Moreover, we wish to use a limited number E of epitopes/antibodies. Thus, an additional multiplexing strategy will be required to read out the phenotype, beyond just a digital Brainbow approach, i.e., a bulk per-cell expression level for each of $F \leq 12$ fluorophores. There are several options for the mode of multiplexing:

Temporal Multiplexing

Temporally multiplexed readout strategies can increase the number of bits that can be read out per fluorophore. In particular, if one has F fluorophores and performs w reagent washes, one can read out wF bits, e.g., for temporal multiplexing via sequential immuno-histochemical wash cycles (*296*), or sequential hybridization (*10*, *297*) to antibody-identifying DNA barcodes. However, because it is assumed that the epitopes are distributed throughout the cell prior to imaging, temporal multiplexing does not increase the number of bits that can be read out per epitope, and is thus only useful given more epitopes E than fluorophores F. In line with the criteria outlined

[2] Again, up to 10 orthogonal secondary antibodies and fluorescent color channels have been demonstrated in immune histochemistry (*350*), to our knowledge.

above, we will limit ourselves to only a single wash, $w = 1$. We will thus assume that we have at least as many fluorophores as epitopes.

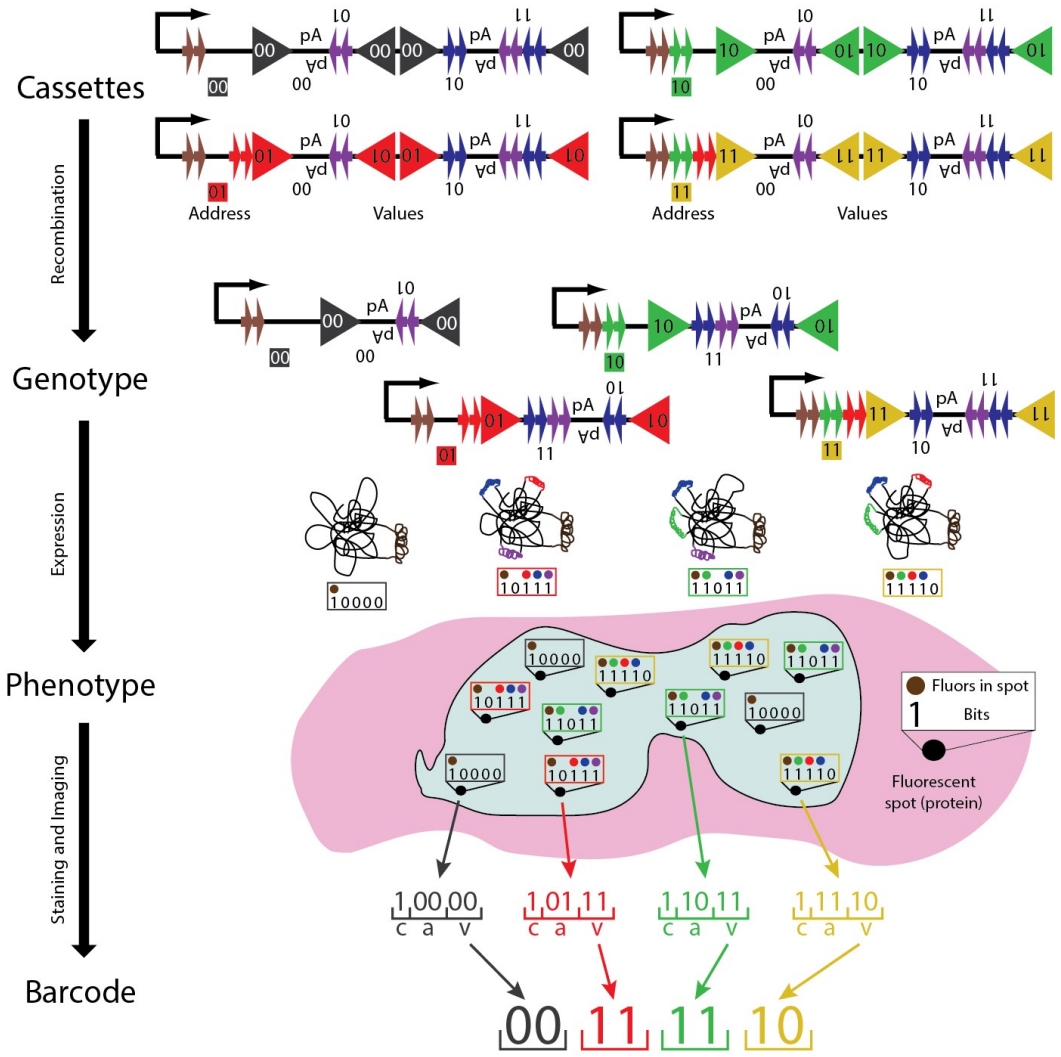

Figure 7-2 Address-Value Connectomic Phenotyping Strategy. Illustration for the case of 4 cassettes, each of which encodes 4 possible proteins, for a total of 44 or 256 labels, which can be read out with 5 epitope-fluorophore pairs (represented here by brown, green, red, blue and purple dots). The initial genome-integrated cassettes (top) undergo recombination, producing a randomized genotype within each cell. The corresponding phenotype consists of four generalized spaghetti-monster fluorescent proteins (smFPs) – designed for efficient immuno-staining – each of which is labeled with a subset of the 5 available epitopes. Two fluorophores (red and green) indicate which cassette gave rise to the protein (the address bits), while two fluorophores (blue and purple) indicate the outcome of recombination (the value bits). The fifth fluorophore is always present. The barcode is given by the value bits, ordered according to the addresses. Readout is performed by antibody staining against the epitopes. The proteins are allowed to remain in the cytosol, and expression must be controlled such that following expansion, there is at most one protein per diffraction-limited spot (or more generally per microscope resolution voxel). This method achieves higher diversity than the structural labeling method for the same number of epitope-fluorophore pairs, because more than 3 fluorophores are used by each cassette. While the protein-epitope readout case is emphasized here, RNA FISH-based or FISSEQ-based readout of address-value barcodes is also possible. pA denotes a polyadenylation sequence that terminates transcription.

An alternative approach to multiplexing involves reusing the same set of peptide epitopes on multiple different protein labels, and then spatially separating these labels in a way that allows them to be distinguished from one another. Then, a labeling scheme with E epitopes and s spatially-separated labels could in principle encode sE bits of information. Unlike in the case of temporal multiplexing, spatial multiplexing strategies increase the amount of information obtained per epitope, as well as per fluorophore.

In an exemplary such scheme, "address-value" multiplexing (Figure 7-2), protein labels generated by different cassettes are allowed to diffuse through the cytosol or along the membrane, and are imaged at the single-molecule level with physical amplification of signal from each single molecule. Spatial multiplexing is achieved by having multiple cytosolic label proteins, each with multiple fluorophores, and expressing these proteins at a low enough density that there is at most one protein per diffraction-limited spot in the post-expansion sample, yet high enough density that a given small region of interest (ROI) contains at least one of each of the multiple distinct labels. The combination of fluorophores present or absent on a protein label then indicates both a) which cassette the protein came from, and also b) the result of recombination.

In order to obtain the maximum amount of genetic diversity, the proteins generated by each cassette must be distinguishable from each other. In the address-value approach, each cassette uses the same epitope/fluorophore pairs to encode the result of recombination, and cassette identity is indicated by a separate, dedicated register of "address" epitope/fluorophore pairs, which are displayed by the corresponding label protein regardless of the result of recombination.

Alternative spatial multiplexing methods might rely on targeting label molecules to distinct cellular structures (Appendix 6, Chapter 13).

Expansion microscopy can be used to ensure at most one label protein per diffraction-limited spot. For example, using a 10-fold linear expansion factor and expressing labels at a density of one per cube of $20\ \text{nm}$20nm on a side in the pre-expansion space, and then imaging with a microscope capable of $200\ \text{nm}$ 3D resolution, we could fit $\left(\frac{100}{20}\right)^3 = 125$ label molecules within an $(100\ \text{nm})^3$ ROI. This should be more than enough to sample at least one of 10-20 distinct species of label molecule. Moreover, after ExM expansion, amplification of label brightness is possible, since each label molecule now has on the order of 200nm of physical space around it that can be occupied with bulky groups such as primary and secondary antibodies, hybridization probes, quantum dots and so forth. For example, with smFP-based antigens (*298*), since each smFP could carry many epitopes, and each primary and/or secondary antibody can carry many fluorophores, one can easily imagine upwards of 100 fluorophores being recruited to each molecular label.

Spatial grouping of label molecules

One challenge with the optical readout strategy presented here is that it depends on being able to find all of the C different label molecules in a local region of interest that manifestly corresponds

to a single cell, i.e., single contiguous local patch of a single neuron. We have chosen peptide-based barcodes because peptides can be expressed to high levels, enabling high densities of label molecules. However, we still require a means to determine if a cluster of nearby label molecules belong to the same neuron, or are split across two or more neighboring neurons. We anticipate that, using high expansion factors to ensure high spatial resolution, and ExM-compatible chemical lipid stains to delineate cell boundaries, such local grouping of label molecules should be possible in many, but probably not all, regions of a neuron. Note that chemical lipid stains can tile the membrane with high density, approaching that of the osmium stains used to label lipids for electron microscopy. In any ROI where this is possible, we will have identified the neuron to which that ROI belongs. This should enable a form of error-correction that would allow morphology-based connectome mapping to bridge across otherwise un-traceable gaps, such as in long-range axon fascicles. In an exemplary scheme, labels could be positioned directly on the inside (i.e., cytosolic rather than extracellular side) of the membrane via an appropriate fusion to a membrane-anchored protein, with the membrane itself marked by a chemical lipid stain.

If we sample more labels per ROI than is strictly necessary, we can detect and correct labeling failures due to, e.g., failure to observe a given fluorophore on a given label molecule due to failure of antibody binding, fluorophore bleaching, or other causes that become relevant at the amplified-single-molecule level. For example, to collect all 18 labels in an 18-cassette scheme, the well-known solution of the Coupon Collector Problem shows that we require 63 samples on average, whereas we have 125 samples if we observe an ROI of $(100\,\mathrm{nm})^3$ with label molecules packed at a density of one per $20\,\mathrm{nm}$ cube.

Optimal distribution between address and value bits

In the address-value barcoding scheme, $\lceil \log_2 C \rceil$ epitopes are needed to encode the identity of the cassette, and it follows that the rest of the epitopes can be dedicated to labeling the possible values that can be assumed by the cassettes upon recombination. With K orthogonal epitopes and corresponding spectrally orthogonal fluorophores available, it follows that the number of possible distinguishable phenotypes is

$$P = \left(2^{K-\lceil \log_2 C \rceil}\right)^C \tag{2}$$

It is interesting to note that, ignoring the ceiling operation (i.e., allowing fractional fluorophores), P achieves its maximum for fixed K when $C = 2K/e$, where e denotes the base of the natural logarithm. Alternatively, denoting by f_C the number of fluorophores used to label cassettes, we find that

$$f_C = K - \log_2 e \tag{3}$$

Perhaps counter-intuitively, the information obtained from an address-value barcoding system is maximized when nearly all fluorophores are used for address bits. Therefore the optimal strategy

with 8 fluorophores would use $2^7 = 128$ cassettes, each encoding only two possible values, thus generating $2^{128} = 3.4 \times 10^{38}$ possible barcodes. A barcoding scheme with 128 cassettes is impractical, but fortunately even distributions with far fewer cassettes are capable of generating huge diversities, see Figure 7-3B.

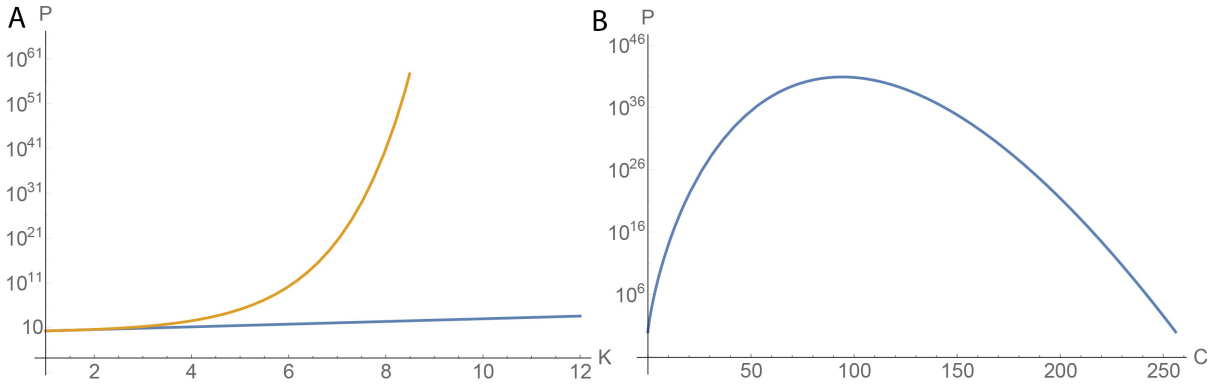

Figure 7-3: Scaling of the Address-Value Barcoding System (A) The number of distinguishable phenotypes P that can be read out in a single wash is plotted as a function of the number of fluorophore-epitope pairs K for the $\rho \gg 1/V$ strategy with 2^K cell colors derived from bulk mixing (i.e., digital BrainBow), and for the $\rho \ll 1/V$ address-value barcoding strategy with the optimal number of cassettes $C = 2^K/e$ at any given value of K, where ρ is the density of proteins and V is the resolution voxel size. The address-value barcoding system is exponential *in log space*. **(B)** The total number of distinguishable phenotypes that can be generated using 8 fluorophores is shown as a function of the number of cassettes that are used. Evidently, the labeling strategy achieves its maximum value for a large number of cassettes, yet the absolute number of barcodes generated is enormous even for many fewer cassettes.

Practical implementations of the address-value barcodes

In practice, given 4-value cassettes (two inversion units), it would be possible to barcode the entire brain of any animal with at most 32 cassettes. With 8-value cassettes, at most 16 cassettes would be necessary. Hence, in the case either of 4-value or 8-value cassettes, 7 fluorophores should be sufficient to barcode the brain of any animal, with either 2 fluorophores dedicated to values and 5 dedicated to cassettes (for 4-value cassettes), or 3 fluorophores dedicated to values and 4 dedicated to cassettes (for 8-value cassettes).

In some cases, it may be useful to have an additional "constant" fluorophore to indicate the presence of the label proteins, irrespective of particular addresses or values. It could also be useful to reserve one or more fluorescent colors for dense chemical (non-genetically-encoded) staining of the cell membrane. In any case, $K < 10$ orthogonal fluorophores should suffice for most applications, and we will quickly become limited by the available orthogonal recombinase sites for achieving a sufficient number C of cassettes, as well as by the ability of each cassette to generate a sufficiently uniform distribution (see Appendix 1, Chapter 13) over a large number of possible values.

Address-Value Barcoding for the Zebrafish Connectome

Overview

We now present two address-value barcoding strategies for the larval Zebrafish connectome which can be implemented using only $K = 6$ fluorescent colors and epitope/antibody pairs (which is readily achievable). The strategies differ in the tradeoff they make between the number of orthogonal recombination sites and the number of values encoded per cassette. If 4-value cassettes are used (e.g., Figure 7-2), 9 such cassettes would achieve 262000 labels, resulting in 68% of neurons having unique labels and an expected degeneracy of $\langle M \rangle = 1.38$ (see Appendix 4, Chapter 13). Alternatively, with 10 cassettes, one could generate $> 10^6$ barcodes, resulting in 91% of neurons being uniquely labeled and an expected degeneracy of $\langle M \rangle = 1.1$. On the other hand, if 8-value cassettes are used, only 6 such cassettes are necessary (e.g., 3 orthogonal Cre/Lox sites and 3 orthogonal Flp/Frt), although this would yield only 118000 effective barcodes under the assumption of 2.8 bits of entropy per cassette, resulting in 43% of neurons being uniquely labeled with an expected degeneracy of $\langle M \rangle = 1.85$. Under the assumption of $100\ \mu m$ axons and one axon per neuron, we crudely estimate (Appendix 4, Chapter 13) that even a degeneracy of $\langle M \rangle = 2$ would only lead to on the order of 10 axon-tracing errors in the Zebrafish brain using tracing algorithms similar to today's automated EM segmentation algorithms (*299*), as compared with tens of thousands of errors in the absence of barcoding[3].

Cassettes

The proteins produced by the cassettes could either be single scaffolds displaying multiple epitopes (e.g. (*281*)), or fusions of multiple scaffolds, each encoding a single epitope. In one possible design, the address epitopes could be displayed on one scaffold coded for by the region directly following the cassette. This scaffold could be connected by a floppy linker to a second scaffold displaying the value epitopes, coded for by the inversion units.

Readout

Regardless of whether one uses 4-value cassettes or 8-value cassettes, 6 fluorophores and epitope/antibody pairs are sufficient to read out the barcode in this approach. Following recombination, the cassettes produce protein scaffolds displaying some combination of the six epitopes. In the case of 4-value cassettes (Figure 7-2), 2 epitopes would indicate the result of recombination while 4 epitopes would indicate the cassette identity; in the case of 8-value

[3] The automated algorithm of (*299*) makes roughly 1 error in every 29 micron segment (see Appendix 4, Chapter 13). So, assuming $100\ \mu m$ long axons, only roughly 5% of axons would be traced correctly in the absence of barcode-based error-correction (2 standard deviations, assuming $3 \pm \sqrt{3}$ errors per axon), i.e., $>$ 95000 errors. Put another way, with no barcoding, 95% of axons would be traced incorrectly, whereas with the proposed barcoding scheme only 0.01% of axons would be traced incorrectly, assuming the underlying morphological tracing error rate was equivalent, even with $\langle M \rangle = 2$ barcode degeneracy.

cassettes, 3 epitopes would indicate the result of recombination while 3 epitopes would indicate cassette identity. Imaging would be performed with confocal microscopy following primary and secondary antibody stains against these epitopes, with a different fluorophore corresponding to each epitope.

Roadmap

Zebrafish/Drosophila

In order to implement the Zebrafish barcoding scheme discussed here, it would be necessary to insert either nine 4-value cassettes or six 8-value cassettes into the Zebrafish and demonstrate orthogonal recombination. This would require the use of 6 fluorophores, which is readily achievable. In addition, the barcoding scheme discussed here would require the demonstration of protein scaffolds that can display up to 6 orthogonal epitopes, all of which can be simultaneously imaged using antibody staining. We believe that these demonstrations should be possible with modest modifications of existing technology. In Appendix 6, we propose an alternative strategy that can be implemented with fewer modifications of existing technology.

Neither of these cases is likely to generate absolutely unique barcodes across all 100k Drosophila or Zebrafish neurons: some barcodes will be found twice in the brain. Appendix 4 (Chapter 13) discusses how even such imperfect barcoding can be used to substantially error-correct automated morphological tracing algorithms.

Mouse

For barcoding at the level of the 10^8 neurons in the mouse brain, an address-value scheme with 6 fluorophores, 14 orthogonal recombination sites and 4 values per cassette would yield 262 million barcodes, sufficient for tracing in the mouse brain. Alternatively, a system with 7 fluorophores, 10 orthogonal recombination sites and 8 values per cassette would yield 282 million barcodes (under the assumption of 2.8 bits per cassette).

Primate

Extending the address-value scheme for the mouse to 14 orthogonal recombination cassettes should be sufficient to barcode primate brains. We discuss in Appendix 2 (Chapter 13) some potential schemes by which 16 effectively orthogonal recombination sites could be achieved through rational design.

Conclusion

We proposed to harness the multicolor nature and high spatial resolution (up to 20smaller than the diffraction limit) of expansion microscopy to permit high-density labeling of neurons with cell-identifying sets of barcode-encoding peptides. We have described how the combination of powerful emerging optical and genetic technologies – multiple orthogonal recombinases, multiple useful epitopes and orthogonal color channels, and amplification of fluorescent signals from single label

protein molecules – could be used to generate a huge diversity of "color codes" for individual neurons.

If all neurons in a brain can be labeled uniquely, as is possible in many of the proposed schemes, then we have converted connectomics from a pure morphological tracing problem (with error rates growing exponentially with tracing length) into a distance-independent barcoding problem, much as would be the case in proposed in-situ nucleic acid barcoding strategies (*277*). Even if the barcode diversity is only comparable to, or even slightly less than, the number of neurons in a brain, we still enable a powerful form of error-correction that is not possible in greyscale electron microscopy, with the potential to hugely reduce the error rates of automated tracing algorithms (Appendix 4, Chapter 13).

Notably, unlike in in-situ sequencing or sequential FISH barcoding approaches, only one round of labeling and imaging is required, so long as we can achieve roughly 7-10 orthogonal color channels. If fewer orthogonal optical channels can be achieved, we can resort to sequential readout approaches to implement the same ideas. This could be done with sequential immuno-staining if enough epitopes are available, and we have also suggested how RNA FISH or RNA FISSEQ could implement similar notions of splitting barcode information over multiple molecules or structures in a region of interest. The RNA FISH case could also be used in a single-wash setting if there are enough orthogonal color channels, effectively using the RNA as a hybridization scaffold rather than its derived protein as an epitope scaffold.

Optical connectomics, of the type described here, appears to require significant yet not prohibitive resources. Following 20x expansion, a $1\ \mathrm{mm}^3$ piece of brain would be expanded to $8000\ \mathrm{mm}^3 = 8 \times 10^{12}\ \mu\mathrm{m}^3$. A confocal microscope with a 60x objective and a detector with an area of $1\ \mathrm{cm}^2$ has a field of view corresponding to $166\ \mu\mathrm{m} \times 166\ \mu\mathrm{m} \times 0.7\ \mu\mathrm{m} = 20000\ \mu\mathrm{m}^3$. Assuming $200\ \mathrm{ms}$ per field of view when imaging in ≈ 10 color channels, we find that 2.54 microscope-years would be required per cubic millimeter of pre-expansion brain. Amortizing $1M microscopes over a three-year project, the raw connectomic data acquisition for even the smallest whole mammal brains would thus cost on the order of $10M to $20M using this scheme.

Notably, with unique or near-unique multicolor labeling of neurons, it is likely that the image analysis problem for axon tracing could become accessible to simple automated algorithms, with multicolor error correction potentially enabling extremely high accuracy. Multicolor optical approaches could enable error-resilient automated tracing approaches through fluorescent barcoding and spatial multiplexing, as well as extensions to molecularly annotated connectomics [17] through temporal multiplexing.

Chapter 8
A New Structure for Scalable Research

Foreword

The ideas contained in this chapter are derived in large part from conversations held in the first 4 months of 2019 with Ed Boyden, Joi Ito, Louis Kang, Jessica Traynor, Daniel Oran, and Laura Deming, with additional feedback from Karl Ruping and Robert Hughes. I also acknowledge Adam Marblestone for reviewing the final essay and offering useful feedback.

Summary

The ordinary lifecycle of a technology begins with the inception of an idea, runs through the creation of a proof of concept, and ends with creation of a product, usually in a for-profit venture. However, there are some projects that could have widespread impact if they were scaled up, but that are not ripe for traditional for-profit investment. Examples include the early development of the integrated circuit and the human genome project, neither of which had clear commercial applications at the time of their inception, but both of which required substantial, focused research and development over ~10 years to generate a commercializable technology. Projects such as these cannot be pursued efficiently either in academia or in a for-profit setting. For these projects, a new research structure is required that funds hyper-focused projects at a high level for a strictly limited period of time, and that emphasizes specific metrics, collaboration, and scale. I lay out the case for this kind of research non-profit, termed a focused research organization (FRO), and propose a systematic program to identify promising targets for and to fund FROs.

Introduction

Academia excels at producing new discoveries and novel ideas, but the vast majority of new ideas and technologies are never reproduced or achieve large-scale impact. In part, this is because academic culture prizes individual recognition (as articulated most clearly by (*300*)): projects that require larger coalitions of researchers are typically unpalatable to academics. However, larger coalitions are vital for the development of scientific ideas (*301*), and it has been noted by the draft roadmap for the second phase of the BRAIN initiative that "dissemination [of new technologies] to the research community will be critical to the BRAIN Initiative's success" (*302*). The question of how to incentivize researchers to participate in larger coalitions and work towards long-term goals is still open.

Traditional for-profit companies can scale scientific or technological approaches and achieve widespread societal and technological impact when the impact is profitable. However, there is a large class of basic research projects, such as the Human Genome Project or the early development of integrated circuits, that likewise need to be scaled in order to achieve impact, but that will only realize their value on timescales beyond the horizons of typical venture investment, or that depend for their impact on the results being public. The Human Genome Project was funded publicly and the data was made available publicly, while the development of semiconductor technology was funded in large part by industrial labs, such as Bell Labs. Modern examples include the mapping of the mammalian connectome (*61*, *303*, *304*), or several other recently proposed neurotechnology-oriented projects (*6*); the translation of nanofabrication technologies from basic materials research to commercial applications (*305*); and the development of approaches for carbon capture or geoengineering (*306*). These projects require large amounts of funding, a focused and specific approach, and the contributions of many individuals with diverse skill sets, but do not lend themselves to funding through traditional academic or for-profit mechanisms.

Here, I argue that in order to support these projects efficiently, a new structure for basic research and development is necessary. I propose that these projects could be pursued in a dedicated structure, termed a focused research organization (FRO), in which the incentives are specifically aligned to enable highly-powered individuals to work together towards to achievement of specific technological goals. These organizations would develop the key IP necessary for company formation and then spin out for-profit companies, recouping the cost of development through the revenues or equity thus generated.

An Example: Scaling Connectomics

As an example of a problem that cannot be scaled either in an academic or a traditional for-profit setting, I will focus on connectomics. As discussed in Chapter 7, the connectome is the set of all connections between neurons in the brain. The cost of the microscopes alone that would be necessary to gather the raw data for the mouse connectome would likely be $5M-$50M, depending on the method used, so the funding scale is beyond what can be mustered in academia. Collecting

the connectome will require applying a single approach to a very large volume, so it will require a specific and focused approach. Moreover, regardless of the method one uses to gather the data, mapping the connectome will require a full-stack approach including basic biology and animal handling, tissue processing, and computation, so it will be necessary to coordinate the efforts of many individuals with disparate skill sets. Finally, the value of the connectome will likely only be realized once the connectome is mapped, so it is not yet possible to lay out a business model or propose a specific therapeutic target as would likely be necessary to establish a for-profit business. Thus, the connectome seems to fit well into the category of projects I define above, which do not lend themselves naturally either to the academic or for-profit models. Indeed, it has previously been recognized that connectomics cannot be scaled within the context of academia, and an ambitious IARPA-funded project termed MICrONS was established in 2016 with the goal of mapping $1\ mm^3$ (0.2%) of the mouse brain (e.g. (*307*)). The MICrONS project achieved some but not all of the objectives I lay out here, and I will discuss below its successes and failures.

Academia Incentivizes Novelty, not Focus

Academics are funded primarily by federal grants, on the basis of work published in academic journals through a process of peer review. Innovation is one of the key criteria by which grant applications and academic publications are judged. To fulfill the innovation criterion, academics are incentivized to distinguish themselves from others and to pursue novel research. In the academic sciences, novelty comes in the form of new discoveries, whereas in academic engineering it comes in the form of proofs of concept.

This emphasis on novelty leads to a constant stream of new ideas coming out of academia. In addition, the freedom to pursue novel ideas and achieve recognition for creativity is used to incentivize academic researchers to work for low salaries. However, at the same time, the emphasis on novelty and the incentive to distinguish oneself from others discourages collaboration and sustained development of ideas once those ideas are published. In particular, this incentive structure leads to three major limitations[4]:

1) **Accountability:** There is little incentive in academia to produce work of higher quality than what is necessary for publication in a journal, which is typically much lower than the quality that is necessary to achieve impact. This is a fundamental source of the reproducibility crisis: results do not need to be reproducible and techniques do not need to be widely adoptable in order to be published, de-emphasizing quality.

[4] I wish to distinguish the critiques I level here from prior critiques, including critiques of the inability of academic funders to fund high risk/high reward research, or the general failings of the academic publishing system. Although the conservative nature of funding organizations is certainly a challenge, I will critique a different aspect of academia, which is the inability of academic labs to achieve goals that require a coherent dedication of large amounts of resources.

2) **Transparency:** Because the currency in academia is recognition, academia is plagued by large first-mover effects, and as a result, academics are incentivized not to share their progress with others. This leads to enormous inefficiencies as many labs often pursue the same work in parallel. Moreover, researchers are unable to learn from the failures of others' experiments.
3) **Collaboration:** Because of the strong emphasis on novelty and individual achievement, it is very challenging to incentivize graduate students to collaborate with each other within a lab, and it is challenging to incentivize labs to work together. Having many authors on a paper dilutes individual recognition, which can be fatal for young and unestablished scientists.

Together, the pursuit of novelty prevents academics from focusing their resources on specific problems in the way that would be necessary to accomplish large-scale research objectives. For example, in the case of the connectome, one could imagine that a project would start as a collaboration between three labs, one doing tissue handling, one focusing on microscopy, and one focusing on computation. However, the tissue handling group could quickly conclude that it would be able to publish a first paper on their barcoding method without the other labs, and would spend a year studying schizophrenia in order to collect the scientific discoveries necessary for that paper. Their paper, a promising technological advance with a great example of a scientific application, might come out in *Science* two years after the project began. Meanwhile, the microscopy lab would likely dedicate their resources to inventing a new kind of microscope rather than optimizing existing microscopes for the purpose, even if the existing microscopes would be technically superior. Their efforts would be rewarded with a paper in *Nature Methods*. Finally, the graduate student in the computational lab working on image processing might lose confidence that she would ultimately get first-authorship credit for her contribution, and switch to a different project analyzing cryo-EM data of the autophagosome. The project would lose institutional knowledge and expertise, but she would rapidly gain two or three papers at *ICML*. In this scenario, each lab does extremely well, but after two years, they are no further towards the connectome than they were at the start of the funding cycle.

The problem was well-articulated in a recent article about the need for a national network of neurotechnology centers to scale methods like brain activity mapping and connectomics. The authors wrote, "Many of these essential operations [e.g. microscopy, computation] may not be perceived, in isolation, as sufficiently cutting-edge to be fundable. Further, many will also be inappropriate for graduate or post- doctoral researchers; instead, to ensure their reliable execution, these activities could be better carried out by professional scientists and engineers. Yet it is generally impossible to sustain skilled and experienced technical personnel through short-term single-investigator funding" (*6*).

Academia occupies a critical point in the innovation cycle, as the primary source of new ideas, but it is unreasonable to expect that it can align the incentives of multiple disparate researchers and labs to pursue complex, highly-coordinated objectives over an extended time period.

Focused Research Organizations

For these projects, that lie between academia and traditional for-profit ventures, I propose the establishment of focused research organizations (FROs). These FROs would be small, well-funded teams, incorporated with specific research objectives in mind. They would have dedicated space and full-time scientific staff, and would be funded at a high level compared to typical academic projects. They would be led by one or more principal investigators, who would be specific to the project: to prevent dilution of attention and to avoid the funds being diverted to fund traditional academic projects, the leaders of the projects would be prohibited from having active academic appointments (although faculty could serve in an advisory role). The operation of these FROs would have three defining characteristics:

Metric-driven: Impact depends more on quality than it depends on novelty. For that reason, FROs should be driven by metrics: the amount of brain tissue that can be processed in a given amount of time, for example, or the number of assays that reproducibly confirm the same hypothesis. These metrics would be determined at the outset of the project (e.g. included in the proposal to form the FRO), and would be used by a board of directors to monitor the progress of the FRO over the funding period. Depending on the formal structure of the FRO, the metrics would be used to insulate the FRO from traditional academic incentives or from market forces, and bonuses could conceivably be tied to the attainment of metrics.

A corollary of being metric-driven, rather than novelty-driven, is that the FRO will be able to align its incentives with academics easily. Collaboration with academics will lead to the introduction of new ideas and insights into the FRO's pipeline, allowing it to improve its progress towards its metrics, while the academics can leverage the data generated by the entity to publish novel scientific findings. The FRO has no incentive to keep its progress secret except as necessary to protect IP: once IP is filed, the FRO should communicate (and indeed, might be required to communicate) with for-profits and academics as much as possible to solicit feedback or insight. For example, whereas getting scooped can be fatal in an academic setting, just as being second to market can be fatal for a for-profit, the FRO has no competitors since any progress towards the final metric goals is considered success.

Team-oriented: Scaling projects requires alignment of the efforts of many talented individuals. This is not possible in an academic setting, because academics are driven by authorship status on papers, and are intrinsically disincentivized from collaborating in large groups. For this reason, it is important that the incentive structure for individuals be closer to the structure found in a company than to that found in academia. Individuals should receive monetary compensation

comparable to that of for-profit companies. In addition, in lieu of the equity they would receive at a for-profit, they should receive equity in the for-profit companies that spin out. Finally, they should have opportunities to move into for-profit spin-outs as a way of affording them career advancement opportunities.

Limited in scope: Tenured academic labs and for-profit companies are both unlimited in their scopes – they can continue to operate as long as they are productive or profitable, respectively. By contrast, the goal of the FRO is to achieve some key metrics necessary to transfer technology to a for-profit. For that reason, it should be strictly limited in scope to avoid mission creep. If the metrics were not met (or at least reasonably approached) at the end of a well-defined period, the project and technology should be reevaluated. On the other hand, if the metrics were achieved, the expertise obtained in the process should transfer into the for-profit companies that spin out.

As an example, a reasonable FRO in the biotech space might require $15M in direct costs over 3 years, sufficient to hire 10-15 people at an average salary of $150,000/yr, with sufficient reagents and capital equipment expenditures.

Comparable Efforts

Non-Profit Research Organizations

Large, highly-focused basic research projects have historically been funded on an individual basis. Very large-scale, highly-coordinated efforts are commonplace in the experimental physical sciences. The LIGO gravitational wave observatory was constructed with a total cost of $620M, and the Large Hadron Collider had a total cost of $13.25bn as of the discovery of the Higgs Boson, with an operating budget of roughly $1bn/yr (*308*, *309*). In the biological sciences, large-scale efforts are much less common, although the Human Genome project was funded with an initial investment of $3.8bn (*6*), and more recently, the ARMI regenerative medicine initiative (also a non-profit) has raised more than $270M.

However, not every effort needs to be funded at such a large scale, and there is ample evidence (especially in biology) that projects funded in the range of $5M-$20M can also achieve transformative results. Several research institutes, such as the Allen Institute for Brain Research, the Howard Hughes Medical Institute, and the Broad Institute, have made outsized contributions to the fields of biology and neuroscience through the establishment of small, highly focused research efforts. The Allen institute provided the neuroscience community with a compendium of in-situ hybridization assays for the great majority of genes in the genome, which has proven to be transformative for research, for a total cost on the order of $50M (*38*, *310*, *311*). It is now focused on obtaining the first cubic millimeter of densely mapped connectome. Janelia has deliberately established project teams with the goal of transforming proofs of concept into workable tools, and has produced the Neuropixels electrophysiology device for a cost of ~$5-10M (*32*), and the gCaMP molecular calcium indicator for an unknown total cost (*312*), both of which are proving

revolutionary for neuroscience. In the area of genomics, the Broad Institute has established several cell atlasing efforts, which are likewise possible due to its relatively corporate structure (with a high ratio of career to academic staff) and provision of core flow sorting and sequencing facilities. These projects have mostly shared the criteria outlined above: strictly limited scope, a focus on the team rather than on individual attainment, and a focus on metrics rather than on novelty. And, notably, most of these projects likely could not have been accomplished in a traditional academic or for-profit setting: as noted by the BRAIN 2.0 Working Group's draft report in 2019, "making 1,000 units [of Neuropixels] exceeds $2 million, well out of the reach of most academic labs," and "highlights a necessary departure from standard business models for dissemination of lab-use neuroscience tools."

Despite these major successes, however, most institutes (including the Broad and Janelia) remain dominated by the academic research model. As noted in a retrospective, Janelia has found that "without an opposing force provided by management, there is a slow, steady drift toward a more conventional environment increasingly focused on maintaining successful programs and documenting individual achievement at the expense of risk taking and collaborative, interdisciplinary work" (*313*). Systematically counteracting this drift and maintaining a culture of high-impact, goal-driven, risk-taking work may require regular disassembly of the research apparatus,

For-Profit Research Organizations

Many industrial labs exist that could also establish the kinds of projects described here. Historically, Bell Labs and Xerox-Parc are the most famous examples, but modern examples include Google X and DeepMind, both owned by Alphabet. However, development of technology within an industrial lab can stymie innovation on the whole, since the IP may not be made readily available for further development, which was the case with Bell Labs before the 1956 consent decree (*314*). Moreover, most industrial research labs have become more focused in the past two decades, working primarily on core product development (*315*).

In addition, for-profit companies (or investors) face challenges associated with the limited lifetime of patents and the time-value of money. Patents in the US are limited to 20 years: even if the genome had been deemed patentable, the patents would now be nearly expired, just as genetic medicine is beginning to come of age. On the other hand, the time-value of money (i.e., the opportunity cost of investing one's money in a research project that is unlikely to generate substantial returns in the short term) was identified in the BRAIN 2.0 whitepaper as one of the major factors working against for-profit development of neurotechnologies (*302*), since it inflates the opportunity cost of projects with long development cycles relative to similarly-priced, shorter projects. This makes it particularly hard for for-profit companies to justify investing in the kinds of projects described here.

Government Programs

DARPA and IARPA have attempted to achieve the goals outlined here by issuing grants (or contracts) focused on *deliverables*, rather than on papers or novelty. The MICrONS project, for example, is a connectomic initiative funded by IARPA in 2016, with $100M divided across three teams for 5 years, and its results support the hypothesis that a dedicated research support structure is necessary for scaling academia effectively. Of the projects, a $18.7M grant to the Allen Institute appears to have succeeded in scaling a connectomic imaging technology to the level needed to achieve the $1\,\mathrm{mm}^3$ reconstruction goal. By contrast, a similarly-sized grant to a collaboration between Carnegie Mellon, Cold Spring Harbor, Harvard, and MIT, has resulted in multiple publications, but the publications from different institutions are largely unrelated to each other (e.g. (*286*, *316*) and a forthcoming publication from the Boyden lab) and generally include authors from only one of the participating institutions, suggesting a failure to establish a highly focused, team-oriented research culture. After 3 years of research, the progress of a third, $28M project at Harvard is unclear. I conclude from this that an existing, non-academic structure for research is likely necessary for large grants to be successful.

Moreover, I propose that although the focus on *deliverables* is preferable to a focus on novelty, a focus on metrics (such as the *rate* of reconstruction, rather than the total reconstructed volume) would have been even more preferable. For example, IARPA would certainly prefer that the Allen Institute project work for 5 years on systems and technology improvements and finally produce a tenth of a cubic millimeter in one day's work, implying the ability to scale to the whole brain in 5000 days (or fewer), than it would for the Allen Institute to apply an existing technology over 5 years to produce a single cubic millimeter, implying the ability to scale to the whole brain in 2,500 years.

Implementation

Implementing the program described here has two challenges: finding a sustainable structure for the research organization, and finding a sustainable funding mechanism.

Structure

Non-profit or for-profit

The specific charter according to which FROs are established will determine their incentive structure and susceptibility to corrupting forces, such as market forces or recognition incentives. As described above, I am specifically interested here in the set of projects for which a traditional for-profit funding model will not work, either because the time needed is beyond the time horizon of a typical venture fund, or because there are is no clear profit model. Nonetheless, the question remains as to whether FROs could be established within a for-profit setting.

If the FRO is established with a for-profit charter, then the charter must be established in a way that will allow the FRO to pursue its founding objectives (i.e. the metrics) without being

distracted by market forces or the whims of investors in the short term. The question, when leveraging a for-profit model for basic science research, is the degree to which the basic research objectives can be aligned with the profit-making incentives of the investors. If they cannot be aligned, the for-profit will find itself under pressure to pivot to more easily achievable goals with clear, marketable products, much in the same way that academics would find themselves under pressure to pivot to lower-hanging paper opportunities. In the specific case of the connectome, one could easily imagine that a for-profit would develop the automated microscopy and tissue handling systems necessary to map the mouse brain at scale, but would then get distracted marketing those systems or selling connectomes of small volumes as a service, rather than mapping the entire brain of a single mouse, which is necessary to construct the complete connectome.

In rare cases, it may be possible to align basic research incentives with the incentives of for-profits. Celera was founded in with the goal of sequencing and then patenting the genome. It received $300M in funding and succeeded in completing a draft of the genome at the same time as the publicly funded, $3bn genome project (although direct comparisons are unfair, since Celera relied on much of the technology developed by the publicly funded project) (*6*, *317–320*). Indeed, if the connectome could be patented, there is no doubt we would be able to raise substantial private funding to acquire it, and likewise for many other large-scale scientific endeavors. Unfortunately, it was announced in March 2000 by Bill Clinton and Tony Blair that the sequences of human genes would not be patentable, and the Supreme Court affirmed in 2013 that products of nature, such as gene sequences, are not patentable, ruling out a Celera model for funding FROs.

Several intermediates between for- and non-profits exist, such as low-profit LLCs, which are for-profit companies structured in a way that allows charitable foundations to donate money to them while also receiving a return (*321*). The directors of social purpose corporations have a fiduciary responsibility to a social purpose set forth in the articles of incorporation. Nonetheless, in both cases, these models are intended for organizations that do not otherwise qualify as non-profits, for example because they are in direct competition with for-profits (*322*). Neither option circumvents the fundamental issue that the FROs by assumption lack a clear plan for making money in the short term, before the metrics are achieved.

By contrast, the non-profit model provides several concrete advantages. Chief among these is the ability to collaborate directly with academics; collaboration between for-profits and academic institutions is often fraught with conflict-of-interest due to restrictions on for-profits benefiting from public grant funding. In addition, non-profits are free and encouraged to share information about their approach, and have more flexibility than for-profits in defining their overarching incentives. Not surprisingly, most existing examples of FRO-like projects (see “Comparable Efforts,” above) took place in a non-profit context.

More creative intermediates also exist. For example (see also "Value Capture," below): 25% of the shares of Novo Nordisk and 75% of its voting shares are owned by the Novo Nordisk Foundation, a non-profit that is funded through the profits of the Novo Nordisk for-profit pharmaceutical company. In this way, Novo Nordisk is still able to attract private investment through the 75% of shares that are owned by other shareholders, but a large percentage of its profits are reinvested in biomedical research through the foundation (*323*). In a case in which a non-profit FRO would spin off for-profit companies that it would partially own (see "Value Capture," below), for-profit investors could be induced to invest in the non-profit by a guarantee to be able to invest in the for-profits down the line, perhaps at a predetermined valuation.

Similar tradeoffs are on show in the case of OpenAI. OpenAI began as a non-profit corporation, inspired by the goal of achieving full transparency. However, it switched to a "capped-profit" corporation in 2019 as a way of attracting capital. Investors may now make money off of it, up to a 100x return, after which any remaining returns will go to an overarching non-profit (*324*). However, this model is unlikely to apply directly to the FRO case, because it is predicated on a clear, existing business model.

Umbrella Organizations

One of the core concepts advocated here is that these focused research organizations should be limited in scope, being completely disassembled at the end of a predetermined time. Assembly and disassembly of a research apparatus incurs immense overhead, so it is not possible for endowed institutions like the Broad Institute or Janelia to completely disassemble and reassemble themselves on a regular basis. However, institutes such as these could serve as hosts for FROs, providing them with lab space and administrative support, to reduce the costs of initiating the projects. Alternatively, one could imagine the creation of a dedicated umbrella organization or a government program with the specific goal of initiating and supporting FROs. The umbrella organization would not conduct research itself, but would provide space, administrative support, and possibly funding for FROs, allowing for the systematic renewal of the research mission without the overhead associated with starting a new institute. Within the structure provided by the umbrella organization, projects could remain focused on big ideas with measurable outcomes, and would be free to fail fast.

Funding

Funding through gifts

Unless an existing institution wants to commit itself to trialing the program described here, initial funding will need to come from philanthropists or donations from for-profits that believe in the necessity of changing the academic research model. Donations from companies or venture funds could be in exchange for a guarantee to be able to license (non-exclusively) the intellectual property generated by the effort, similarly to the funding structure for the MIT Media Lab.

Eventually, I hope that the federal government would recognize that in order to derive maximum value from the academic work it invests in, it should invest in maturing the most promising projects to the level of commercialization. The government could start a project to initiate e.g. 100 projects per year, each at a funding level of $5M/yr for 3 years with no (or strictly limited) options for renewal. This would only require an annual budget of $1.5bn, a small fraction of the total federal research budget. Federal grants to these projects would be clearly differentiated from academic grants: for example, the presence of a clear, quantitative, and ambitious final metric would replace innovation as a central criterion for evaluation of proposals, and in contrast to many large federal grants, all participating researchers would be required to be primarily employed by the same research organization to ensure opportunities for close coordination. Projects that required additional funding to achieve maximum impact could subsequently seek out philanthropic funding on the basis of their achievements. There is a clear interest in funding these projects among federal agencies. The ARMI institute, a non-profit, raised $80M in defense funding for work on regenerative medicine (*325*).

Value Capture

There is ample evidence that academia fails to capture the value it creates. For example, companies funded by MIT graduates have $1.9tn in annual revenues; but revenues attributable to MIT IP are only roughly $2bn/yr, and MIT only captures roughly 1%-2% of that value (*326*, *327*). There may be many reasons for this disparity, such as the limited lifetime of patents (*328*). Regardless, only 0.2% of the revenues of companies founded by MIT graduates would suffice to completely fund MIT's operating budget, including replacing tuition and federal research funding. Non-profit FROs could actively seek to found for-profit companies, taking either a small percentage (e.g. 0.5%) of the revenues of the for-profit or a large percentage of its shares. These returns would almost certainly be realized long after the FRO has disbanded, but they could be returned to the umbrella organization to fund the creation of future FROs. A similar model has been adopted by Novo Nordisk, as described above. This conclusion mirrors a recent conclusion from the Brookings Institution that most university technology transfer offices consume more resources than they produce, and that universities should focus less on licensing technology and more on founding startups (*329*).

Conclusion:

Given the considerations above, I propose the creation of a new, non-profit organization with the specific goal of establishing the research facilities and the funding pipelines necessary to initiate and support metric-driven, team-oriented FROs. The FROs should be funded at a total level of $5M/year for 3 or 4 years, after which they should terminate, to allow for a ground-up reevaluation of their approach and goals. In addition to the board of directors of the overarching organization, each FRO should have its own board of directors, responsible for monitoring progress towards the metrics and disbursing funding. The FROs should each have their own COO

with experience from industry, to ensure that they function more as focused research efforts than as open-ended academic ventures. The umbrella organization should take responsibility for initiating FROs, and should focus on spinning off companies based on the research of the FROs. It should maintain a large stake in companies spun off in this way, as a way of ensuring its future existence and tying its future growth to the growth of its spinoffs. Along with capturing more of the potential value in academic research, this system would have additional beneficial effects for scientific culture. It would open a new pathway into basic science research for graduate students, postdoctoral researchers, or researchers from the private sector with good ideas and management experience but few high-impact publications.

Chapter 9
Supplementary Information to Chapter 3

Materials and Methods

Overview

Our standard workflow, elaborated upon below, consists of gel synthesis, followed by incubation in a patterning solution, typically a solution of fluorescein and a hydroxide in water. Subsequently, the gel was patterned using 780nm excitation on a 2-photon microscope. Following patterning, the patterning solution was removed, and different reagents (depending on the experiment) were deposited in the patterned locations. In the case of the silver patterning, gold nanoparticles thus anchored to the gel could then be grown by aqueous silver intensification, using the LI silver chemistry. Finally, the gels were shrunken by exposure to solutions of HCl or divalent cations and possibly dehydrated. For some experiments, different patterning reagents, deposition reagents, or shrinking processes were used, as described below. The experimental procedure for each figure is summarized in Table 9-3.

Throughout, all washes were performed on an orbital shaker at 80RPM except during the shrinking and dehydration steps.

Gel Synthesis

Gels were synthesized as described elsewhere (*37*). In short, the monomer solutions are mixed from stock solutions of 10x PBS, 5M NaCl, 38% (w/w) sodium acrylate, 50% (w/w) acrylamide, and 2% (w/w) N,N'-methylenebisacrylamide in concentrations given in Table 9-1 and Table 9-2, for the 10x gel and 20x gel monomer solutions respectively. Solutions were aliquoted and stored at -20 ˚C. Prior to casting, the monomer solutions were kept at to 4 ˚C to prevent premature gelation. Concentrated stocks of ammonium persulfate (10% w/w) and tetramethylethylenediamine (TEMED) (10% v/v) were diluted 50x into the monomer solutions. The resulting gelation solution was then mixed thoroughly and added to a gel mold that was ~0.17 mm tall and ~1 cm wide. Molds consisted of a glass slide for the bottom and a No. 1.5 coverslip for the top, using two additional coverslips as spacers. The mold was placed at 37 ˚C for 1.5 hours to allow for gelation. Following gel synthesis, the gel was washed twice in ~2-3 million times its initial volume in water for 30 minutes to ensure full expansion.

Preparation for Patterning:

Following expansion, expanded gels were cut into 2cm squares and transferred into a glass-bottom dish (Mattek, P50G-1.5-30-F) and incubated in 2ml patterning solution (below) twice for 30 minutes each time. Except where otherwise indicated (Figure 3-1M,Figure 3-2D,Figure 9-4A), we used the 10x gel solution. Following incubation, a 40mm diameter coverslip (Fisher Scientific 22-038-999) was placed over the well of the glass-bottom plate with the gels inside and excess

patterning solution was withdrawn, in this configuration the coverslip pressed the gel against the bottom of the plate helping to reduce sample drift and slowing evaporation.

For direct deposition of streptavidin into the gel, as in Figure 3-1J, Figure 9-1D, and Figure 9-3, the patterning solution consisted of 333μM biotin-4-fluorescein (Biotium Cat. 90062) and 1.25mM rubidium hydroxide (Sigma, 402393-25G).

For depositing NHS-activated fluorophores or reagents, such as biotin-NHS (Sigma, H1759), as in Figure 3-1B,D,F,H,I,K,L, Figure 9-2D (red bar), Figure 3-3, Figure 3-4, Figure 9-1A-C, Figure 9-2, Figure 9-5 and Figure 9-6, the patterning solution consisted of 500μM 5-aminomethyl fluorescein hydrochloride (Life Technologies, A-1353) and 2mM sodium hydroxide in water.

For depositing with maleimide-activated fluorophores and nanoparticles into the gel, as in Figure 3-2B-H and Figure 9-7, the patterning solution was made by reacting fluorescein-NHS (Life Technologies, 46409) to cysteamine (Sigma, M9768-5G) at 1mM concentration in water for at least 30 minutes prior to incubation.

Patterning:

Gels were patterned using an inverted Zeiss LSM 710 confocal microscope with a Chameleon Ultra II femtosecond pulsed IR laser set to 780nm, using a 40x 1.1NA water immersion objective. Within the Zen software, custom ROIs were defined for acquisition. The surface of the gel was identified by a decrease in fluorescence relative to the external patterning solution. Standard patterning conditions were 0.79μs pixel dwell time and a pixel size of 350nm, amounting to a patterning speed of 44cm/s, in pre-shrink dimensions. Unless stated otherwise, all patterns were generated using 2x line scanning. For Z-stacks, a 2μm step size was used.

Laser power varied depending on the intensity of patterning desired. The optimal laser power for patterning depends strongly on the laser collimation, objective, gel composition and patterning solution composition. However, because fluorescein retains some of its fluorescence upon attachment to the gel, it is possible to optimize the patterning power quickly, by patterning rectangular prisms with different powers (as in Figure 9-2B,D). In this case, the patterns will begin to bulge outwards as the power increases, and one typically wants to choose the highest power at which bulging is not evident. It is important that this calibration be performed using patterns with similar depth to those that will ultimately be patterned, because the degree of patterning when patterning several adjacent layers in the axial dimension will in general be greater than the degree of patterning when patterning a single layer, because the patterning voxels

from successive layers may overlap. Unless stated otherwise, we used 128mW laser power, as measured using a power sensor (Thor Labs, S170C) in the image plane.

For patterns in Figure 3-1B,D,F,H,K,M, Figure 3-2B-H except D (red bar), Figure 3-3, and Figure 3-4A-C a Z-stack exposure was taken starting 10µm below the gel interface continuing 50µm inside of the gel to ensure that the patterns were at the surface of the gel for SEM visualization performed at the end of the process.

For patterns in Figure 3-1I,J,L, Figure 3-2D (red bar), Figure 3-4D-F, Figure 9-1A-C, Figure 9-2, and Figure 9-5, Z-stacks were performed starting 50µm inside the gel. Figure 3-1I,J,L, Figure 3-2D (red bar), Figure 9-1A-C, Figure 9-2 and Figure 9-5 were done using Z-stacks that extend 50µm further into the gel.

For the patterns in Figure 3-1I and Figure 9-2B,D, the laser powers are as follows, from left to right, in mW. Top row: 52, 60, 68, 76. Second row: 84, 91, 99, 107. Third row: 114, 121, 128, 136. Fourth row: 143, 149, 155, 161.

For the patterns in Figure 3-2B-D except Figure 3-2D (red bar), each line was scanned either once or twice using the 40x objective, with variable laser power. The condition was indicated by tick marks above and to the left of the triangles, as follows: 1 tick mark, 12.5% laser power with 1x line scanning. 2 tick marks, 12.5% laser power with 2x line scanning. 3 tick marks, 17.7% laser power with 1x line scanning. 4 tick marks, 17.7% laser power with 2x line scanning. 5 tick marks, 25% laser power with 1x line scanning. For patterns in Figure 3-2E-H, we used 17.7% laser power with 2x line scanning. To ensure that the patterns were at the surface of the gel for SEM visualization, patterns in Figure 3-2 except Figure 3-2D (red bar) were generated as Z stacks with 2µm step size beginning below the surface of the gel and extending 50µm into the gel.

For Figure 9-7, we used 25% laser power with 0.39µs pixel dwell time, with a 25x glycerol immersion objective.

Deposition:

We applied a specific and complementary chemistry for deposition depending on the reactive group patterned into the gel. Following patterning, the gels were washed four times in water for fifteen minutes each time to remove excess patterning solution.

For depositing fluoronanogold-streptavidin (nanoprobes #7416, hereafter referred to as fluoronanogold) onto patterns of 5-aminomethyl fluorescein as in Figure 3-1D,F,H,I,L, Figure 3-3, Figure 3-4, Figure 9-2A-D, Figure 9-5, and Figure 9-6 the gel patterns were first stained with biotin-NHS (Sigma, H1759). To do this the gels were washed in 1x PBS for 15 minutes before

performing the conjugation with 100µM biotin-NHS in 1x PBS for three hours. Subsequently, biotin-NHS was washed out three times in water for 30 minutes. Then, gels were washed once in 1xPBS and positioned in the middle of the Mattek glass well, to prevent the gel or the fluoronanogold solution from coming into contact with the plastic rim of the dish. Fluoronanogold was diluted 30x to 2.7µg/ml into 300µL of 1x PBS and placed on top of gel. The samples were then left to stain for twelve hours on a shaker at room temperature in the dark. Fluoronanogold was then washed out four times in 0.1x PBS for an hour each time before two additional 10 minute washes in water.

For depositing Atto 647N-NHS onto patterns of 5-aminomethyl fluorescein, as in Figure 3-1K,M, gels were washed twice in 1x PBS for 15 minutes each time. Subsequently, Atto 647N-NHS (Sigma,18373-1mg-F) was diluted to 50µM concentration in 1x PBS and washed onto the gel for at least 4 hours. Because Atto 647N is positively charged and tends to partition into the negatively charged gel, gels were then washed twice in 200mM NaOH for at least 30 minutes each time, followed by three washes in 1x PBS for 30 minutes each time, followed by three washes in water for 15 minutes. By contrast, after staining aminomethyl fluorescein with a negatively charged dye, excess dye could simply be washed out in water.

For depositing DNA onto patterns of 5-aminomethyl fluorescein, as in Figure 9-1A-C, gels were functionalized with biotin NHS at 1mM concentration in 1x PBS overnight, followed by three washes in water and two more washes in 1xPBS to remove excess reagent and prepare for the streptavidin deposition. Atto 647N-labeled streptavidin (Sigma, 94149-1mg) was then washed onto the gel at 40µg/ml in 1x PBS with 3% Bovine Serum Albumin overnight. The gel was then washed in 2.5mM Tris-HCl, pH 8, three times for at least 1 hour each time to remove excess streptavidin. DNA could then be deposited within streptavidin-functionalized gels by washing the gels in a solution with 10µg/mL biotinylated DNA in 1x PBS for 3 hours. DNA was subsequently removed by washing in water 3 times, for at least 15 minutes each time.

For depositing maleimide-activated gold nanoparticles into patterns of fluorescein-cysteamine, as in Figure 3-2E,F,G,H, gels were washed twice in 1x PBS for 15 minutes each time. Subsequently, maleimide-functionalized 1.4nm gold nanoparticles (Nanoprobes, 2020A) were diluted to 5µM concentration in 1x PBS and washed onto the gel overnight. Gels were then washed twice in water for at least 30 minutes each time, transferred to a new container, and washed in water three more times for at least 30 minutes each time to remove excess gold.

For depositing maleimide-conjugated fluorophores onto patterns of fluorescein-cysteamine, as in Figure 9-7, gels were washed twice in 1x PBS for 15 minutes each time. Subsequently, maleimide-functionalized dyes were washed into the gel in PBS at 100µM concentration, and left to stain overnight. Gels were then washed three times in water, for at least 30 minutes each time.

Intensification:

Following deposition of fluoronanogold, gels were transferred to a 35mm diameter petri dish (Corning 353001). The gels were then washed in 50mM EDTA pH 5.5 for 30 minutes. Gels were then immersed in 2mL LI silver solution (Nanoprobes #2013) and placed in a shaking incubator at 20 ˚ C and 80 rpm for a variable amount of time, as described below. To halt intensification, gels were washed briefly in water once ~1-2 minutes and then three more times for 10 minutes. Remaining silver ions in the gel were removed prior to shrinking by washing in 50mM sodium citrate for one hour. Subsequently, the gel was washed four times in water for 10 minutes each time.

For a given batch of samples, we determined the intensification time necessary to achieve the optimal density of silver by performing intensification on test samples for each of 40, 45, 50, 55, and 60 minutes. These test samples were then shrunken according to the protocols below, dehydrated, and imaged on a Zeiss Ultra Plus or Supra55 FESEM. Samples that were grown for too long would show bulging at the edges of the patterns as a result of steric hindrance during the shrinking process. Thus, the optimal intensification time for the batch was determined as the maximum growth time that did not lead to visible distortion in the SEM images. The remaining samples in the batch were then intensified for the optimal amount of time. Although there was significant batch-to-batch variability in the amount of intensification time necessary to achieve high-quality metallized patterns, the within-batch variability was found to be small, and this process robustly generated well-metallized patterns without distortion.

Following intensification of the remaining samples for the optimal growth time, samples could be imaged as in Fig. 1F on a Nikon TI microscope with brightfield illumination, using an Orca Flash 4.2 camera set to 16 bit gain 1/4, a 0.5NA condenser, and a Nikon Plan Fluor 20x objective.

Shrinking:

For Figure 3-1K,M, Figure 3-2, with the exception of Figure 3-2D (red bar), and Figure 9-4, gels were shrunken by washing first in 2mM HCl, followed by 20mM HCl and 200mM HCl, all in a glass chamber. Subsequent experiments determined that the 20mM and 200mM HCl washes were unnecessary to achieve full shrinking.

For gels in Figure 3-1L,H, Figure 3-3, Figure 3-4A-C, Figure 9-2E, Figure 9-3, Figure 9-5B, and Figure 9-6, gels were shrunken using acid by transferring to a glass container and washing in 2mM HCl with 0.05% Tween-20 for 6 hours and again for another hour. Finally, gels were washed in 2mM HCl for 30 minutes to remove residual Tween-20. Liquid was then removed and gels were left out in open air until completely dry, typically for 2 hours.

For Figure 3-4F and Figure 9-5A, the gel was shrunken, but not dehydrated, by washing in 2mM HCl with 0.05% Tween-20 before imaging.

For Figure 9-1A-C and Figure 9-7 the gel was shrunken by washing 3x in 10x PBS for 15 minutes followed by washing in 1M MgCl2 3x for 15 min, and these gels were not dehydrated prior to imaging.

Sintering:

To ensure conductivity of the silver structures, sintering was performed using the same microscope, laser, and objective used for patterning. First, dehydrated samples were mounted on carbon tape, such that only the edge of the substrate was attached to the tape, and placed face down on a Mattek dish. This allowed patterns to be located on the microscope using transmission illumination. The samples were then imaged and brought into focus using 1.5mW 2-photon illumination intensity, with excitation at 780nm.

For samples in Figure 3-3 and Figure 3-4B, sintering was then performed by capturing a single image of the field of view containing the pattern with a power of 15mW, using the same objective, pixel size and dwell time as used for patterning. The data in Figure 9-6B represents a mixture of samples for which sintering was performed with 15mW or 20mW exposures. As the difference between the two groups was not found to be statistically significant, the data from the two groups was lumped to improve the utility of the regression.

Imaging:

For Figure 3-1D,I,K, Figure 3-4D,E,F, and Figure 9-2C,D, samples were imaged on a Zeiss LSM710 with a 32x 0.8NA water immersion objective in either fluorescence confocal mode, or reflection confocal mode in the case of Figure 3-4E. The image in Figure 9-5A was obtained on the same microscope with a 40x 1.1NA water immersion objective.

The post-shrink measurements of samples in Figure 3-1L and Figure 3-2D (red bar) were obtained using the LSM710 with a 63x 1.4NA oil immersion objective with the sample immersed in oil, to minimize optical aberrations.

The post-shrink image in Figure 9-5B was obtained using the LSM710 with a 40x 1.3NA oil immersion objective with the sample immersed in oil to minimize optical aberrations.

For Figure 3-1B, Figure 3-2E, and Figure 9-2A,B multi-photon imaging at 780nm was performed on the Zeiss LSM710, typically while the gels were still in the patterning solution. This imaging was performed using much lower laser power than the power needed for patterning.

For Figure 3-1M and Figure 3-2B-D fluorescence imaging was performed using a Perkin Elmer spinning disk (CSU-10 Yokogawa) confocal microscope. We used a Hamamatsu Orca-ER cooled

CCD camera, and either a 10x 0.5NA objective or a 40x 1.15NA Plan Apo long working distance water-immersion objective (Nikon).

Transmission optical images, including Figure 3-1F and images used for analysis in Figure 9-6B, were taken on a Nikon TI microscope with Koehler illumination, using an Orca Flash 4.2 camera set to 16 bit gain 1/4, a 0.5NA condenser, and a Nikon Plan Fluor 20x objective.

Images for Figure 9-1 and Figure 9-7 were taken on a Nikon TI widefield microscope, using an Orca Flash 4.2 camera and a variety of objectives.

Scanning electron microscope images of the AuNP patterns (Figure 3-2F-H) were taken using a FE-SEM (UltraPlus, Zeiss) with an Energy selective Backscatter (EsB) detector. Images from Figure 3-1H, Figure 3-3B,C, and Figure 3-4A,B were taken using the same FE-SEM (UltraPlus, Zeiss) with the SE2 detector. The atomic force microscopy (AFM) of the gel surface in Fig. S3B was taken with tapping mode in air (Cypher ES, Asylum Research) with a silicon probe (AC240TS, Olympus). Images for Figure 3-3A and Figure 3-4C were taken on a Zeiss FE-SEM (Supra), with an SE2 detector.

Analysis:

Figure 3-1L,M: Data for the lateral shrink measurements in Figure 3-1L,M was obtained by comparing the feature sizes of patterns as specified on the patterning microscope to the size of patterns after shrinking and dehydration. Samples were chosen on the basis of the availability of high-resolution optical or SEM images of the shrunken state, and came from a variety of different experiments. The axial shrink amount for 10x gel was deduced by patterning a cross consisting of 354µm long lines of 14µm thickness and 300µm depth. The height of the shrunken pattern was then measured on a confocal microscope and compared with the patterning dimensions to determine the amount of shrink in the axial dimension.

The calculation of estimated binding sites patterned using our process was done using data from Figure 9-2A as follows. The concentration of 5-aminomethyl fluorescein used to incubate the gels is known to be 500µM. We measured the fluorescence both inside and outside and used this ratio to deduce that the internal concentration is 300µM. Then using what was known to be the brightest pattern in Fig. S2A and S2B we calculated based on the difference in fluorescence in the pattern from the background that the concentration must be greater than 79.2µM. We say greater than because the patterning process bleaches an unknown fraction of the fluorescein molecules. Thus, any measurement we make is likely lower than the actual values for sites patterned. To calculate the final concentration of 277.2mM after shrinking we simply multiplied by the volumetric shrink factor demonstrated in Figure 3-1L.

Figure 3-2D: Isotropy was measured for samples into which circles had been patterned. Yellow and blue bars: bar graphs of the lateral isotropy of shrink for six 10x gels, and four 20x gels. Lateral

isotropy was defined as the ratio of the longest axis of patterned circles (C, inset) to the shortest axis, in the shrunken and dehydrated state. The isotropy was measured by visually determining the longest axis of the circle, and comparing the diameter on that axis to the diameter on the orthogonal axis. A mixture of gels patterned with aminomethyl fluorescein and fluorescein-cysteamine were used. Gels were chosen for inclusion in the dataset on the basis of the availability of images for analysis, prior to measuring the isotropy. No gels were excluded. Dots are measurements for individual circles within a single gel; bars indicate mean + standard deviation across individual circles within a single gel. Bars are rank ordered from left to right by degree of anisotropy, for each shrink factor. Red bar: The axial isotropy for six 10x gels is shown in red. For axial isotropy, we a produced a pattern with a "+" cross-section extending 300µm axially. Analogously to the lateral isotropy measurements, then, the axial isotropy for a given pattern was then defined as max(S/S', S'/S), where S is the ratio of the axial to lateral shrink factors, and S' is the ratio of the mean axial to mean lateral shrink factors. For the axial isotropy data, dots represent single measurements made on six different gels, and bar indicate mean + standard deviation across gels.

Figure 3-2G,H: The widths of lines visualized with SEM were measured by using ImageJ to rotate the image so that the lines were oriented vertically, and then taking the mean pixel value over the vertical dimension for a clean segment of line. The average was performed over the longest clean segment of line available in the image, usually several hundred pixels. The full width at half maximum (FWHM) was then measured in pixels, and converted into a distance using the scale bar provided by the SEM imaging software. The baseline used in the FWHM measurement was found by linear interpolation between the baseline levels immediately on either side of the line profile (Figure 3-2G). A vertical line was drawn between the highest point in the profile and the interpolated baseline, and the midpoint of this line was chosen as the half-maximum. Lines were excluded from our analysis when the magnitude of the background (for example due to charging) prevented a determination of the FWHM. In addition, a subset of the lines in the resolution pattern were excluded in every gel due to a consistent and reproducible error in the Zen software that caused an extra line to be patterned directly below those lines, leading to a larger FWHM. We reasoned that these lines could be excluded, because they represent a limitation of the software rather than a limitation of the patterning and shrinking process.

Figure 3-3D,E, Figure 9-6A: For all conductive samples, conductivity was measured using a four-point probe setup with a semiconductor parameter analyzer. The parameter analyzer was set to measure the voltage and current at an electrode (V1, I1) placed on one side of the sample, and to measure the voltage at two other electrodes, one (V2) placed adjacent to the first electrode, and the other (V3) placed adjacent to a ground electrode. Voltage measurements were performed for many different values of V1, typically spanning a range between 1mV and 100mV. The measurements were occasionally noisy due to poor vibrational isolation. For this reason, in all

cases, we calculated the conductivity as follows. The total resistance Rtot was determined by linear regression of V1 against I1. In addition, we regressed V3 against V1 to obtain R0/Rtot, and V2 against V1 to obtain (R0 + Rsample)/Rtot, where Rsample is the resistance of the sample, and R0 is the contact resistance at the ground electrode. We then obtained the resistance Rsample algebraically. For samples on which the measurements were clean, the values of Rsample calculated in this way aligned closely to the values obtained simply by regressing V4-V2 against I1. However, we found that our method was also capable of calculating the resistance in the presence of significant vibrations.

Figure 9-6B: We measured the opacity of silverized patterns in the expanded state following silver intensification using a transmission light microscope with Koehler illumination. Intensified silver patterns appeared dark on the transmission microscope due to absorption by the silver patterns. We calculated the opacity by measuring the average intensity O outside the pattern and the average intensity I inside the pattern, and then defined the opacity as 1 – I/O, where I is the light intensity passing through the metallized region and O is the light passing outside the metallized region.

Multimaterial Patterning:

For multimaterial patterning as in Figure 3-1J and Figure 9-3 the patterning solution consisted of 333μM biotin-4-fluorescein (Biotium Cat. 90062) and 1.25mM rubidium hydroxide (Sigma, 402393-25G). The solution was washed into a fully expanded 10x gel for 30min prior to each round of patterning.

To pattern the gel with biotin-4-fluorescein we used 255mW laser power, with a 40x 1.1NA objective, and 1x line scanning. Patterns were generated as Z stacks with 2μm step size beginning below the surface of the gel and extending 100μm into the gel. Gels were then washed four times in water for 20 minutes each time following patterning to remove excess patterning solution. Then, gels were washed once in 1x PBS for 20 minutes, after which Alexa 488-labeled streptavidin (for Figure 9-3) (Thermofisher, S11223), was diluted to a concentration of 33μg/ml in 1xPBS, added to the gel and left to stain for 12 hours; or fluoronanogold (for Figure 3-1J) was diluted into PBS, added to the gel and left to stain for 12 hours. The gel was then washed in 0.1x PBS three times for two hours each time and then twice in water for 20 minutes, to remove excess streptavidin conjugates.

Subsequently, the gels were immersed again in the biotin-4-fluorescein solution, and patterned for a second time as above. Excess fluorescein was removed by washing and streptavidin conjugation proceeded identically to the first round. For Figure 3-1J, the second round of deposition used 33μg/ml Qdot655 Streptavidin (Thermofisher, Q10151MP), while for Figure 9-3 the second round of patterning used Atto 647N-labeled streptavidin (Sigma, 94149-1mg). After washout the gels for

both Figure 3-1J and Figure 9-3 were shrunken but not dehydrated in 1xPBS by washing once for 30min before imaging with a Zeiss LSM710 and a 40x 1.1NA objective.

In a multimaterial experiment of this type, some reagent from the second round of staining may be deposited on reactive groups patterned during the first round of patterning. To determine the magnitude of this "cross-talk," we relied on the data collected in Figure 9-3A. The magnitude of the background-subtracted Atto 647N fluorescence signal at the first patterning location was found to be 18.5% of the magnitude at the second patterning location. Because some of that signal may be due to spectral overlap of the B4F or Alexa 488 fluorophores with the Atto 647N fluorophore, this places an upper bound on the amount of cross-talk associated with this multimaterial protocol at 18.5%. A similar measurement in which the two regions were overlapping (Figure 9-3B) yielded 21% cross-talk.

Rehydration and hybridization to DNA gels:

For the experiment in Figure 9-1D, samples were treated identically to those in Figure 9-3, up through the Alexa 488-streptavidin stain. Subsequently, gels were washed in 1x PBS and incubated for 3 hours with biotinylated DNA carrying an Atto 565 dye at 10μg/ml. The sequence of the DNA target was GATTATCCGTGACACAGTAGACTA, and the fluorophore was on the 3' side. Subsequently, the gel was washed in water three times, for 20 minutes each time. It was then placed in 50mM sodium citrate. The gel was then transferred to a solution of 5mM citric acid, and washed twice with this solution, for 30 minutes each time. It was then put in 2mM HCl and 0.05% tween for 1 hour, rinsed in 2mM HCl without tween, and then dehydrated. The gel was then imaged on a widefield epifluorescence microscope with a 20x objective. Imaging confirmed the presence of the DNA in the gel at this point.

The gel was then rehydrated by washing in PBS twice for 30 minutes each time. Subsequently, the gel was incubated in a solution of a probe DNA oligo carrying an Atto 647N dye at 10μg/ml. The sequence of the probe DNA was CTACTGTGTCACGGATAATT, and the fluorophore was on the 5' side. The gel was then washed twice in PBS for 30 minutes each time, and was then imaged. Imaging at this step confirmed that the 647-labeled probe DNA oligo was in the gel at the same location as the target. Moreover, we observed a substantial reduction in the fluorescence of the 565-labeled target oligo, which we attribute to quenching of the Atto 565 fluorophore by the probe DNA oligo, possibly by FRET.

To confirm that the probe oligo was attached to the target oligo by DNA hybridization, we subsequently immersed the gel in 200mM NaOH for 2 hours. We then washed once in PBS and

imaged again. We observed a large reduction in signal in the 647 channel, and a recovery of signal in the 565 channel, consistent with a loss of the probe DNA oligo.

Figures

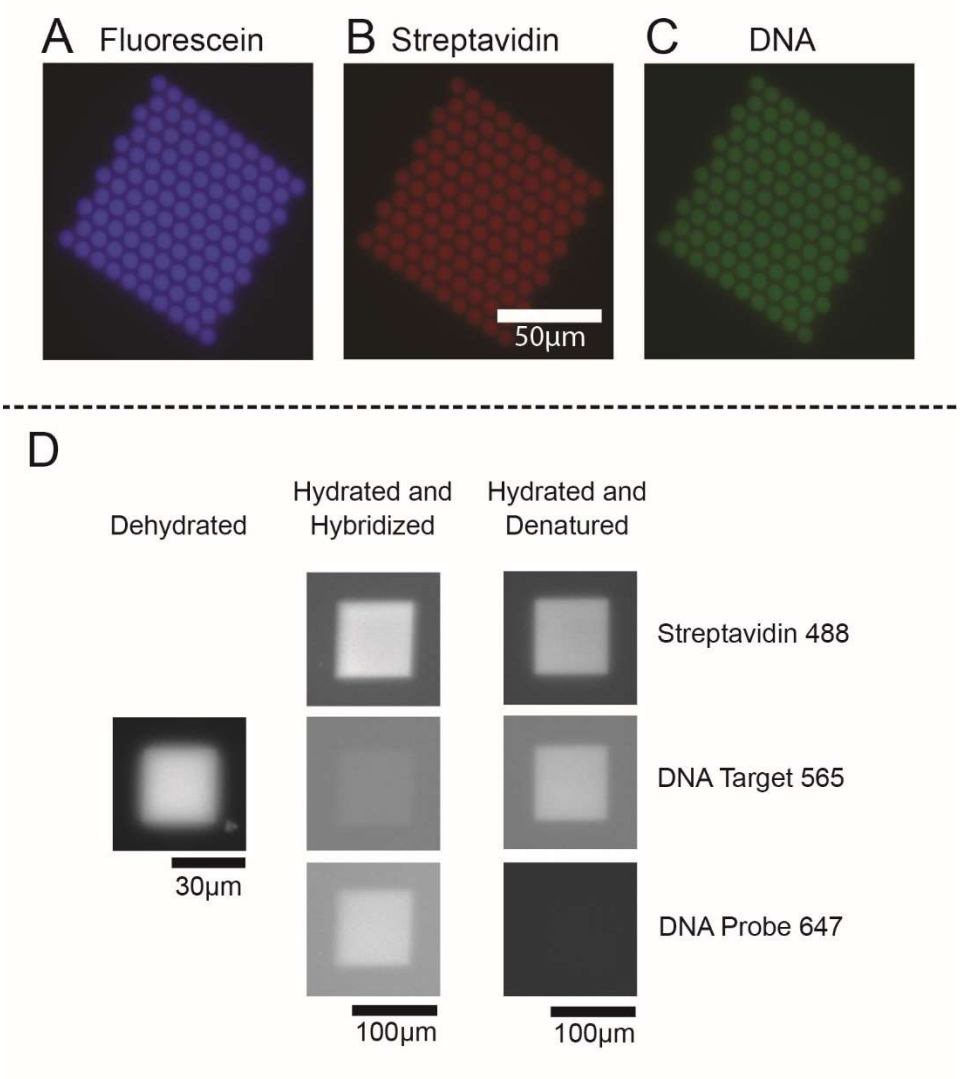

Figure 9-1: Material conjugations of various kinds, shown after shrink but not dehydration, and imaged with epifluorescent microscopy. **(A)** Image of fluorescein patterned into the gel in a defined, "microarray"-type pattern. Scale bar on (B). **(B)** Image of fluorescently labeled streptavidin deposited in the same sample. **(C)** Image of fluorescently labeled DNA deposited in the same sample. Scale bar on (B). **(D)** To test whether DNA and streptavidin survive the HCl shrinking and dehydration protocol, a DNA oligo functionalized with biotin and Atto 565 was attached to Alexa 488-labeled streptavidin. Subsequently, the gel was shrunken with HCl, dehydrated, and imaged (left panel, showing DNA present in the dehydrated state). It was then rehydrated, and a complimentary "probe" oligo labeled with Atto 647N was washed into the gel (center). We attributed the decrease in the fluorescence of the Atto 565 signal at this stage to quenching of the Atto 565 fluorophore by the probe DNA oligo, possibly by FRET. The gel was subsequently washed in 200mM NaOH to denature the hybridized DNA and imaged (right), confirming a loss of the 647 signal and a recovery of the 565 signal.

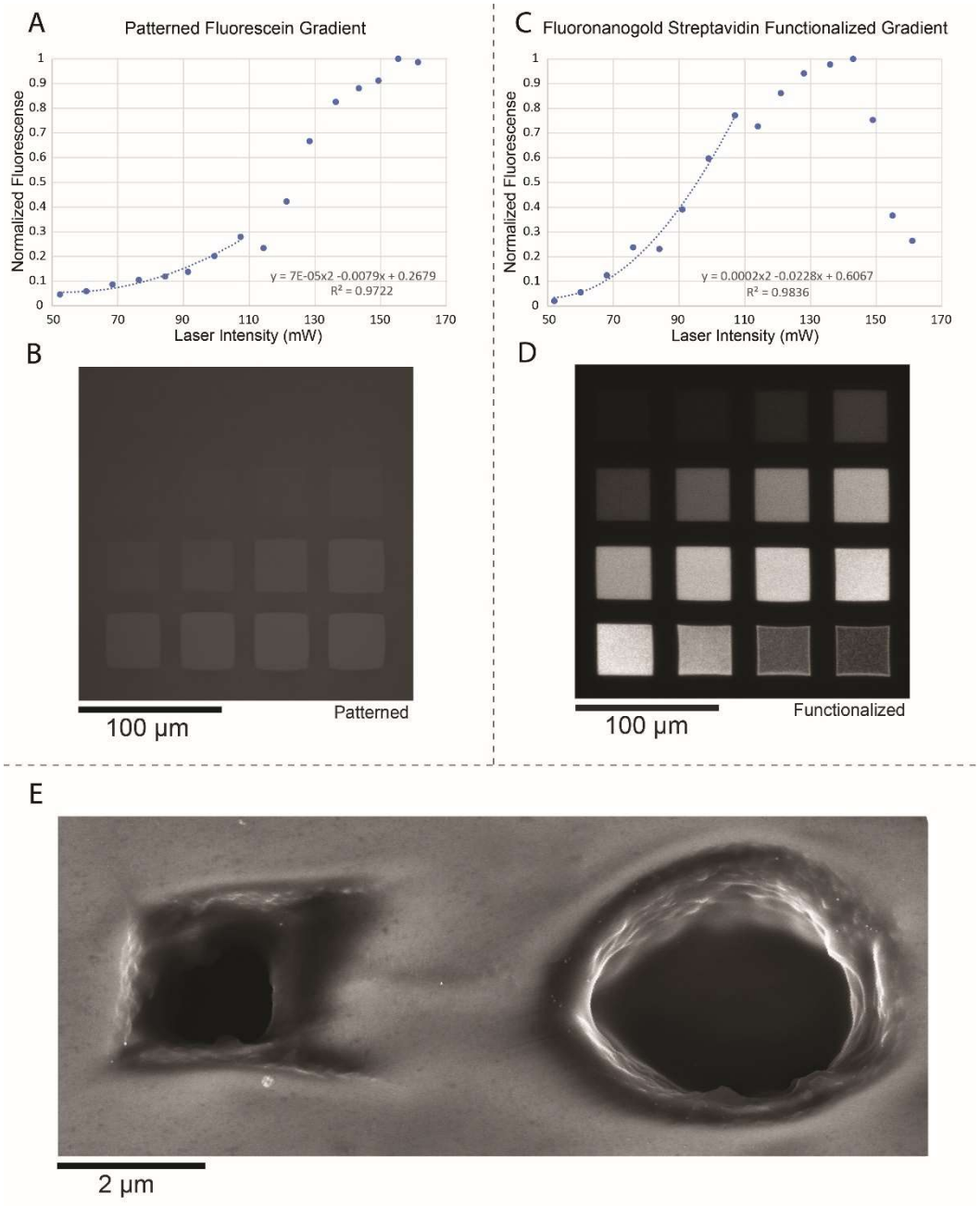

Figure 9-2 Sixteen squares were patterned into a single gel, with each square being patterned with a different laser power. Gels were imaged immediately after patterning, prior to washing the patterning solution out of the gel, and were subsequently functionalized with fluorescent streptavidin. For laser powers below a critical threshold, the density of the deposited material is approximately quadratic in the laser power used. At higher powers, the density of deposited material shows an inversion and the patterns bulge inwards, coinciding with ablation of the gel substrate, although other processes such changes in the solubility of the gel or the fluorescein due to laser heating may play a role. **(A)** The average intensity of bound fluorescein at each square is shown as a function of the laser power used in patterning. A quadratic fit is shown for powers less than 110mW. A quadratic dependence of the fluorescence of bound fluorescein on laser power is expected, because the rate of two-photon excitation depends quadratically on the laser intensity. **(B)** The raw two-photon image of the squares is shown, powers increase from left to right and top

to bottom (see Methods). Note that bulging of the squares while they are in the patterning solution appears to correspond to the regime in which the patterned intensity no longer increases with increasing power. Also note that the intensity of fluorescein in the patterned region does not decrease with increasing laser power, unlike in the case of the deposited material (D). If the inversion phenomenon is due to gel ablation, the lack of inversion in the fluorescein signal could be explained by fluorescein partitioning out of the gel, into the void left by the ablation. **(C)** The average intensity of conjugated streptavidin is shown as a function of the laser power used in patterning. A quadratic fit is shown for powers less than 110mW. **(D)** The raw confocal image of the squares is shown, powers increase from left to right and top to bottom. Note that contraction of the squares following deposition appears to correspond to the inversion region. **(E)** SEM image of 20x shrunken and dehydrated gel, showing ablation of the gel substrate corresponding to a patterned square and circle upon the use of excessive laser powers. In the course of developing the current manuscript, we found that gel ablation in this way could be used to generate complex three-dimensional structures, but those structures would not typically survive the dehydration process.

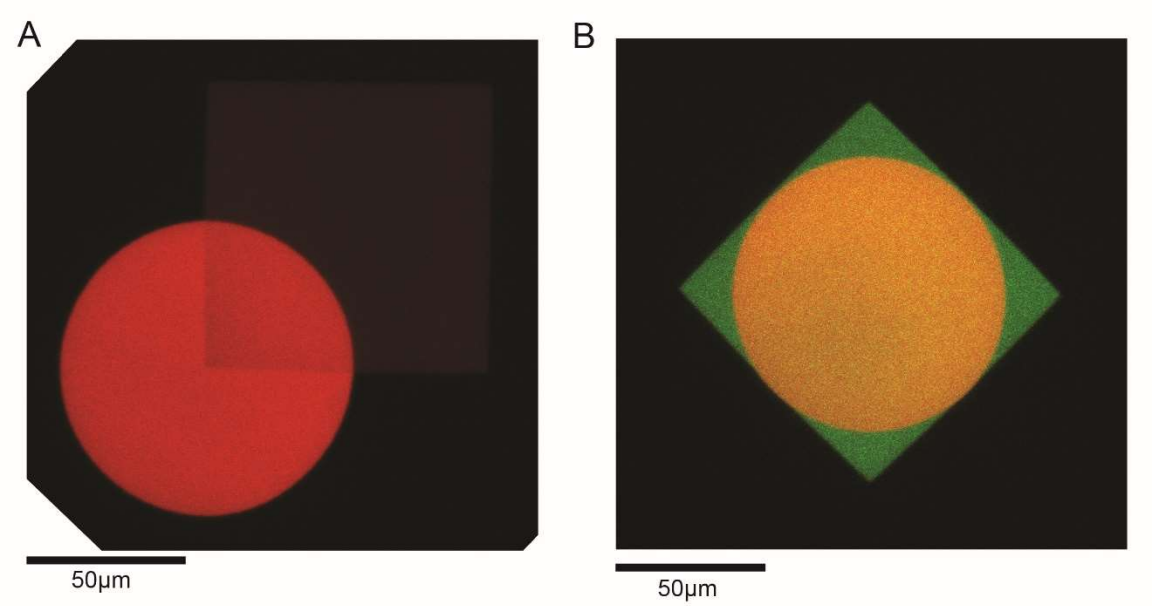

Figure 9-3: (A) Image of fluorescent streptavidin patterned in the second round of a multimaterial patterning experiment. Only the fluorescence associated with the second round deposition is shown. The square (top right) was patterned in the first round, and the circle (bottom left) was patterned in the second round. The intensity of the square pattern is 18.5% of the intensity of the non-overlapping circle pattern, indicating that the crosstalk between patterning rounds is at most 18.5%. **(B)** A similar pattern, showing the channels associated with both the first (green) and second (red) patterning rounds at once. In this case, the cross-talk was measured to be 21%.

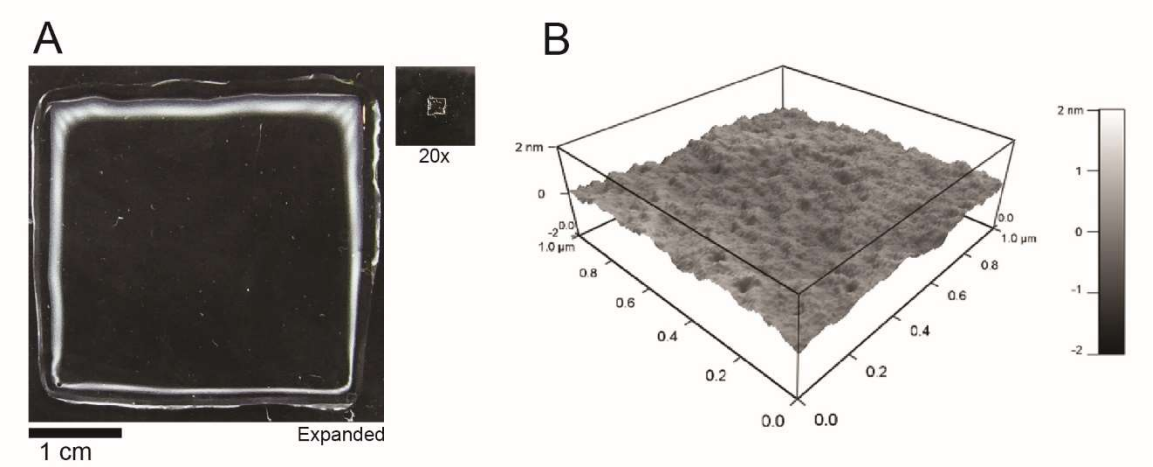

Figure 9-4: (A) Photographic images of a gel before (left) and after (right) shrinking and dehydration (20x gel). **(B)** Atomic force microscopy (AFM) smoothness measurement performed on a 10x shrunken and dehydrated gel, unpatterned, showing surface smoothness in the nanometer range across length ranges of ~1 micron.

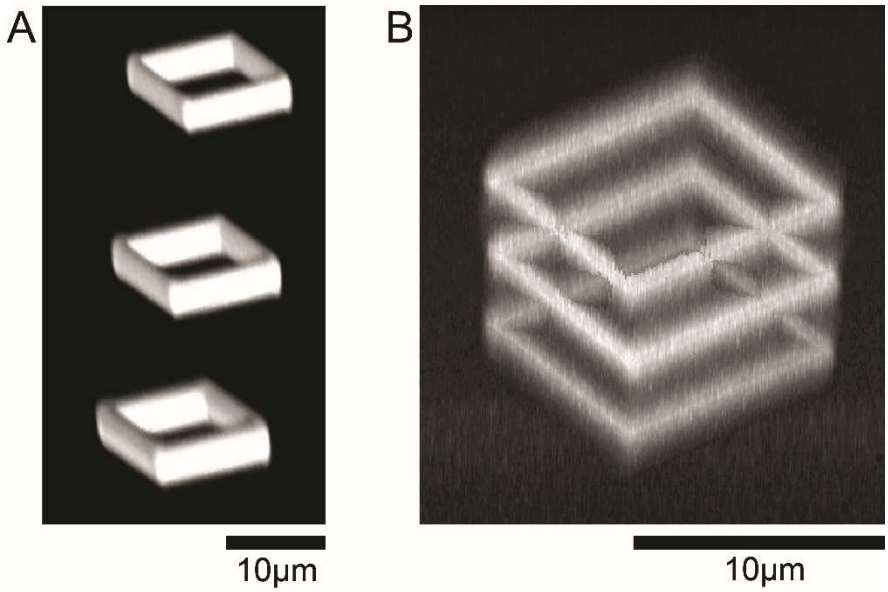

Figure 9-5 (A) Fluorescence image of a rectangular prism, imaged after HCl shrinking but prior to dehydration. **(B)** The same pattern imaged after dehydration, showing additional shrinking in the axial dimension, which causes the rectangular prism to become a cube.

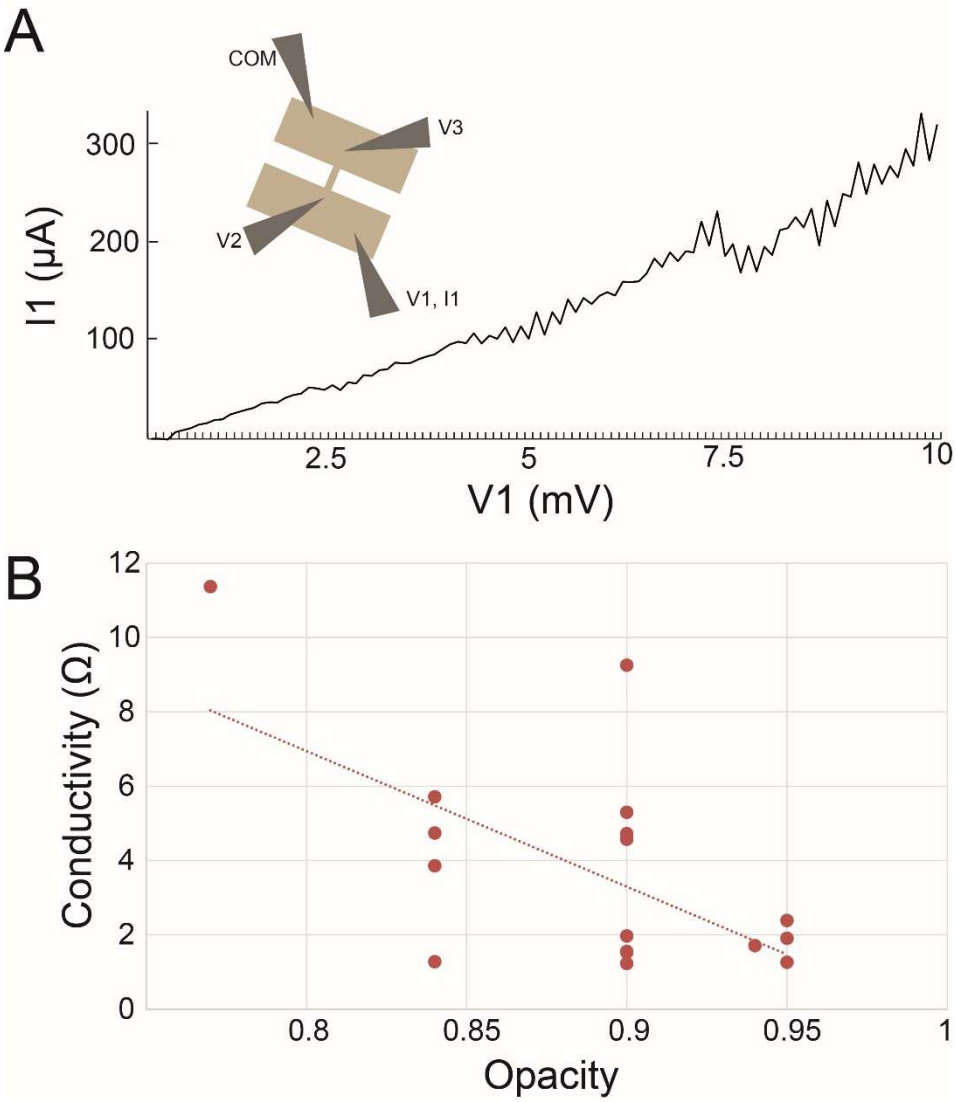

Figure 9-6: (A) Current-voltage (IV) curve shown for one sintered silver wire, determined by a four-point probe measurement as shown in the inset. **(B)** The conductivity of silver wires (N=18, N=8 sintered with 20mW laser power; N=10 sintered with 15mW laser power) as a function of the opacity following intensification, measured as 1 – I/O, where I is the light intensity passing through the metallized region and O is the light passing outside the metallized region. The best fit is shown as a dashed red line, with R^2=0.36 and F=8.99. The linear relationship is significant at the α=0.01 level.

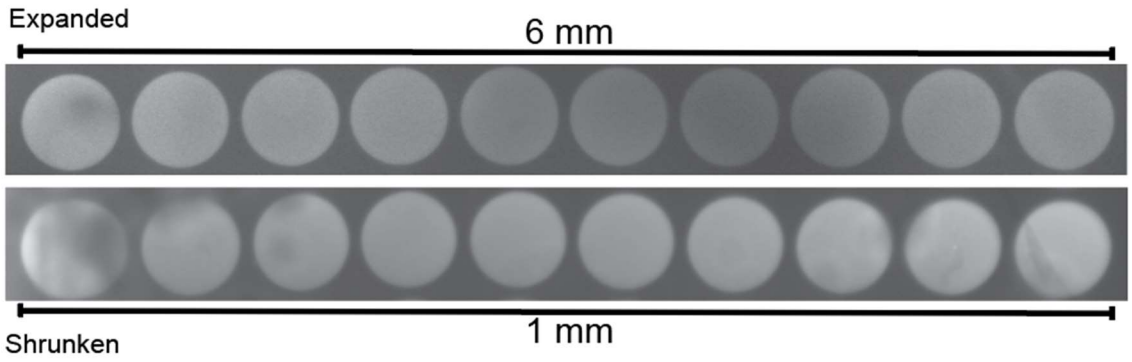

Figure 9-7: A large-area pattern of circles shown in the expanded (top) and shrunken but not dehydrated (bottom) states. In total, the pattern in the shrunken state covered an area of roughly $1mm^2$; a subset of the total pattern is shown here since the pattern was repetitive. This sample was shrunken by a linear factor of 6 in a $MgCl_2$ solution. The differences in brightness observed in the expanded state are due to refraction of the excitation light off the edges of the gel, and are not significant. The inhomogeneities in the shrunken image are defects that arose during handling.

Component	Stock Conc.	Amount (mL)
Sodium Acrylate	38% (w/w)	2.25
Acrylamide	50% (w/w)	0.5
Bisacrylamide	2% (w/w)	0.375
NaCl	5M	4
10x PBS	10x	1
Water		1.475
Final		9.6

Table 9-1 Formulation of the 10x gel mix. To this monomer solution, we would add 200uL of 10% (w/w) APS and 200uL of 10% (v/v) TEMED to initiate polymerization, or 2uL of both APS and TEMED into 96uL of the monomer solution.

Component	Stock Conc.	Amount (mL)
Sodium Acrylate	38% (w/w)	2.25
Acrylamide	50% (w/w)	0.5
Bisacrylamide	2% (w/w)	0.075
NaCl	5M	4
10x PBS	10x	1
Water		0.9
Final		8.725

Table 9-2: Formulation of 20x gel monomer solution. To this monomer solution, we would add 182uL of 10% (w/w) APS and 182uL of 10% TEMED to initiate polymerization, or 2uL of both APS and TEMED into 96uL of the monomer solution. Note that this monomer solution was concocted by accident, so the volume does not sum to 9.6mL as for the 10x gel mix.

Sample	Gel Stock	Patterning Solution	Patterning Parameters	Deposition	Intensification	Shrinking	Dehydrated	Imaging
1B,D,F, H	10x	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	128mW, 2x line scanning	30x dilution of fluoronanogold to 2.7µg/ml in 1x PBS	Yes	2mM HCl 0.05% Tween-20 for 6 hours; 2mM HCl 0.05% Tween-20 for 1 hour; 2mM HCl without tween for 30 minutes.	Yes	**B:** LSM710, 40x 1.1NA water immersion objective, visualizing fluorescein fluorescence with two-photon excitation at 780nm immediately after patterning, while gel is still in patterning solution. **D:** LSM710, 32x 0.8NA objective, single-photon excitation, imaging fluoronanogold following deposition, while gel is expanded. **F:** Transmission optical microscopy on a widefield microscope following intensification, while gel is expanded. **H:** SEM image of gel following shrinking and dehydration.
I	10x	"	Varies, see methods.	30x dilution of fluoronanogold to 2.7µg/ml in 1x PBS	None	None	No	Imaging fluoronanogold in expanded state. LSM 710 single photon excitation; 32x 0.8NA objective.
J	10x	333µM biotin-4-fluorescein and 1.25mM rubidium hydroxide in water	255mW, 1x line scanning	First Round: 30x dilution of fluoronanogold to 2.7µg/ml in 1x PBS. Second round: 33µg/ml Qdot655-streptavidin in 1x PBS.	None	1x PBS	No	Imaged after shrinking in PBS. LSM 710 single photon excitation; 40x 1.1NA water immersion objective. Two channels superimposed.
K	10x, 20x (inset only)	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	128mW, 2x line scanning	50µM Atto 647N in PBS	None	2mM HCl, followed by 20mM HCl, followed by 200mM HCl.	No	LSM710, 32x 0.8NA objective, single-photon excitation, imaging Atto 647N, while gel is expanded or shrunken and dehydrated (insets).
L	10x	"	128mW, 2x line scanning	30x dilution of fluoronanogold to 2.7µg/ml in 1x PBS.	None	2mM HCl 0.05% Tween-20 for 6 hours; 2mM HCl 0.05% Tween-20 for 1 hour; 2mM HCl without	Yes	Images of nanogold fluorescence were taken using a LSM710 with a 63x 1.4NA oil immersion objective with the sample immersed in oil to minimize optical aberrations, with single-photon excitation.

						tween for 30 minutes.		
M	20x	"	128mW, 2x line scanning	None	None	2mM HCl, followed by 20mM HCl, followed by 200mM HCl.	Yes	Images of fluorescein fluorescence were taken using a spinning-disc confocal following shrinking and dehydration.
2B-H	10x, 20x (2D, blue bars only)	1mM Fluorescein-NHS, 1mM cysteamine, in water, prepared 30 minutes prior to use.	Varies, see methods.	5µM Maleimide-activated gold nanoparticles in PBS.	None	2mM HCl, followed by 20mM HCl, followed by 200mM HCl.	Yes	**B,C:** Imaging fluorescein fluorescence using a spinning-disc confocal either in the expanded state, or following shrinking and dehydration. **E:** LSM710, 40x 1.1NA water immersion objective, visualizing fluorescein fluorescence with two-photon excitation at 780nm immediately after patterning, while gel is still in patterning solution. **F:** Imaging gold nanoparticles deposited in the gel using an FE-SEM.
2D (red bar)	10x	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	128mW, 2x line scanning	30x dilution of fluoronanogold to 2.7µg/ml in 1x PBS.	None	2mM HCl 0.05% Tween-20 for 6 hours; 2mM HCl 0.05% Tween-20 for 1 hour; 2mM HCl without tween for 30 minutes.	Yes	Images of nanogold fluorescence were taken using a LSM710 with a 63x 1.4NA oil immersion objective with the sample immersed in oil to minimize optical aberrations, with single-photon excitation.
3, 4A-C	10x	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	128mW, 2x line scanning	30x dilution of fluoronanogold in 1x PBS	Yes	2mM HCl 0.05% Tween-20 for 6 hours; 2mM HCl 0.05% Tween-20 for 1 hour; 2mM HCl without tween for 30 minutes.	Yes	**3A,B, 4A,C:** SEM with an SE2 detector, prior to sintering. 3**C, 4B:** SEM with an SE2 detector, following sintering.

4D-F	10x	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	128mW, 2x line scanning	30x dilution of fluoronanogold in 1x PBS	Yes	2mM HCl with 0.05% Tween-20	No	**D:** LSM710, 32x 0.8NA objective, single-photon excitation, imaging fluoronanogold following deposition, while gel is expanded. **E:** LSM710, 32x 0.8NA objective, single-photon excitation, imaging reflected light following intensification, while gel is expanded. **F:** LSM710, 32x 0.8NA objective, single-photon excitation, imaging fluoronanogold shrinking but not dehydration.
S1A-C	10x	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	128mW, 2x line scanning	40µg/mL Atto 647N-labeled streptavidin in PBS, followed by biotinylated DNA. See methods.	None	Three washes in 1xPBS, followed by a wash in 1M $MgCl_2$.	No	Imaged on a widefield fluorescence microscope.
S1D	10x	333µM biotin-4-fluorescein and 1.25mM rubidium hydroxide in water	255mW, 40x 1.1NA water immersion objective. 1x line scanning	33µg/ml fluorescent streptavidin conjugates in 1x PBS, followed by biotinylated DNA, see methods.	None	Citric acid, followed by 2mM HCl with 0.05% tween.	Yes, followed by rehydration, see methods.	Fluorescent streptavidin and fluorescent DNA were imaged with an epifluorescence microscope and 20x magnification in the dehydrated state, and then in PBS.
S2B,D	10x	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	Varies, see methods.	30x dilution of fluoronanogold in 1x PBS	None	None	None	**B:** LSM710, 40x 1.1NA water immersion objective, visualizing fluorescein fluorescence with two-photon excitation at 780nm immediately after patterning, while gel is still in patterning solution. **D:** Imaging fluoronanogold in expanded state. LSM 710 single photon excitation; 32x 0.8NA objective.
S2E	20x	1mM Fluorescein-NHS, 1mM cysteamine, in water, prepared 30 minutes prior to use.	Exact power is unknown, but it was within the inversion region of S2C.	N/A	None	2mM HCl, followed by 20mM HCl, followed by 200mM HCl	Yes	SEM with an SE2 detector
S3	10x	333µM biotin-4-fluorescein and 1.25mM rubidium	255mW, 40x 1.1NA water immersion objective. 1x line scanning	33µg/ml fluorescent streptavidin conjugates in 1x PBS.	None	1x PBS	No	Imaged after shrinking in PBS. LSM 710 single photon excitation; 40x 1.1NA objective.

		hydroxide in water						
S4A	20x	None	None	None	None	2mM HCl, followed by 20mM HCl, followed by 200mM HCl	Yes	Photographic images before (left) and after (right) shrinking.
S4B	10x	None	None	None	None	2mM HCl, followed by 20mM HCl, followed by 200mM HCl	Yes	AFM image
S5	10x	500µM 5-aminomethyl fluorescein and 2mM NaOH in water.	128mW, 2x line scanning	30x dilution of fluoronanogold to 2.7µg/ml in 1x PBS	None	2mM HCl 0.05% Tween-20 for 6 hours; 2mM HCl 0.05% Tween-20 for 1 hour; 2mM HCl without tween for 30 minutes.	Yes	**A**: LSM710, 40x 1.1NA water immersion objective**,** visualizing fluoronanogold with single photon excitation after shrinking but before dehydration. **B**: Images of nanogold fluorescence were taken using a LSM710 with a 40x 1.3NA oil immersion objective with the sample immersed in oil to minimize optical aberrations, with single-photon excitation.
S7	10x	1mM Fluorescein-NHS, 1mM cysteamine, in water, prepared 30 minutes prior to use.	25% laser power, 0.39µs pixel dwell time, glycerol immersion objective.	100µM Alexa-488 maleimide in PBS.	None	1M $MgCl_2$	No	Widefield fluorescence images in expanded or shrunken states.

Table 9-3: Summary of experimental procedures used to generate each figure. Figures labeled with "S" refer to chapter 9, while those without "S" refer to chapter 3.

Chapter 10
Supplementary Information to Chapter 4

Materials and Methods:

Beads:

Bead barcodes were synthesized by the ChemGenes Corporation on one of two polystyrene supports (Agilent PLRP-S-1000A 10 μm particles or 10 μm custom polystyrene from AMBiotech). Oligonucleotide synthesis was performed as described for Drop-seq (*25*). Beads were used with one of the two following sequences:

Sequence 1:
5'- PEG Linker- TTTT-PC-
GCCGGTAATACGACTCACTATAGGGCTACACGACGCTCTTCCGATCTJJJJJJTCTTCAG
CGTTCCCGAGAJJJJJJJNNNNNNNNT30

Sequence 2:

5'- Linker-
TTTTTTTTCTACACGACGCTCTTCCGATCTJJJJJJJJTCTTCAGCGTTCCCGAGAJJJJJJJ
NNNNNNNNT30

"PC" designates a photocleavable linker; "J" represents bases generated by split-pool barcoding, such that every oligo on a given bead has the same J bases; "N" represents bases generated by mixing, so every oligo on a given bead has different N bases; and "T30" represents a sequence of 30 thymidines.

Puck Preparation:

Pucks were prepared in batches of 20 to 30, which were then stored dehydrated at 4C. Glass coverslips (Bioptechs, 40-1313-0319) were attached to a miniature centrifuge (USA Scientific 2621-0016) using double sided tape. Subsequently, the coverslip was cleaned by spraying with 70% ethanol and wiping with lens paper (VWR 52846-007). A spray-on silicone (Techspray 2102-12S) formulation was then applied to the coverslip, the cover to the minifuge was closed, and the minifuge was turned on for 10 seconds to spin coat the silicone onto the glass. The minifuge was then turned off and the cover opened, and liquid tape (Performix 24122000) was sprayed onto the coverslip. The minifuge was again closed and turned on for 10 seconds. The coverslip was then carefully removed from the minifuge, and a gasket (3 mm diameter holes from Grace Biolabs, CW-50R-1.0) was placed on top of the coverslip and pressed down. Beads were pelleted and washed twice in 500ul ultrapure water (Thermofisher, 10977015), and resuspended to a final concentration of 100,000 beads/uL. 10 μl of bead solution

was pipetted into each position on the gasket. The coverslip-gasket filled with beads centrifuged at 40C, 850g for at least 30 minutes until the surface was dry.

The gasket was carefully removed from the dried coverslip. Gentle pipetting of water directly onto the pelleted beads removed all beads except for those directly in contact with the liquid tape layer. The resulting bead monolayer was allowed to dry, generating the final puck. Beads removed in this way could be stored at 4C for later use. As much water was removed from the resulting pucks as possible, and the pucks were left to dry.

Puck Sequencing:

Puck sequencing was performed using SOLiD chemistry in a Bioptechs FCS2 flow cell using a RP-1 peristaltic pump (Rainin), and a modular valve positioner (Hamilton MVP). Flow rates between 1mL/min and 3mL/min were used during sequencing. Imaging was performed using a Nikon Eclipse Ti microscope with a Yokogawa CSU-W1 confocal scanner unit and an Andor Zyla 4.2 Plus camera. Images were acquired using a Nikon Plan Apo 10x/0.45 objective. After each ligation, images were acquired in the following channels: 488nm excitation with a 525/36 emission filter (MVI, 77074803); 561nm excitation with a 582/15 emission filter (MVI, FF01-582/15-25); 561nm excitation with a 624/40 emission filter (MVI, FF01-624/40-25); and 647nm excitation with a 705/72 emission filter (MVI, 77074329). The final stitched images were 6030 pixels by 6030 pixels.

Sequencing consisted of three steps: (1) primer hybridization; (2) ligation; and (3) stripping. During primer hybridization, a primer was flowed into the flow cell at 5 µM concentration in 4x SSC for 20 minutes. Subsequently, the flow cell was washed in 3 mL of SOLiD buffer F. Following buffer F wash, ligation mix (recipe below) was flowed into the chamber and allowed to sit for 20 minutes, before being flowed back into its original reservoir. Ligation mix was reused for ~10 ligations before being replenished. Following ligation, the flowcell was washed again in buffer F. Then, to cleave the fluorophore off the ligated SOLiD oligo, we flowed 1.5 mL of SOLiD buffer C into the chamber, followed by 1.5 mL of SOLiD buffer B, and repeated this cleave step once again. We then washed the flowcell in buffer F and repeated the ligation step. After the second ligation step, 10 mL of 80% formamide in water was flowed into the flowcell and left for 10 minutes. The flowcell was then washed in instrument buffer, and the process repeated with the next primer.

In order to sequence bead barcodes, we performed 2 ligations on each of 10 primers (Table 10-1), of which 6 were “constant” bases (i.e., the first ligation on a primer recessed by 2 or more

nucleotides, which only sequence the primer and thus contain no information about the barcode sequence). The final bead barcodes were 14 bases long.

Each 3mm puck presented in this manuscript consists of roughly 70,000 beads, with a total cost of less than $0.10. Moreover, roughly 250uL of SOLiD SR-75 sequencing oligo is required to sequence a batch of 30 pucks. With other necessary reagents, each 3 mm puck requires roughly $10 of SOLiD sequencing reagents. However, each 3mm puck also requires roughly 300 million reads, or ~$200-$500 worth of sequencing using the Illumina Novaseq platform. Thus, the dominating cost of Slide-seq is the cost of short-read sequencing.

Ligation mix:

1x T4 DNA Ligase Buffer (Enzymatics)

6 U/uL T4 DNA Ligase (Rapid) (Enzymatics)

40x dilution of SOLiD SR-75 sequencing oligo (Life Technologies).

Image Processing and Basecalling:

All image processing was performed using a custom-built processing suite in Matlab. Briefly, we acquired one image per puck after each ligation, and each image contained four color channels. First, color channels were co-registered to each other by thresholding the images and maximizing the cross-correlation between the thresholded images. Subsequently, for each puck, the images of each ligation were registered to the image of the first ligation using a SIFT-RANSAC image registration algorithm based on the VLFeat SIFT package in Matlab (*330*). Registered images were then base-called on a pixel-wise basis, as follows. First, the intensities in the Cy3 channel were multiplied by a factor of 0.5 and subtracted from the intensities in the TxR channel, which accounts for crosstalk between the channels resulting from the excitation of TxR using the 561nm laser. Furthermore, for even-numbered ligations, the image of the previous ligation was multiplied by a factor of 0.4 and then subtracted on a channel-by-channel basis from the image of the even ligation. Each pixel was then called by intensity. For pucks made using the 180402 bead batch, we further enforced the expected base balance by including an additional step in which the intensities of the dimmest channels were progressively increased until each channel accounted for between 20% and 30% of the pixels in the center of the image.

Beads were subsequently identified from the base-called images as follows. Each pixel was assigned a number, the base 5 representation of which corresponds to the bases that were called at

that pixel on each ligation. Every such number that occurred on at least 50 connected pixels in the image was determined to be a bead, represented by the centroid of the connected cluster.

SOLiD barcodes were then mapped to Illumina barcodes using a custom-built Matlab application that identifies the pairwise distance between all members of the two sets of barcodes. Pairs of SOLiD and Illumina barcodes were saved for further analysis if: (1) the two barcodes were separated by at most two Levenshtein distance units; (2) there were at least 10 transcripts identified in Illumina sequencing with that barcode; and (3) the mapping between the barcodes was unique, i.e. if there were no other barcodes at equal or lower edit distance to either barcode.

Tissue Handling:

Fresh frozen tissue was warmed to -20 C in a cryostat (Leica CM3050S) for 20 minutes prior to handling. Tissue was then mounted onto a cutting block with OCT and sliced at a 5° cutting angle at 10 µm thickness. Pucks were then placed on the cutting stage and tissue was maneuvered onto the pucks. The tissue was then melted onto the puck by moving the puck off the stage and placing a finger on the bottom side of the glass. The puck was then removed from the cryostat and placed into a 1.5 mL eppendorf tube. The sample library was then prepared as below. The remaining tissue was re-deposited at -80 C and stored for processing at a later date.

Library preparation:

RNA Hybridization:

Pucks in 1.5 mL tubes were immersed in 200 µL of hybridization buffer (6x SSC with 2 U/µL Lucigen NxGen RNAse inhibitor) for 15 minutes at room temperature to allow for binding of the RNA to the oligos on the beads.

First Strand Synthesis

Subsequently, first strand synthesis was performed by incubating the pucks in RT solution for 1 hour at 42 C.

RT solution:

75 µL H2O

40 µL Maxima 5x RT Buffer (Thermofisher, EP0751)

40 µL 20% Ficoll PM-400 (Sigma, F4375-10G)

20 µL 10 mM dNTPs (NEB N0477L)

5 µL RNase Inhibitor (Lucigen 30281)

10 µL 50 µM Template Switch Oligo (Qiagen #339414YCO0076714)

10 µL Maxima H- RTase (Thermofisher, EP0751)

Tissue Digestion:
200 µL of 2x tissue digestion buffer was then added directly to the RT solution and the mixture was incubated at 37C for 40 minutes.

2x tissue digestion buffer:

200 mM Tris-Cl pH 8

400 mM NaCl

4% SDS

10 mM EDTA

32 U/mL Proteinase K (NEB P8107S)

Library Amplification
The solution was then pipetted up and down vigorously to remove beads from the surface, and the glass substrate was removed from the tube using forceps and discarded. 200 µL of Wash Buffer was then added to the 400 µL of tissue clearing and RT solution mix and the tube was then centrifuged for 3 minutes at 3000 RCF. The supernatant was then removed, the beads were resuspended in 200 µL of Wash Buffer, and were centrifuged again. After repeating this procedure an additional 2 times, the beads were moved into a 200 µL PCR strip tube, pelleted in a minifuge, and resuspended in 200 µL of water. The beads were then pelleted and resuspended in library PCR mix and PCR was performed.

Wash Buffer:

10 mM Tris pH 8.0

1 mM EDTA

0.01% Tween-20

Library PCR mix:

23 µL H2O

25 µL of 2x Kapa Hifi Hotstart ready mix (Kapa Biosystems KK2601)

1 µL of 100 µM Truseq PCR handle primer (IDT)

1 µL of 100 µM SMART PCR primer (IDT)

PCR program:

95 C 3 minutes

4 cycles of:

98 C 20 s

65 C 45 s

72 C 3 min

9 cycles of:

98 C 20 s

67 C 20 s

72 C 3 min

Then:

72 C 5 min

4 C forever

PCR cleanup and Nextera Tagmentation

The PCR product was then purified by adding 30 µL of Ampure XP (Beckman Coulter A63880) beads to 50 µL of PCR product. The samples were cleaned according to manufacturer's instructions and resuspended into 10 µL of water. 1 µL of the library was quantified on an Agilent Bioanalyzer High sensitivity DNA chip (Agilent 5067-4626). Then, 600 pg of PCR product was taken from the PCR product and prepared into Illumina sequencing libraries through tagmentation with Nextera XT kit (Illumina FC-131-1096). Tagmentation was performed according to manufacturer's instructions and the library was amplified with primers Truseq5 and N700 series barcoded index primers. The PCR program was as follows:

72°C for 3 minutes

95°C for 30 seconds

12 cycles of:

95°C for 10 seconds

55°C for 30 seconds

72°C for 30 seconds

72°C for 5 minutes

Hold at 10°C

Samples were cleaned with AMPURE XP (Beckman Coulter A63880) beads in accordance with manufacturer's instructions at a 0.6x bead/sample ratio (30 µL of beads to 50 µL of sample) and resuspended in 10uL of water. Library quantification was performed using the Bioanalyzer. Finally, the library concentration was normalized to 4nM for sequencing. Samples were sequenced on the Illumina NovaSeq S2 flowcell with 12 samples per run (6 samples per lane) with the read structure 42 bases Read 1, 8 bases i7 index read, 50 bases Read 2. Each puck received approximately 200-400 million reads, corresponding to 3,000-5,000 reads per bead.

Calculation of Bead Packing:

To estimate the packing fraction of the beads, we imaged 10 pucks with 488 nm light on the same microscope mentioned above after deposition onto the surface and prior to in situ sequencing. The signal was normalized to background and the image was binarized. The percent packing was reported as the fraction of the image occupied by the beads divided by the theoretical packing fraction of 0.9069 for dense packing of uniform spheres on a 2D surface. The mean and standard deviation of packing are reported in Figure 10-1D.

Clustering Analysis:

For clustering of the pucks shown in Figure 4-1C,D and Figure 10-2, highly variable genes were identified by running FindVariableGenes() in the Seurat package in R, using a y.cutoff of 0.7 in liver, 0.6 in kidney and olfactory bulb, and 0.5 in hippocampus and cerebellum. For the hippocampus and cerebellum analyses, variable genes identified from the published datasets from these tissues were also included. Non-negative matrix factorization was performed using the NNLM package in R, on standardized, log-transformed values, with a k of 8 in liver and kidney, 6

in olfactory bulb, 13 in cerebellum, and 20 in hippocampus. For each bead, the largest factor loading from NMF after L2 normalization was used to assign cluster membership.

Diffusion Analysis and Comparison of smFISH, scRNAseq and Slide-seq:

An image of Slide-seq bead signal density was generated through plotting the pixel intensity of each bead as a linear representation of the number of UMIs captured (180602_17, 180602_20, 180611_6). Single molecule FISH was performed on the serial section using HCR v3.0 (Molecular Technologies) with probes against three strong CA1 markers (*Slc17a7, Ociad2, Atp2b1)* and co-stained with DAPI. Images were taken of the tissue sections and profile was taken across a region of CA1 for the Slide-seq image, the DAPI image, as well two of the three genes (*Slc17a7*, *Atp2b1*) used in the FISH data. The full width half maximum (FWHM) of the profile was then calculated for 10 such profiles across the CA1 for both Slide-seq and the serial tissue sections in both DAPI and FISH (Figure 10-5).

To quantify the efficiency of mRNA capture, we compared the counts of these genes in Slide-seq, scRNAseq and smFISH. FISH images were taken using a 40x 1.15 Nikon Plan Apo water immersion objective. Two fields of view (FoV, 652 µm x 652 µm) were imaged across CA1 for each of the genes tested (*Slc17a7, Ociad2, Atp2b1)* for each of the pucks for a total of six regions. Transcript counts for smFISH data were obtained by using StarSearch (rajlab.seas.upenn.edu/StarSearch/). Slide-seq data from the same FoV on the puck corresponding to the serial section was counted for the same marker genes (*Slc17a7, Ociad2, Atp2b1)*. Using the DAPI image for each of the FoV in CA1, we estimated the number of cells present in the FoV. Finally, a random sample of CA1 neurons from Drop-seq was taken equal to the number of cells present in the field of view and the sums for the three genes listed were taken across all single cell barcodes. The result of the total counts is shown as a bar plot (Figure 10-4E) highlighting the differences in counts between the technologies.

Comparison to Bulk sequencing:

To compare the capture of Slide-seq to bulk RNAseq dat (Figure 10-4C), we used a stranded mRNA Truseq kit (Illumina #20020594) to prepare stranded PolyA selection libraries from a dissected sagittal mouse hippocampus. The libraries were sequenced and transcripts per million (TPM) for each gene were generated using Salmon post alignment with STAR (*331*). For Slide-seq data, average transcripts per million (ATPM) was computed by summing counts for each gene, across all beads on a puck, and dividing by the sum of all UMIs on the puck, and dividing by 1 million (total UMI count/1million). The per-gene distribution for each of these values (bulk TPM and Slide-seq ATPM) was plotted and linear regression was performed giving an $R = 0.89$.

Comparison to scRNAseq:

To compare the capture of Slide-seq to scRNAseq, as in Figure 10-4A,B,D, we extracted cells assigned to the CA1 cluster from hippocampal atlas data (*24*). For five hippocampal pucks (180531_13, 180531_17, 180531_22, 180602_20, and 180620_4) we isolated beads in CA1 by hand cropping. We then plotted the distributions of the number of transcripts per bead for each of the three genes considered (*Slc17a7*, *Atp2b1*, and *Ociad2*), and the total number of transcripts per bead, for the atlas (Figure 10-4A) and Slide-seq (Figure 10-4B) data. Figure 10-4D was likewise generated by plotting the mean expression per bead in the atlas CA1 data against the mean expression per bead in the Slide-seq CA1 region for every gene in both the atlas and Slide-seq datasets. Note that for Figure 10-4A,B,D, expression levels for Slide-seq are averaged over the 5 pucks listed above.

Calculation of UMI per cell estimates:

For calculation of the total UMIs captured normalized to total cells in Figure 10-3D and Figure 10-4F, we used DAPI images of serial stained tissue sections to estimate the total number of cells within a puck. Segmentation was performed in ImageJ by first scaling signal to background and binarizing the image followed by applying a 1.5 µm Gaussian Blur and a watershed transform. Nuclei were counted only if they had a diameter greater than 2 µm and less than 12 µm. The total number of UMIs from the puck was then divided by the number of nuclei obtained to generate the statistic total transcripts/total cells.

Cell Type Deconvolution (NMFreg):

For each bead, the contribution of each cell type to the RNA on that bead was computed using a custom method, implemented in Python, termed NMFreg (Non-Negative Matrix Factorization Regression). The method consisted of two main steps: first, single-cell atlas data previously annotated with cell type identities (*24*) was used to derive a basis in reduced gene space (via NMF), and second, non-negative least squares (NNLS) regression was used to compute the loadings for each bead in that basis.

To perform NMF on the single-cell data, highly variable genes were first selected as in (*24*), and NMF was performed using a specified number of factors (see below). Each factor was then assigned to the cell type whose cells from (*24*) most frequently had their largest loading on that factor. Next, for each Slide-seq bead, we first computed the bead loadings in the basis using NNLS. The resulting matrix of factor loadings (with dimensions of the number of beads by the number of factors) was scaled so each factor had unit variance. Finally, the cell type of the bead was assigned based on the identity of the maximum factor loading.

For the implementation of NMFreg in Figure 4-2B,C, an adult mouse single-cell cerebellum dataset (*24*) was used to define the NMF basis, using a k (factor number) of 25. The published cluster identities from this tissue were modified to remove clusters of cells outside of the Slide-seq-assayed anatomical region (e.g., cells from midbrain not seen on the puck) and to reduce the number of subpopulations. Specifically, all endothelial populations were merged together into one population, as were non-Bergmann astrocytes and oligodendrocytes. Interneurons not annotated as unipolar brush or Golgi (clusters 3-1, 3-2, 3-3, and 3-4)—which could not be assigned to a specific type in the published dataset—were also grouped together. Only Slide-seq beads with more than 15 unique genes were used in NNLS regression. For the implementation of NMFreg in Figure 2D, an adult hippocampus scRNA-seq dataset (*24*) was used in NMF setting k to 30 with 5 variable gene cutoff for bead inclusion. The first-level published cluster identities were used for bead assignment to cell types.

For the implementation of NMFreg in Figure 4-2D, Figure 4-3 and Figure 4-4, the data were processed using published cerebellum (Figure 4-3) or hippocampus (*24*) (Figure 4-2D, Figure 4-4) datasets. In Figure 4-2D, Figure 4-3, and Figure 4-4, Slide-seq beads were used for NNLS regression if they had at least 5 variable genes. For Figure 4-4, hippocampus cluster 13 was interpreted as marking mitosis.

Often, multiple cell types may be present on a bead. Thus, for the purpose of calculating the number of cells of each type appearing on the puck, as in Figure 4-2C and Figure 10-7, we determined that a cell type was present on a bead if the L2 norm of the vector of factor loadings for that cell type was at least half of the L2 norm of the vector of all factor loadings for that bead. Figure 10-7 shows the numbers plotted in Figure 4-2C as a function of this cutoff.

Confidence Thresholding:

The bead factor loadings returned by NMFreg are in general less pure than the factor loadings obtained for single-cell sequencing data, possibly reflecting both the sparsity of the Slide-seq data and RNA contributions of other adjacent cell types. In Figure 10-8, in order to determine whether a given bead could be confidently assigned to its highest contributing cell type, we computed a cell-type-specific, single-cell-derived threshold. The threshold for a given cell type was the maximum loading of this cell type among all single cells not assigned to this cell type in single cell atlas data. A bead was said to be confidently assigned if the L2 norm of the vector of factors corresponding to that cell type exceeded the threshold. This comparison was made after normalizing so that the sum of the L2 norms of the vector of factors for each cell type would be equal to 1.

For Figure 10-8A-E, we first performed NMFreg using only beads with at least 100 total transcripts. This decreases the number of beads called by 72.6% +/- 13.7% (mean+/-std over 7 cerebellar pucks). Interestingly, there was no relationship between the number of UMIs per bead and the confidence score of the bead (Figure 10-8F). Note that for the computation in Figure 10-8F, NMFreg was performed on all bijectively mapped beads, which must have at least 10 transcripts.

The diameter of Slide-seq beads is 10 µm (original feature size). For the analysis in Figure 10-8A-D, in an attempt to investigate the importance of the size of the features, we generated larger beads *in silico*, selecting artificial feature sizes of 20, 40, and 100 µm. Aggregate array features were performed by taking bead centroid locations obtained through SOLiD sequencing and forming a grid of defined size over the locations of the beads and aggregating beads within each region of the grid and treating the resulting data as a single bead.

Robustness of NMFreg:

To evaluate the robustness of the NMFreg cell type assignments, we calculated a consistency metric (Figure 10-6B,C) by running NMFreg for 30 values of k (the number of factors) between 18 and 48, or for 30 different random seeds. For each Slide-seq bead, the consistency was then defined as the fraction of NMFreg runs on which the bead was assigned to the most common cell type across conditions tested. These results were plotted as a cumulative distribution function of the consistency score per bead.

3D volume reconstruction of hippocampus:

For Figure 4-2D, beads assigned to hippocampus scRNA-seq clusters 4, 5, and 6 (CA fields and DG) (*24*) from serial hippocampal Slide-seq sections were plotted in space. Sequential slices were roughly aligned by the density and shape of beads localized to hippocampal morphology. Alignments were refined with the ImageJ plugin TurboReg (*332*). Volumes were reconstructed in 3D by generating a 3D image stack with a sphere of diameter 12.5 µm with intensity proportional to number of UMIs centered on each bead centroid.

Hippocampal Subtype Images:

Metagenes for Figure 4-2E were identified from cell type specific atlas expression. The metagenes are listed in Table 10-2.

Metagenes were plotted via density plots (see below) on their corresponding Atlas clusters. Beads corresponding to hippocampal atlas clusters 4, 5, and 6 (CA1, CA2/3, and DG) were displayed in light gray as a counterstain.

Density Plots:

For the density plot images in Figure 4-2E, Figure 4-3 (black backgrounds), Figure 4-4 (black backgrounds), Figure 10-9E,F, Figure 10-11A,C,F,G and Figure 10-12, we formed an image as follows. Each point P in the 6030 x 6030 images was assigned an intensity equal to the sum of the intensities of all beads with centroids lying within 44-pixel square centered on P. For Figure 4-4B,C, each bead assigned to the indicated NMFreg cluster was assigned a unit intensity, while the intensity for each bead in Figure 4-3C,D,F,G was taken as the total number of transcripts belonging to genes in the indicated metagene. Finally, the images were passed through Gaussian filters with a standard deviation of 12 pixels.

For the images with blue backgrounds in Figure 4-4, each bead was represented by a square of length 70 pixels on each side, with intensity equal to the total number of transcripts belonging to the set of genes indicated in the legend. Overlapping squares summed their intensities in the overlap region. For Figure 4-4G-K, all the images within a given panel are normalized to the same values (i.e., the same colors represent the same values in all four images).

Significant Gene Calling:

To determine whether a transcript had a significantly non-random spatial distribution within a particular set of beads (for example, within the set of beads called as Purkinje neurons by NMFreg), we first calculated the matrix of pairwise Euclidean distances between all beads in the set. We then compared the distribution of pairwise distances between the beads expressing at least one count of that transcript (Figure 10-10A) to the distribution of pairwise distances between an identical number of beads, sampled randomly from all mapped beads within the set with probability proportional to the total number of transcripts on the bead (Figure 10-10B). (Rigorously, therefore, the spatial significance gene algorithm determines whether the spatial distribution of a particular transcript differs significantly from the spatial distribution of all transcripts.) Specifically, we generated 1000 such random samples, and for each sample calculated the distribution of pairwise distances. We then calculated the average distribution of pairwise distances, averaged across all 1000 samples (Figure 10-10B, bottom). Finally, we calculated the L1 norm between the distribution of pairwise distances for the true sample of beads and the average distribution (Figure 10-10C), and the L1 norm between the distribution of pairwise distances for each of the 1000 random samples and the average distribution (Figure 10-10D). We defined p to be the fraction of random samples having distributions closer to the average distribution (under the L1 norm) than the true sample, and considered any genes with values $p \leq 0.005$ (Figure 10-10E). Often, as many as 4000 genes would pass the filters described above, leading to a high false-positive rate. For this reason, various methods were used to enrich for true positives (described in detail below), for example by using multiple biological replicates, or by identifying clusters of correlated genes within the set of spatially significant genes.

Genes were identified as spatially non-random using a custom Matlab application (see Figure 10-10). In regions in which cells are densely packed, one often finds markers from multiple different cell types on a single bead. In some instances, when seeking to identify spatially patterned genes within a cell type, our algorithm identified markers of cell types in spatial proximity. For example, in cerebellum, granule cell markers were sometimes identified as spatially non-random within a set of oligodendrocytes due to the proximity of the granular layer and the cerebellar white matter. For this reason, genes were identified as candidates for the statistical significance analysis within a particular cluster if they had an average expression of at least 0.1 transcripts per bead within that cluster in the atlas reference dataset, or if the variance within that cluster in the atlas reference dataset was at least 0.01 transcripts squared and the ratio of the variance to the squared expression was at least 7.5 (an empirically determined value). Moreover, candidate genes for the statistical significance analysis were required to have at least one transcript on at least 15 beads in Slide-seq.

Overlap Analysis:

To identify genes that are significantly correlated or anticorrelated with other genes, we applied a custom Matlab algorithm. For simplicity of description, we consider the case of determining the genes that are correlated or anticorrelated with a particular gene, gene A. For each gene in the genome, we generated a "true" image in which each bead with at least one transcript of the gene was represented by a square of side length 100 pixels (~64 microns). Images were then binarized, so overlapping squares did not sum. Then, for each gene, we additionally generated 50 "random" images in which the same number of transcripts were redistributed across all beads with probability proportional to the number of reads per bead. We then calculated the pixel-wise inner product between the image of gene A and the 50 random images every other gene, and calculated the mean and standard deviation of the inner products. We then compared the mean and standard deviation to the inner products of the image for gene A with the true image of every other gene, obtaining a Z score for each gene. All genes with Z scores greater than 3 were deemed correlated, while those with Z scores less than 3 were deemed anticorrelated.

Regional Significance Analysis:

For several of the analyses in Figure 4-3 and Figure 10-11, we used the following procedure to determine whether the expression of a gene within a given region of the puck was significantly enriched or depleted. We divided Puck 180819_12 into 5 regions (Figure 10-12): a dorsal region, a ventral region, a nodulus region, a nodulus-uvula region (consisting of the nodulus and the anterior uvula), and a VI-VII region, corresponding to the posterior side of lobule VI and the anterior side of lobule VII. The significance of a gene was then determined by a Fisher exact test

performed on the contingency matrix [A, N-A; B, M-B], where A is the number of counts of the gene in the designated region, B is the number of counts outside of the designated region, N is the total number of counts of any gene in the designated region, and M is the total number of counts of any gene outside of the designated region. As in the case of the significant gene-calling algorithm, this analysis could be performed on a subset of the beads on the puck. This procedure provides a list of genes with a significantly different pattern of expression within the designated region than outside of the designated region, regardless of whether the expression is elevated or depressed.

In Figure 4-3B, *Kctd12* and *Car7* did not pass the *p*-value cutoff, but are displayed as squares to demonstrate their location relative to *Aldoc*.

Identification of spatially variable genes in the cerebellar granular layer:

We identified *Gprin3* by finding all of the genes with significant expression ($p<0.001$, Fisher exact test) in the ventral part of puck 180819_12 compared to the dorsal part of the puck, for which more than 80% of the transcripts were in the ventral portion. This yielded three hemoglobin genes, *Th*, *Cemip, Gprin3, Mab21l2,* and *Syndig1l*. The three hemoglobin genes and *Th* were discarded because they were not expressed in granule cells.

Identification of Aldoc- and Plcb4-associated genes in the cerebellar Purkinje layer:

To identify the *Aldoc* and *Plcb4*-associated genes, we ran the significant gene calling algorithm on 14 cerebellar pucks (3 coronal, 11 sagittal), restricted to beads called as cluster 2 (Purkinje cells), cluster 7 (Bergmann glia), or the union of cluster 2 and 7 together. In this way, we identified 669 genes that were significant on at least one of pucks. This method presumably included many false positives, due to the high false discovery rate of the spatial significance algorithm. For that reason, we came up with the following procedure to restrict the set of spatially significant genes to those that correlated more with *Aldoc* than with *Plcb4*, or more with *Plcb4* than with *Aldoc*, on the grounds that false positives or genes unrelated to the Zebrin staining pattern would not correlate more with one than with the other. To identify genes correlating preferentially with *Aldoc* or *Plcb4*, we used the significance overlap algorithm to identify, for each of the 669 genes, the other genes in the set that correlate spatially with that gene on at least one puck. We then calculated, for each pair of genes in the set of 669, the magnitude of the intersection of the sets of correlating genes. To construct the matrix in Figure 4-3A, we restricted that overlap matrix to the set of genes that have a larger intersection with *Aldoc* by at least 3 genes, or a larger intersection with *Plcb4* by at least 3 genes.

For the purposes of displaying the matrix thus obtained in Figure 4-3A, we first normalized the *i,j*th entry of the matrix by dividing as follows:

$$p_{i,j} \leftarrow \frac{p_{i,j}}{\sqrt{p_{i,i}p_{j,j}}}$$

We then divided each column of the resulting matrix by the sum of the column. Finally, because the resulting matrix was asymmetric, we summed the matrix and its transpose. For purposes of display, we then performed Ward clustering in Matlab and ordered them by cluster.

Identification of *Hspb1* pattern:

To generate Figure 4-3B, we included genes if they had significant expression in the nodulus-uvula region at $p<0.001$ (Fisher exact test). We excluded *Ttr*, which was not expressed in Purkinje cells. For purposes of display, *Kctd12* and *Car7* were added to the graph as squares to help illustrate the clustering of *Aldoc*-like genes and *Cck*-like genes.

Identification of *B3galt5* pattern:

To generate Figure 10-11E, we included genes if they had significant expression in the nodulus at $p<0.05$ (Fisher exact test) and significant expression in the VI-VII region at $p<0.05$ (Fisher exact test).

Identification of injury-correlated genes:

To identify all genes that correlated spatially with *Hba-a1*, *Hba-a2*, and *Hbb-bs*, we ran the overlap analysis on pucks 180819_1, 180819_2, 180819_3, 180819_4, 180819_13, 180819_14, and 180728_15. The first four pucks were taken from a single mouse in the coronal orientation, while the last three pucks were taken from a second mouse in the sagittal orientation. We considered all genes that correlated with at least one of those three genes on at least 2 pucks. The only genes identified in this way, besides hemoglobin, were *Lars2* (a marker of rRNA, see Identification of rRNA below) and *Fos*.

To identify genes correlating with *Vim*, *Ctsd*, or *Gfap* at the 3-hour timepoint (pucks 180819_16, 180819_18, 180819_19, and 180821_3) or the 2-week timepoint (pucks 180819_5, 180819_6, 180819_7, and 180819_8) (Table 10-2), we ran the overlap algorithm. All four pucks for each timepoint were taken from a single mouse in the sagittal orientation. The corresponding list in Table 10-2 is the set of all genes that correlate with at least one of *Vim*, *Ctsd*, or *Gfap* on at least two of the pucks.

Distance Measurements for Injury Site:

The distance measurements in Figure 4-4D,E were performed by plotting beads in each of cluster of interest with radius linearly proportional to the number of transcripts per bead, with one

transcript corresponding to a 25 pixel diameter and 500 transcripts corresponding to a 125 pixel diameter. Beads with more than 500 transcripts were plotted with a 125 pixel diameter. This was done to ensure that beads with more transcripts were weighted more heavily when calculating the spatial profile of the cell types. We then drew boxes around the injury and took line profiles (i.e., summed along one axis) across the injury site, to generate the profiles in Figure 4-4D,E.

For measurements of the mitosis layer thickness, we took two measurements from one puck (Puck_180821_3, both sides of the injury site) and one measurement from a second puck (Puck_180819_19, the bottom side of the injury site). For measurements of the astrocyte scar thickness and the microglial penetration thickness, we took six measurements: two on each side of the scar from each of three pucks (Puck_180819_5, Puck_180819_6, and Puck_180819_7).

For the distance measurements in Figure 4-4K, we plotted grayscale versions of the images in Figure 4-4K using the IEG metagene listed in Table 10-2, and took line profiles similar to those taken for the measurements in Figure 4-4D,E. We took measurements from each side of the injury for puck 180819_7 (Figure 4-4G, bottom). We additionally took measurements from one side of the injury on pucks 180819_5 and 180819_6. We only used one side from those pucks on the grounds that the injury site was very close to the edge of the puck on one side.

Two of the three-day injury pucks (180819_16 and 180819_18) were excluded from all distance measurements on the grounds that the tissue damage was not readily identifiable on the puck.

One two-week injury puck (180819_8) was excluded from all distance measurements on the grounds that the tissue slice was more lateral than the other tissue slices. It showed neither enrichment of the immediate early genes around the injury site, nor a dip in astrocyte density in the middle of the scar, leading us to suspect that it was at the edge of the wound.

Identification of rRNA in pucks:

During analysis of the 2-hour injury pucks, we observed many counts of the *Lars2* gene correlating with hemoglobins and *Fos* at the injury site (Figure 10-14). Upon investigation of the *Lars2* gene, we found using RepeatMasker (http://www.repeatmasker.org/) that it has a rRNA-derived repeat in its 3' UTR, leading us to hypothesize that the counts we observed of Lars2 might in fact be misaligned rRNA reads (*333*). Moreover, we found that the spatial distribution of *Lars2* counts across the puck is highly correlated to the counts of rRNA, supporting this hypothesis. We thus used *Lars2* as a proxy for rRNA expression in Figure 4-4A.

Staining and Validation of the Cortical Injury protocol:

To validate the cortical injury procedure in Figure 10-13, we stained with (Abcam ab53554) against Glial Fibrillary Acidic Protein (*Gfap*), a marker of activated astrocytes and microglia that should be enriched near the site of injury. To further validate our finding of *Vim* as a gene strongly upregulated at the site of injury we also stained with (Abcam ab20346) against *Vim* showing that it is expressed precisely at the injury. Sections were sectioned at 10 μm and post fixed in 4% PFA for 10 minutes. Post fixation they were washed three times in PBS before being co-stained with the antibodies listed above for two hours at 37C. Post primary antibody incubation sections were washed three times for five minutes in 10 mL of 1x PBS. Sections were then stained with the appropriate secondary antibodies (Abcam ab150135 and ab175700) for one hour in 1XPBS. Sections were then washed three times for five minutes in 1X PBS and co-stained with DAPI and imaged using a 20x 0.75 Nikon Plan Apo objective.

Gene Ontology Analysis:

For Figure 4-4G-J, we first identified (using the tool at http://geneontology.org/) gene ontology annotations that were significantly enriched within the set of genes that correlated with the injection site only at the 2 week timepoint or only at the 3 day timepoint (see "Identification of injury-correlated genes," above). Each image in Figure 4-4G-J is a heatmap showing the total gene counts summed over all genes in each annotation. For each of Figure 4-4G-J, both heatmaps were normalized to the maximum value in either the top or bottom heatmap. Thus, the values shown for the 2-week and 3-day pucks are on the same scale, and the units are arbitrary.

The annotation used for Figure 4-4G was "mitotic cell cycle." Figure 4-4H was "antigen processing and presentation via MHC class Ib." The annotation used for Figure 4-4I was "gliogenesis." The annotation used for Figure 4-4J was "oligodendrocyte development."

Animal Handling:

Animals were group housed with a 12-hour light-dark schedule. All procedures involving animals at MIT were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals under protocol number 1115-111-18 and approved by the Massachusetts Institute of Technology Committee on Animal Care. All procedures involving animals at the Broad Institute were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals under protocol number 0120-09-16.

Traumatic Brain Injury Model:

Animals for the TBI model were anesthetized and processed according to a standard intracranial injection protocol as a model for injury. Specifically, mice were anesthetized using isofluorane and stereotactically restrained. Subsequently, an incision was made in the scalp and a hole was made

in the skull using a dental drill. A Hamilton needle (32 gauge, 7803-04) was lowered to 2 mm below the surface of the skull, and was then promptly retracted. The wound was closed using Vetbond, and the animal was allowed to recover. Mice were treated with Buprenorphine-SR and Meloxicam for analgesia. Mice were sacrificed by cardiac perfusion 2 hours, 3 days, or 2 weeks following the injury.

Transcardial Perfusion:

Animals were anesthetized by administration of isoflurane in a gas chamber flowing 3% isoflurane for 1 minute. Anesthesia was confirmed by checking for a negative tail pinch response. Animals were moved to a dissection tray and anesthesia was prolonged via a nose cone flowing 3% isoflurane for the duration of the procedure. Transcardial perfusions were performed with ice cold pH 7.4 HEPES buffer containing 110 mM NaCl, 10 mM HEPES, 25 mM glucose, 75 mM sucrose, 7.5 mM MgCl2, and 2.5 mM KCl to remove blood from brain and other organs sampled. The appropriate organs were removed and frozen for 3 minutes in liquid nitrogen vapor and moved to -80C for long term storage.

Human Sample Information:

Human cerebellum tissue assayed in Figure 10-3 was obtained from the Sepulveda Research Corporation through the NIH NeuroBioBank. The tissue was received without identifiable information, and did not meet the definition of human subjects research (project # NHSR-4235).

Figures

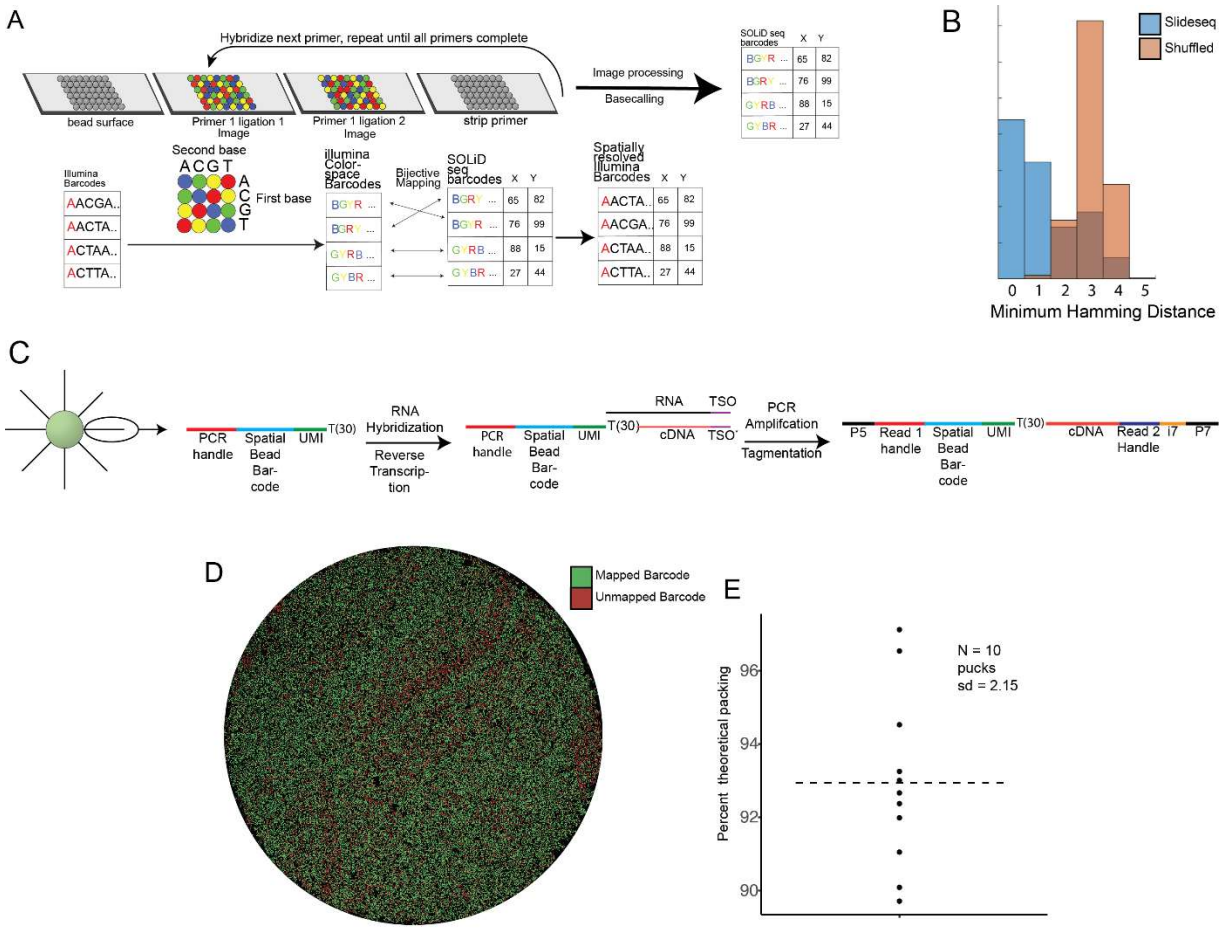

Figure 10-1 (A) Top: schematic of the in situ sequencing and base-calling system established for generation of barcoded surfaces ("pucks"). Bottom: schema for mapping of Illumina barcodes to SOLiD barcodes. **(B)** Minimum hamming distance between Illumina colorspace-converted barcodes and barcodes from a puck sequenced in situ using SOLiD chemistry (Blue, puck barcodes, Orange, shuffled puck barcodes). **(C)** Structure of the library at each stage of the preparation. **(D)** Barcode mapping across the puck. Beads colored green have a barcode bijectively matched between Illumina and SOLiD sequencing. Red beads are SOLiD-called barcodes not detected by Illumina sequencing. **(E)** Beanplot shows the packing fraction of the beads on the surface, as a fraction of the maximum theoretical density. The average packing fraction is 85%, which is 93% of the theoretical maximum.

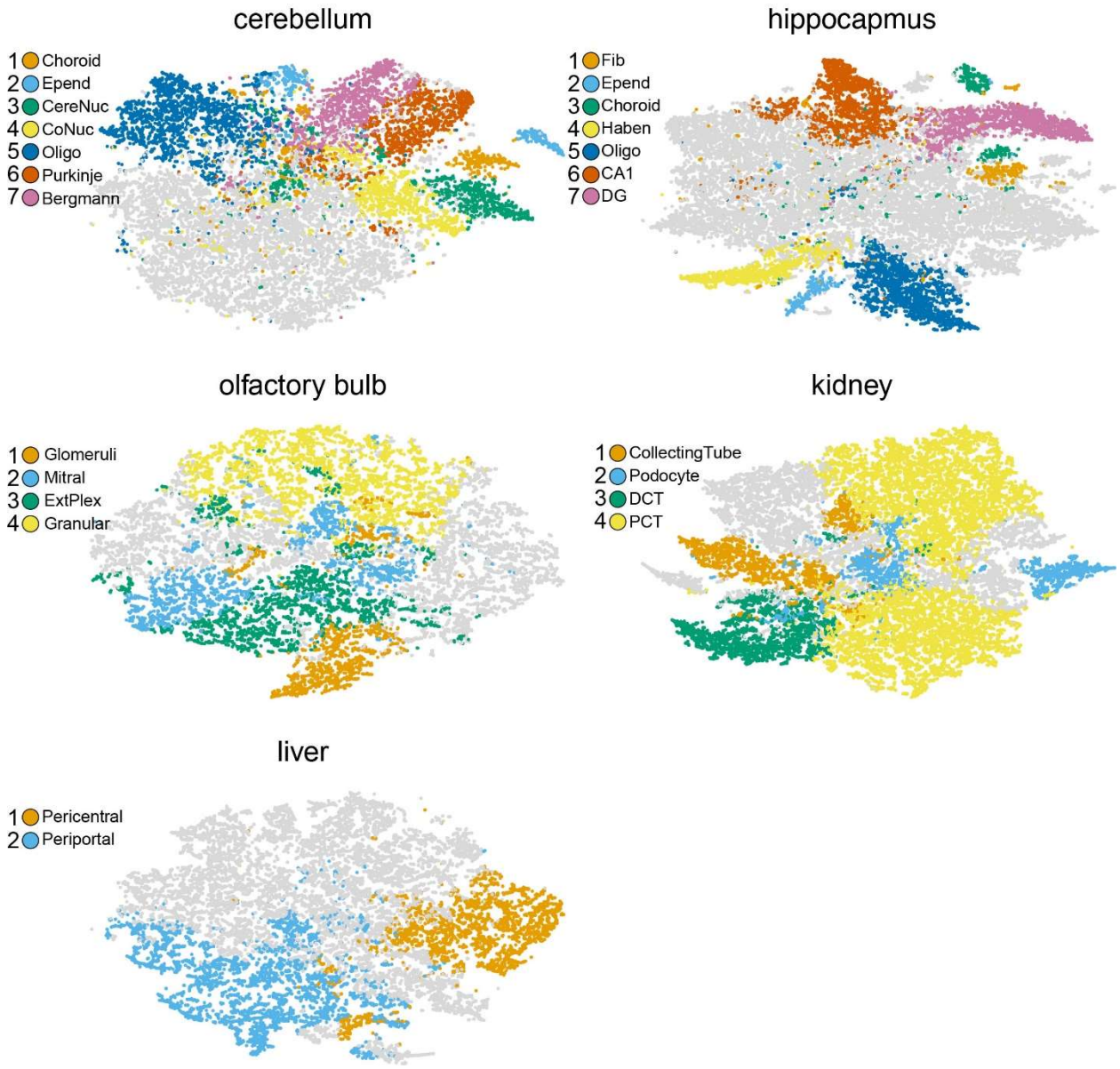

Figure 10-2 Paired tSNE from Slide-seq data for various tissue types: Shown are tSNE embeddings of the tissues assayed in Fig 1C. Coloring of clusters is consistent with Fig 1C. Cluster identities were annotated as follows: Cerebellum: (1) Choroid plexus (2) Ependymal (3) cerebellar nucleus neurons (4) Cochlear nucleus (5) Oligodendrocyte (6) Purkinje cells (7) Bergmann glia. Hippocampus: (1) Fibroblast-like (2) ependymal (3) choroid (4) habenula (5) oligodendrocyte (6) CA1 neurons (7) dentate gyrusneurons. Olfactory bulb: (1) Glomerular layer (2) mitral layer (3) external plexiform layer (4) granule cell layer. Kidney: (1) Collecting tube (2) podocytes (3) Distal convoluted tubule (4) Proximal convoluted tubule. Liver: (1) Pericentral lobule layers (2) periportal lobule layers

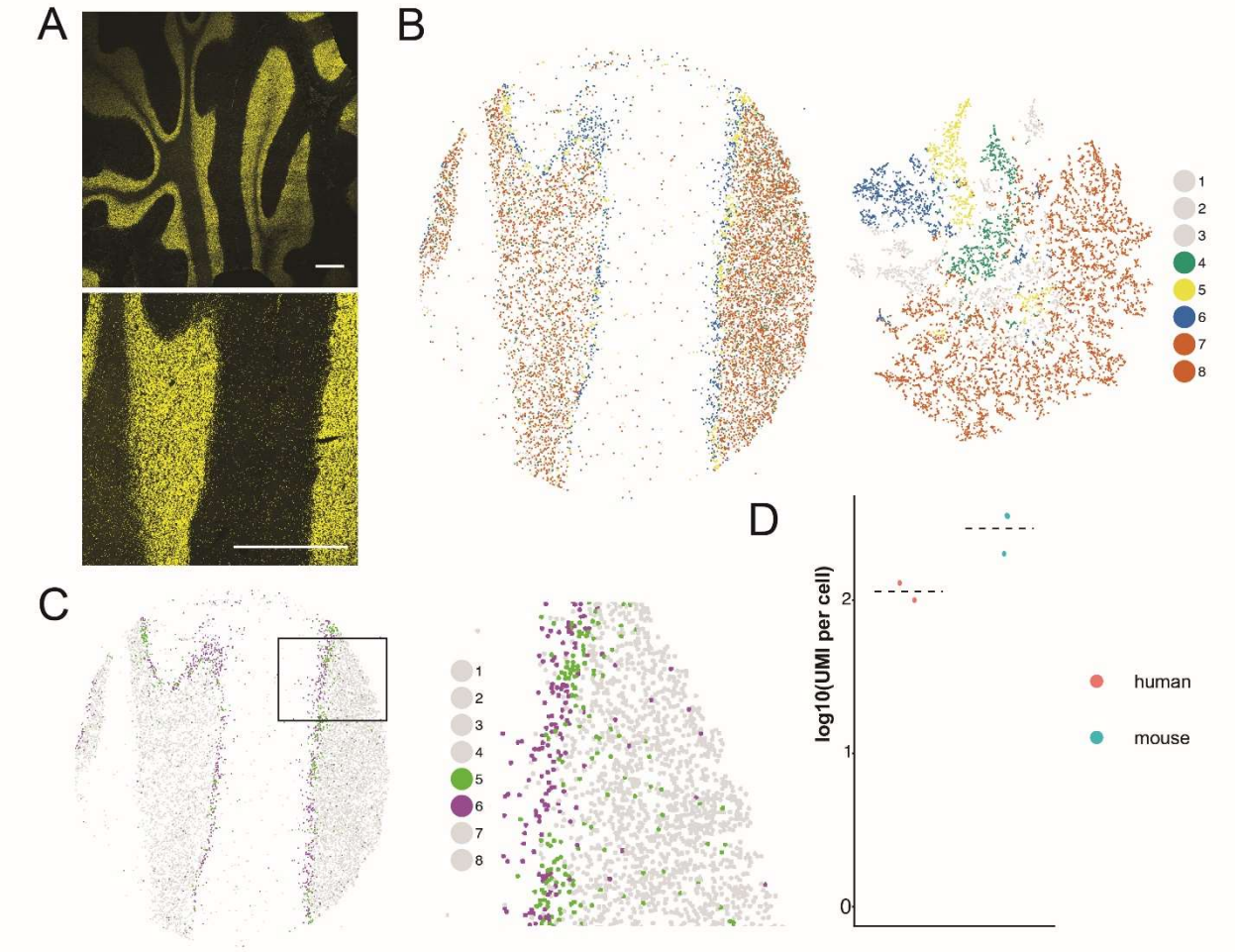

Figure 10-3 (A) Top: DAPI image of 10um section of a human cerebellum (scale bar 2 mm). Bottom: Region of the tissue placed onto a puck (white boxed region in top image, scale bar 2 mm). **(B)** Left: Slide-seq reconstruction of tissue with each bead colored by a cluster label. Right: NMF clustering of beads plotted by tSNE. Cluster identities shown: 4: Oligodendrocytes, 5: Purkinje Neurons, 6: Bergmann Glia, 7: Granular Cells, 8: Granular Cells **(C)** Left: Image in B recolored to highlight the striping pattern of Purkinje Neurons and Bergmann Glia. Right: Magnified image highlighting the alternation between beads called as Bergmann glia (purple) and Purkinje neurons (green), boxed region on left image. **(D)** Comparison of UMI counts per cell between mouse cerebellum (N = 3, 301±88 UMIs, mean ± std), and human cerebellum (N = 2, 115±21 UMIs, mean ± std).

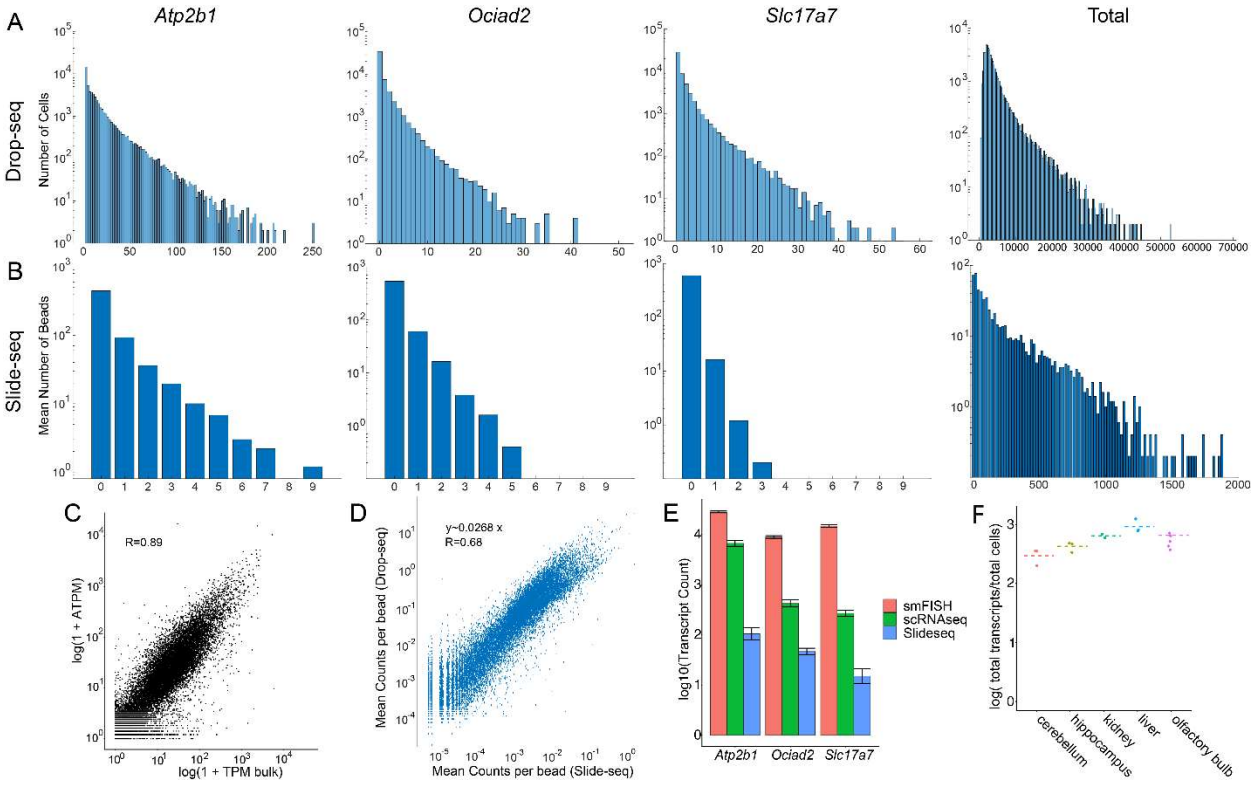

Figure 10-4: **(A)** Histograms of counts of three CA1 marker genes (left and center) and total counts (right) in Drop-seq profiles assigned to a CA1 cell identity using data from Saunders et al. (*24*) **(B)** Gene count distributions on Slide-seq beads in Fig. 2D (mean number of transcripts, averaged over five pucks from Figure 4-2D). **(C)** Comparison of Slide-seq expression data to bulk RNAseq. X axis represents log10(1+TPM) of bulk sagittal hippocampus RNA seq data. Y axis represents log10(1+ATPM) of Slide-seq data, see methods (R = 0.89). **(D)** For ~20,000 genes, the mean counts per cell in CA1-assigned Drop-seq beads is plotted against the mean counts per bead in CA1 Slide-seq beads. Note that although the scatterplot is displayed in log space, the fit was performed in linear space to estimate the efficiency of Slide-seq in comparison to Drop-seq. (I.e. we fit the model y~ax+b, rather than y~a x^b as is standard). A fit performed on log-adjusted transcript counts yielded an R value of 0.68. The slope of 0.0268 in the linear fit suggests that Slide-seq has 2.7% the capture of Drop-seq. **(E)** Comparison of transcript counts of three genes (*Atp2b1, Ociad2, Slc17a7)* across smFISH, scRNAseq, and Slide-seq across a field of view of CA1 (for smFISH and Slide-seq) and for the equivalent number of cells in scRNAseq. **(F)** Quantification of the number of transcripts per cell in Slide-seq data across five different tissues including hippocampus (N = 4, 427±79, mean ± std), cerebellum (N = 3, 302±88, mean± std), kidney (N = 2, 641±64, mean±std), liver (N = 3, 942 ± 255, mean ± std), and olfactory bulb (N = 6, 718±359, mean±std).

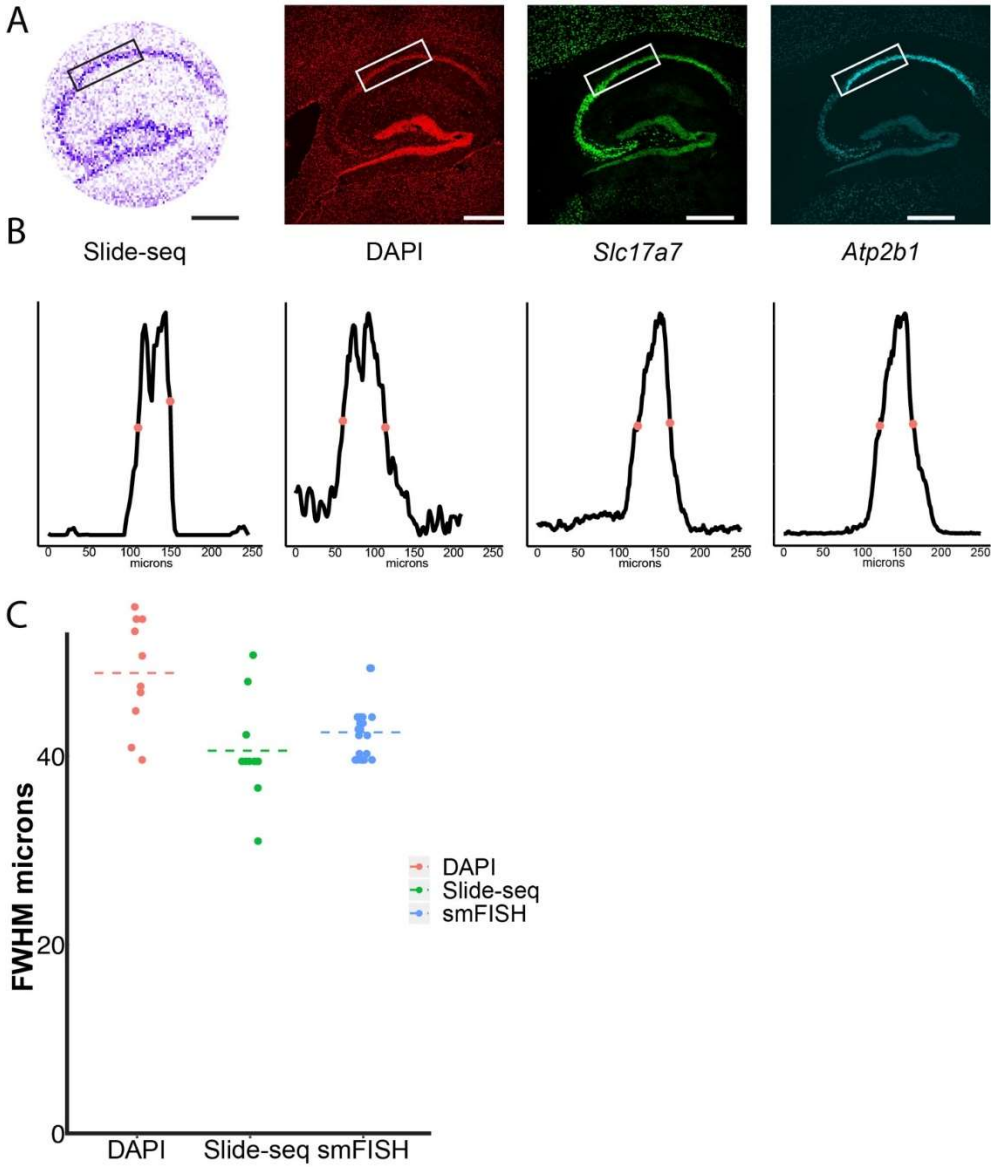

Figure 10-5: (A) Left: Slide-seq reconstruction of mouse hippocampus, shaded by the number of transcripts captured per bead. Middle left: DAPI image of a tissue section adjacent to the Slide-seq puck. Middle right and right: Images of smFISH staining for *Slc17a7* and *Atp2b1* from adjacent section. Box on each image represents a region taken for diffusion analysis. **(B)** Representative plots of the full width at half maximum (FWHM) for the samples above. Red dots represent the half-maximum (see Methods). **(C)** Beanplot of independent FWHM measurements of the CA1 for DAPI, Slide-seq and smFISH. Two CA1 markers were used for smFISH quantification (*Atp2b1* and *Slc17a7*). Dotted line represents mean. Scale bars: 500µm

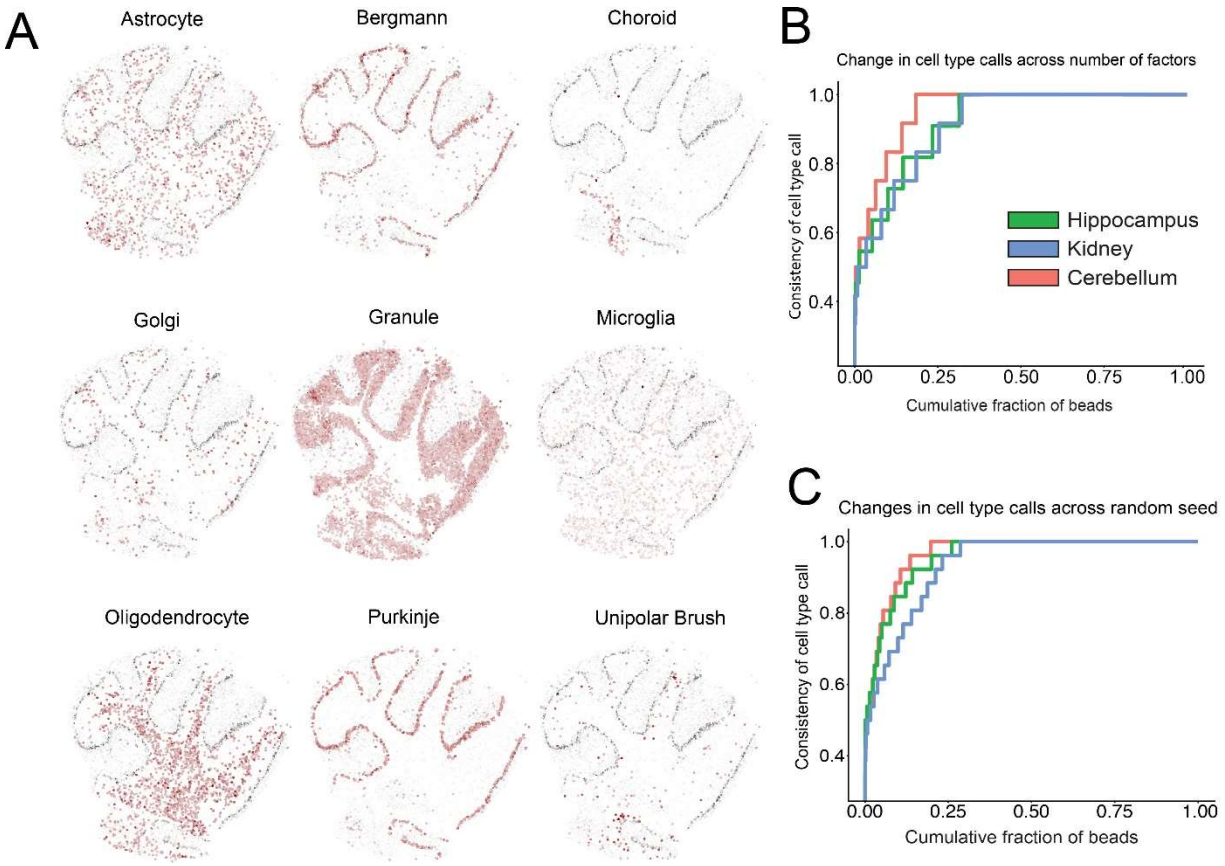

Figure 10-6: (A) Loadings of individual cell types, defined by scRNA-seq cerebellum (*24*) on each bead, as in Figure 4-2B, but for additional cell types. **(B)** Cumulative distribution plot showing the consistency in the bead identities assigned from NMFreg. The consistency is calculated by running NMFreg for 30 values of k (the number of factors) between 18 and 48. For each bead, the consistency is then defined as the fraction of NMFreg runs on which the bead was assigned to the modal cell type across all factors tested. Data is shown across three different tissue types. **(C)** As in B, but here the consistency is calculated by running NMFreg with 30 different random seeds.

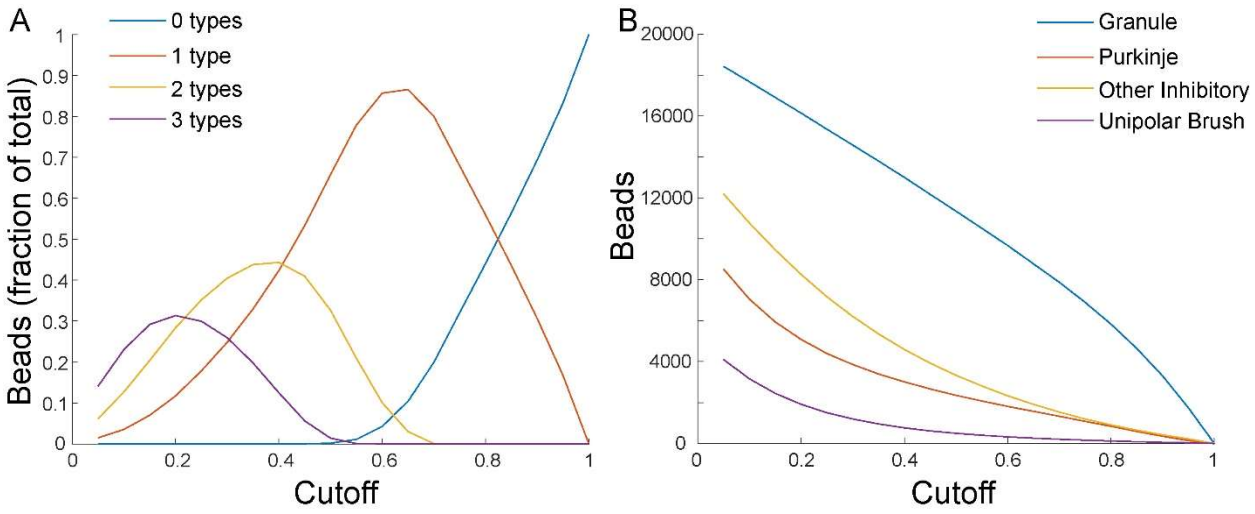

Figure 10-7: (A) A plot of the fraction of beads from cerebellar pucks analyzed in Figure 4-2C, with zero cell types (blue), one cell type (red), two cell types (yellow), or three cell types (purple) as a function of the cutoff C. A cell type is defined to be present on a bead if the L2 norm of the vector of factor loadings mapping to that cell type is greater than or equal to C times the L2 norm of the vector of all factor loadings for that bead. For Figure 4-2C, a cutoff of 0.5 was used. The plot shows mean across seven cerebellar pucks. **(B)** The mean number of beads representing granule cells (blue), Purkinje cells (red), other inhibitory neurons (yellow), and unipolar brush cells (purple) as a function of the cutoff C. The decrease in the number of each kind of cell is roughly linear for C>0.7, but is nonlinear for values of C<0.7, for which multiplets are possible.

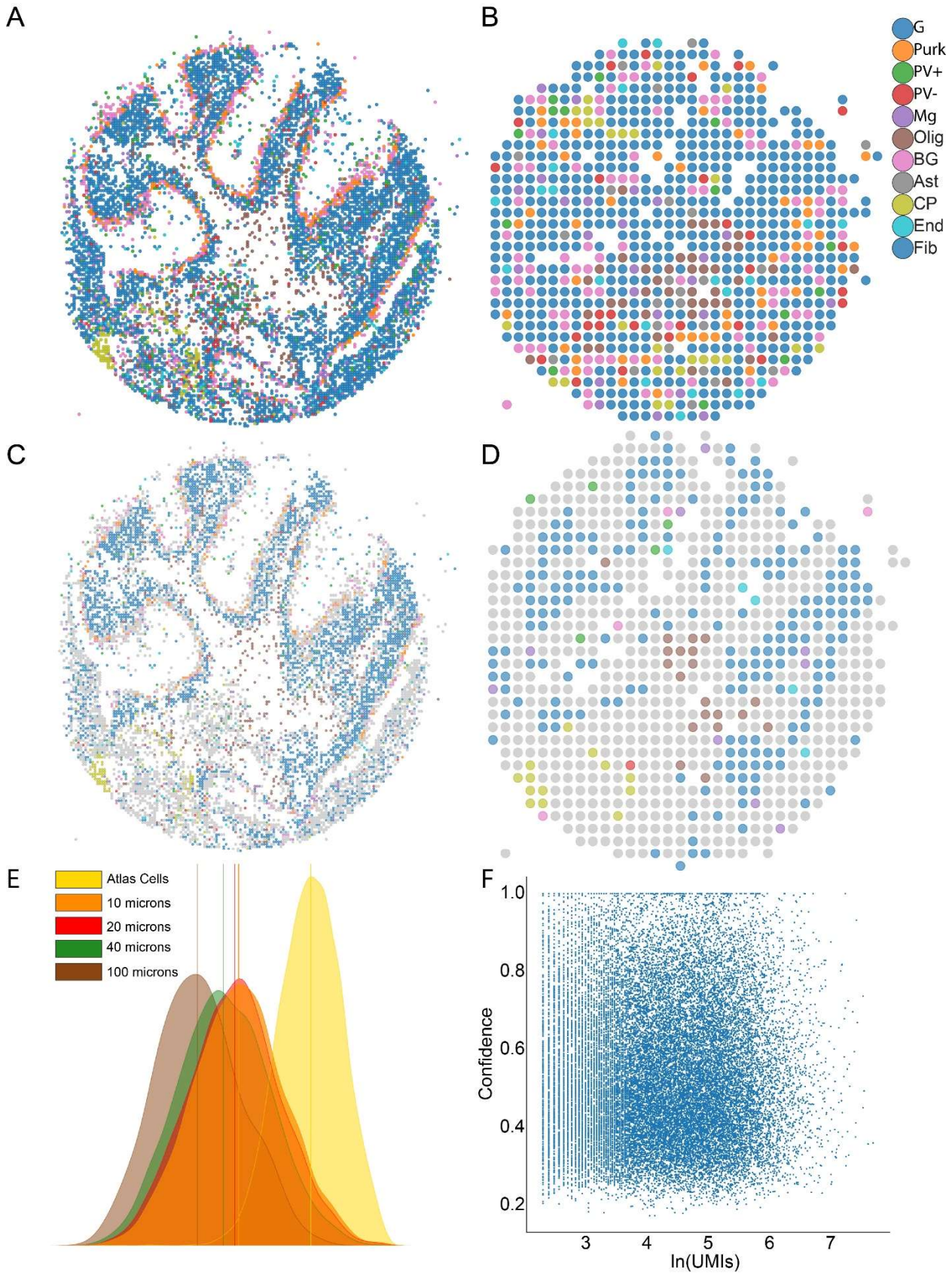
A
B
G
Purk
PV+
PV-
Mg
Olig
BG
Ast
CP
End
Fib
C
D
E
Atlas Cells
10 microns
20 microns
40 microns
100 microns
F
Confidence
1.0
0.8
0.6
0.4
0.2
3
4
5
6
7
ln(UMIs)

Figure 10-8: Analysis of larger feature sizes, aggregated *in silico*. **(A)** All beads were aggregated into 20 µm-diameter features and the resulting features were assigned cell types by NMFreg. Beads are colored according to the cluster to which they were assigned. Legend: G=Granule cells, Purk=Purkinje, PV+=Parvalbumin-positive interneuron, PV-=Parvalbumin-negative interneuron, Mg=Microglia, Olig=Oligodendrocytes, BG=Bergmann Glia, Ast=Astrocytes, CP=Choroid Plexus, End=Endothelium, Fib=Fibroblasts. **(B)** As in (A), but for 100 µm diameter aggregated features. **(C)** Same as (A), but all features that fail to pass the confidence threshold are colored in gray. **(D)** As in (C), but for 100 µm features. Upon aggregating features into 100 µm diameter features, we retain the ability to identify choroid plexus, white matter, and granule cells, but no other cell types with confidence. **(E)** The distributions of L1 norms between the factor loading distributions and the uniform distribution are shown for atlas cells, the original Slide-Seq data (10 µm), 20 µm aggregated features, 40 µm aggregated features, and 100 µm aggregated features, showing the decrease in cell type purity as the feature size increases. **(F)** The number of UMIs (natural log) versus the confidence, defined as the L2 norm of the vector of factors mapping to the cell type as which the bead was called, after normalizing so that the sum of the L2 norms for all cell types is 1. There is no relationship between the number of UMIs and the bead confidence.

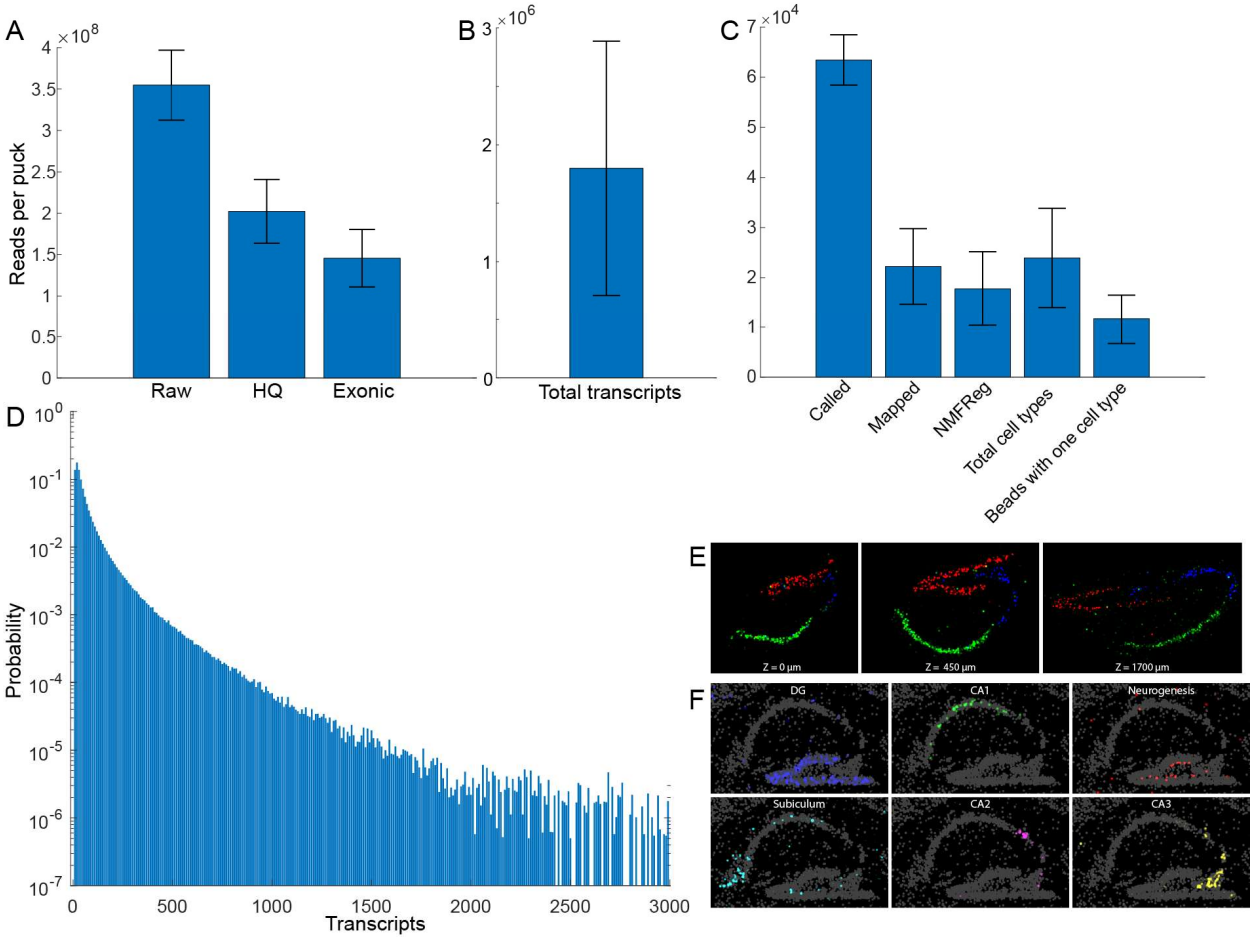

Figure 10-9: (A) The number of raw reads, high-quality reads, and exonic reads per puck for 10 randomly selected pucks from the 66 hippocampal pucks in Figure 4-2D. **(B)** The total number of transcripts per puck for the 66 hippocampal pucks in Figure 4-2D. **(C)** For the 66 hippocampal pucks in Fig. 2D, from left to right, all reported on a per puck basis: the number of beads identified by SOLiD basecalling; the number of SOLiD bead barcodes mapped to Illumina bead barcodes (see "Image Processing and Basecalling", above); the number of bijectively mapped barcodes that were processed by NMFreg (i.e., that had at least 5 variable genes); the total number of cell types passing the 0.5 L2 norm cutoff following NMFreg (Figure 10-7); the number of beads with a single cell type passing the 0.5 L2 norm cutoff. **(D)** A probability density plot (i.e. normalized histogram) of the number of transcripts per bead, averaged over all 66 pucks. All error bars show standard deviation. **(E)** Cell type calls of three representative sections from the dataset with the position on the mediolateral axis denoted at the bottom of the image. **(F)** Metagene profiles on a sagittal hippocampus section representing cell subtypes.

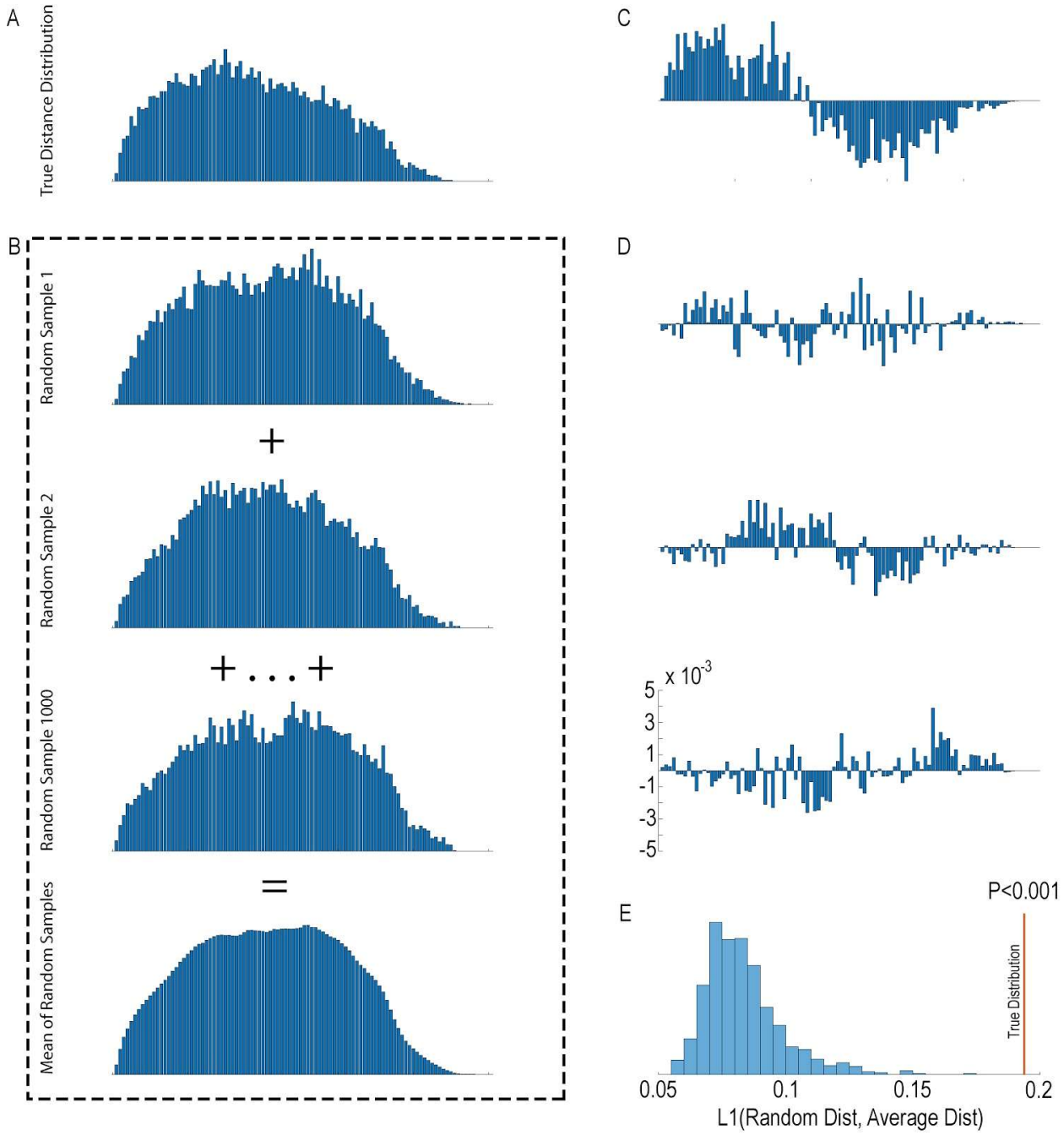

Figure 10-10: Schematic of the algorithm for identifying spatially non-random genes. The algorithm can be run on any specified subset of beads to identify genes with significant nonrandom distribution within that subset. All histograms displayed here are calculated beads defined as granule cells on a coronal cerebellar puck (Figure 4-3A). **(A)** For each gene of interest, we calculate the distribution of the Euclidean distances between all beads in the specified subset expressing at least one transcript of the gene, shown here for *Rasgrf1*. **(B)** We then randomly sample an equivalent number of beads from the subset with probability proportional to the number of reads per bead, without replacement. We perform this sampling 1000 times, and for each sample, calculate the distribution of pairwise Euclidean distances between the beads thus chosen. We take the elementwise mean of all 1000 samples to obtain the average distribution of pairwise

distances across random samples. **(C)** We then take the elementwise difference between the distance distribution for the gene of interest and the average distribution, **(D)** as well as between the distance distribution for each of the random samples and the average distribution. **(E)** A histogram of the sum absolute values of the distributions shown in in (D), i.e., the L1 norm between distance distributions of the random samples and of the average sample. The L1 norm serves as our test statistic: if the gene of interest is distributed proportionally to the number of transcripts per bead, the L1 norm will be uniformly distributed. For *Rasgrf1*, the L1 norm of the true distribution is greater than the L1 norms of any of the random samples, so $p<0.001$ (permutation test, see Methods). (Because there are only 1000 samples for reasons of computational complexity, the smallest observable p value is $p<0.001$).

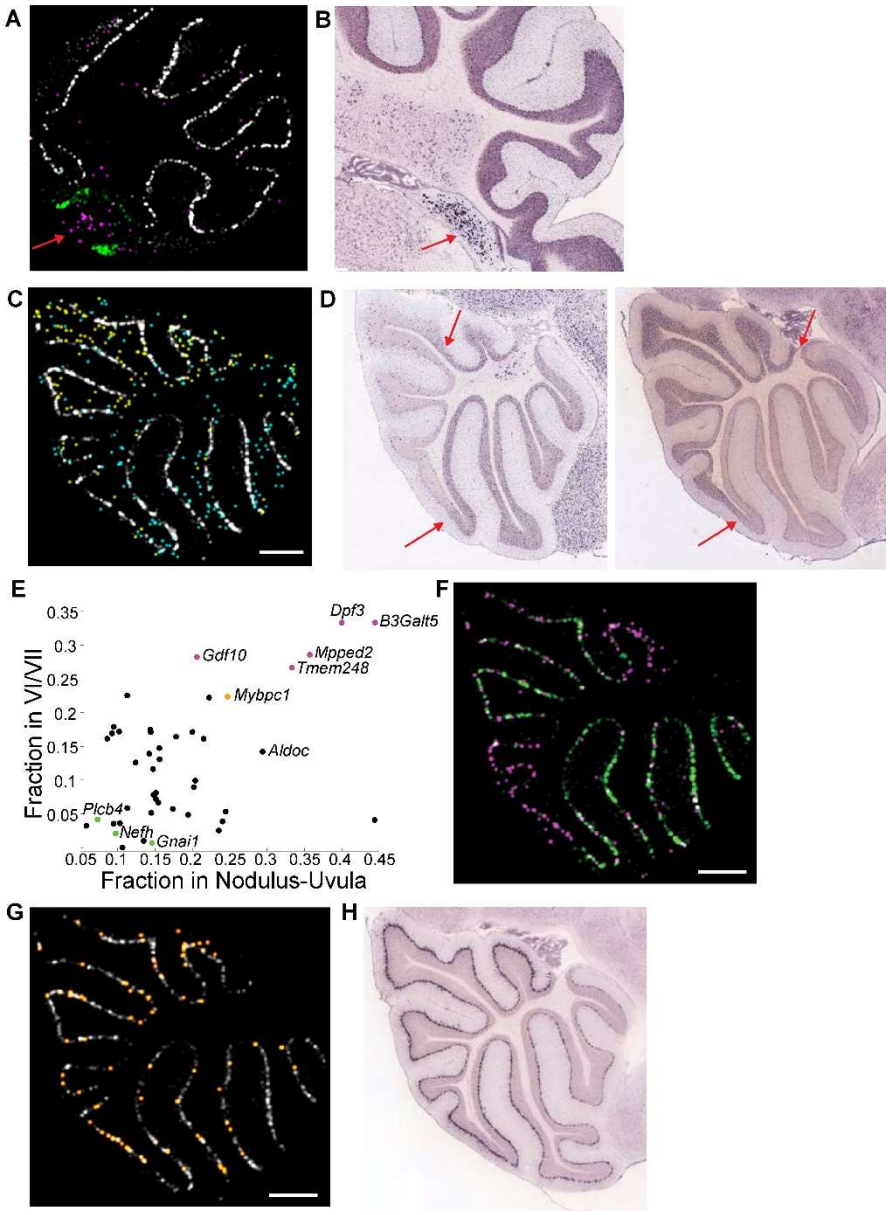

Figure 10-11 : (A) A coronal cerebellar puck is shown, with Purkinje-assigned beads in white, choroid-assigned beads in green, and beads expressing *Ogfrl1* in magenta. Red arrow indicates cluster of *Ogfrl1*-positive beads. **(B)** An Allen Atlas (*38*) *in situ* hybridization atlas image of *Ogfrl1*, from a similar brain region. Red arrow indicates *Ogfrl1* expression in the cochlear nucleus. **(C)** A sagittal cerebellar puck showing counts of *Pcp4* (gray), *Rasgrf1* (blue), and a metagene consisting of *Gprin3*, *Cemip*, *Mab21l2*, and *Syndig1l* (yellow). **(D)** Allen atlas images of *Rasgrf1* (left) and *Gprin3* (right). Arrows indicate point of boundary of expression within the granular layer for each gene. **(E)** As in Figure 4-3B, but for genes with significant expression both in the nodulus ($p<0.05$, Fisher exact test) and the VI/VII boundary ($p<0.05$, Fisher exact test). **(F)** A *Gnai1* metagene in green, and a *B3galt5* metagene in magenta. (G) *Mybpc1* in orange. **(H)** An Allen atlas image for *Mybpc1* (*38*). All scale bars show 250 µm; *Pcp4*, a ubiquitous marker for Purkinje cells, is in gray in (A), (C), (F), and (G). All metagenes are listed in Table 10-2.

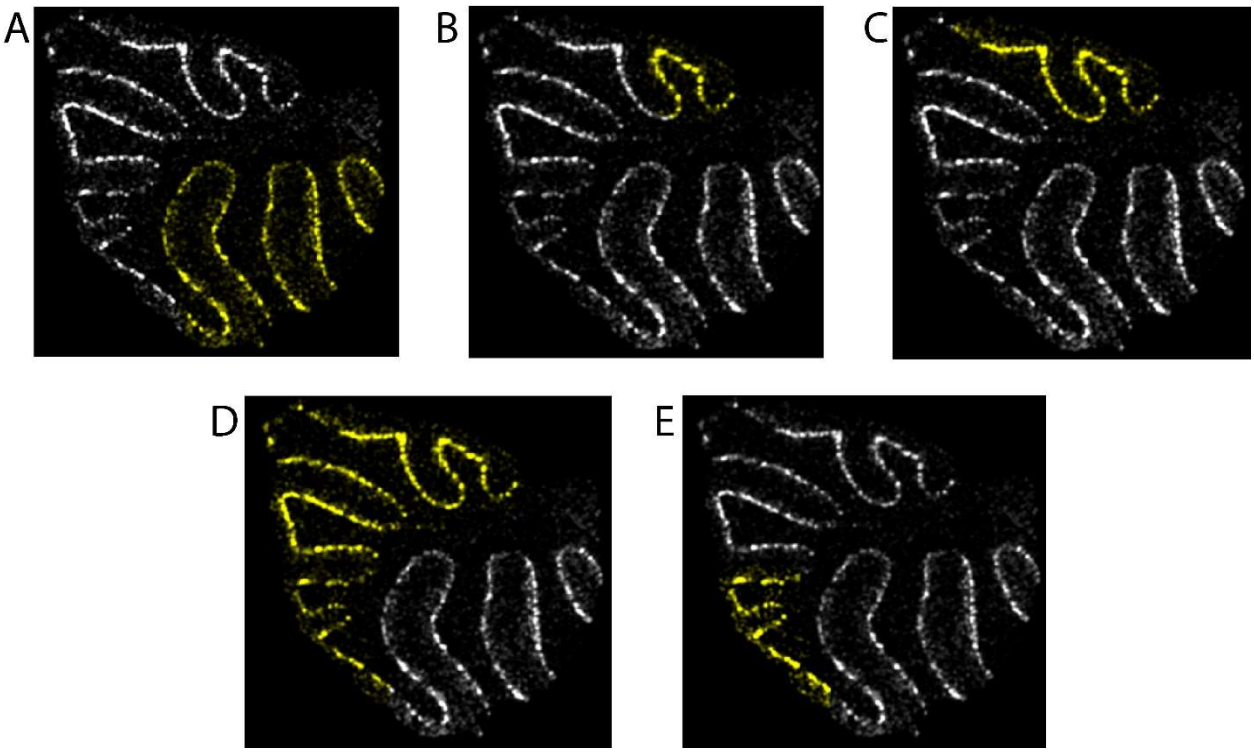

Figure 10-12: Regions chosen for analysis in Figure 4-3. Yellow indicates beads included in the region designation, while white indicates beads excluded from the region. A metagene consisting of Pcp4 and Pcp2 is plotted. **(A)** The dorsal region. **(B)** The nodulus region. **(C)** The nodulus-uvula region. **(D)** The ventral region. **(E)** The VI/VII region.

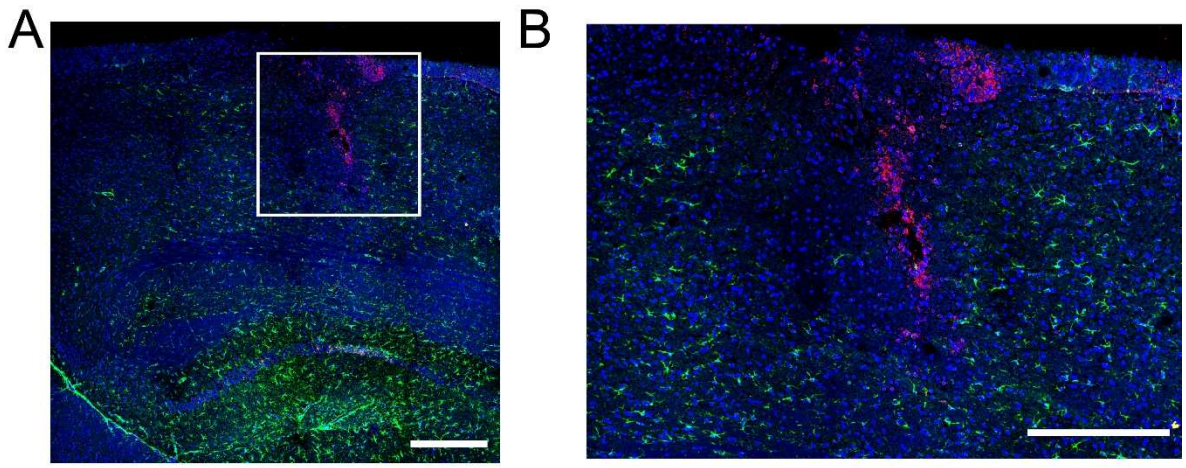

Figure 10-13: (A) Section of sagittal hippocampus at the site of cortical injury 3 days post injury stained with DAPI to stain nuclei (blue), *Gfap* (green), and *Vim* (magenta) revealing the precise location of the injury (white box). **(B)** Magnified image of boxed region in (A). (Scale bars: 500 μm)

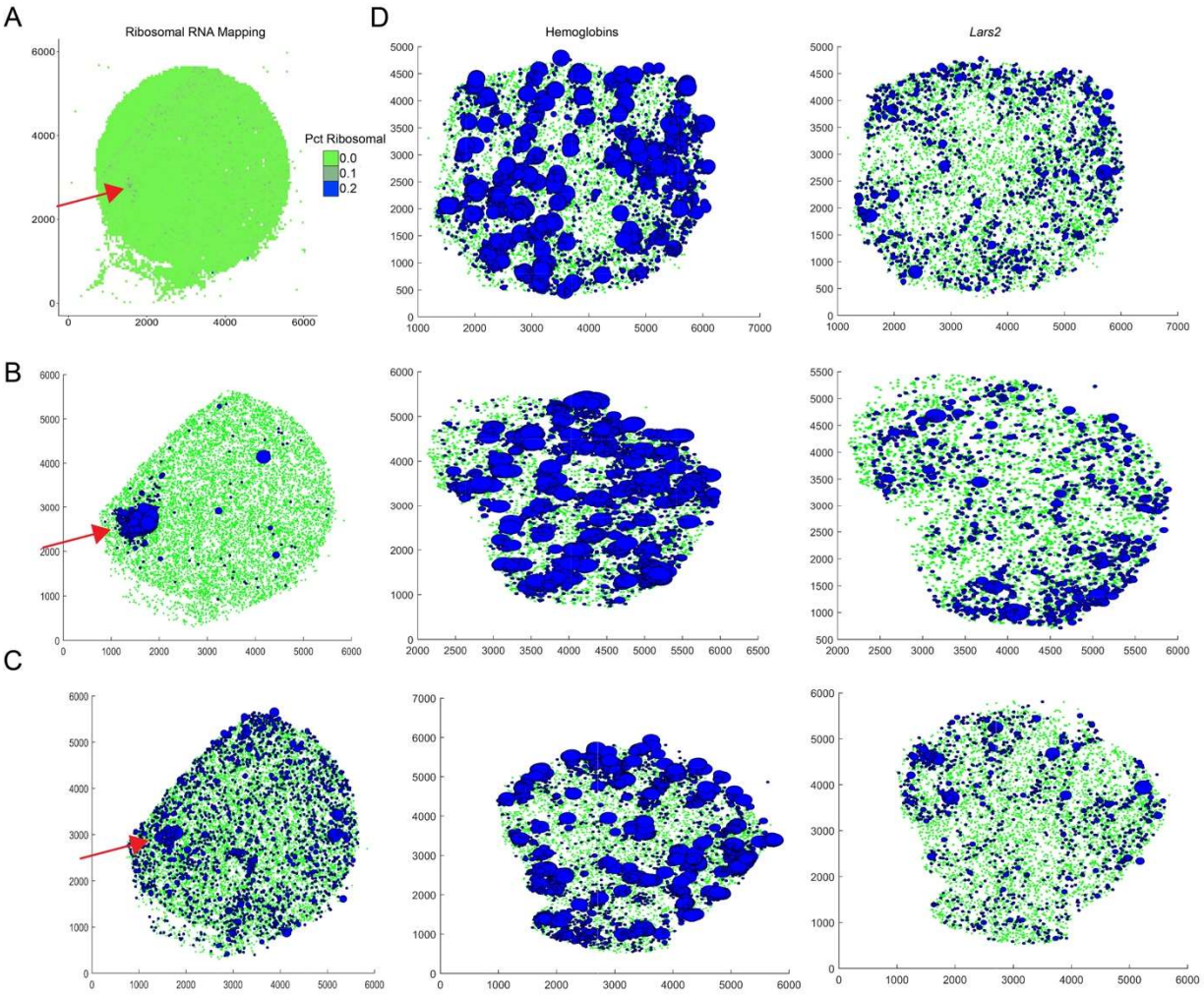

Figure 10-14: (A) Plot showing the percentage of reads at each bead mapping to ribosomal RNA, prior to alignment, for the 180819_3 puck (same as in Figure 4-4) **(B)** Plot showing beads expressing hemoglobins. All beads expressing at least one transcript of *Hba-a1*, *Hba-a2*, *Hbb-bs*, or *Hbb-bt* are shown in blue, with radius proportional to the total number of hemoglobin transcripts. All other genes are shown in green. **(C)** As in B, but for *Lars2* transcripts, which are believed to represent rRNA. (See "Identification of rRNA in pucks" in Methods.) **(D)** Three cerebellar (non-injected) pucks, showing hemoglobin transcripts (left) and *Lars2* transcripts (right). The correlation between hemoglobin and *Lars2* in B and C is in great excess over the correlations observed in D.

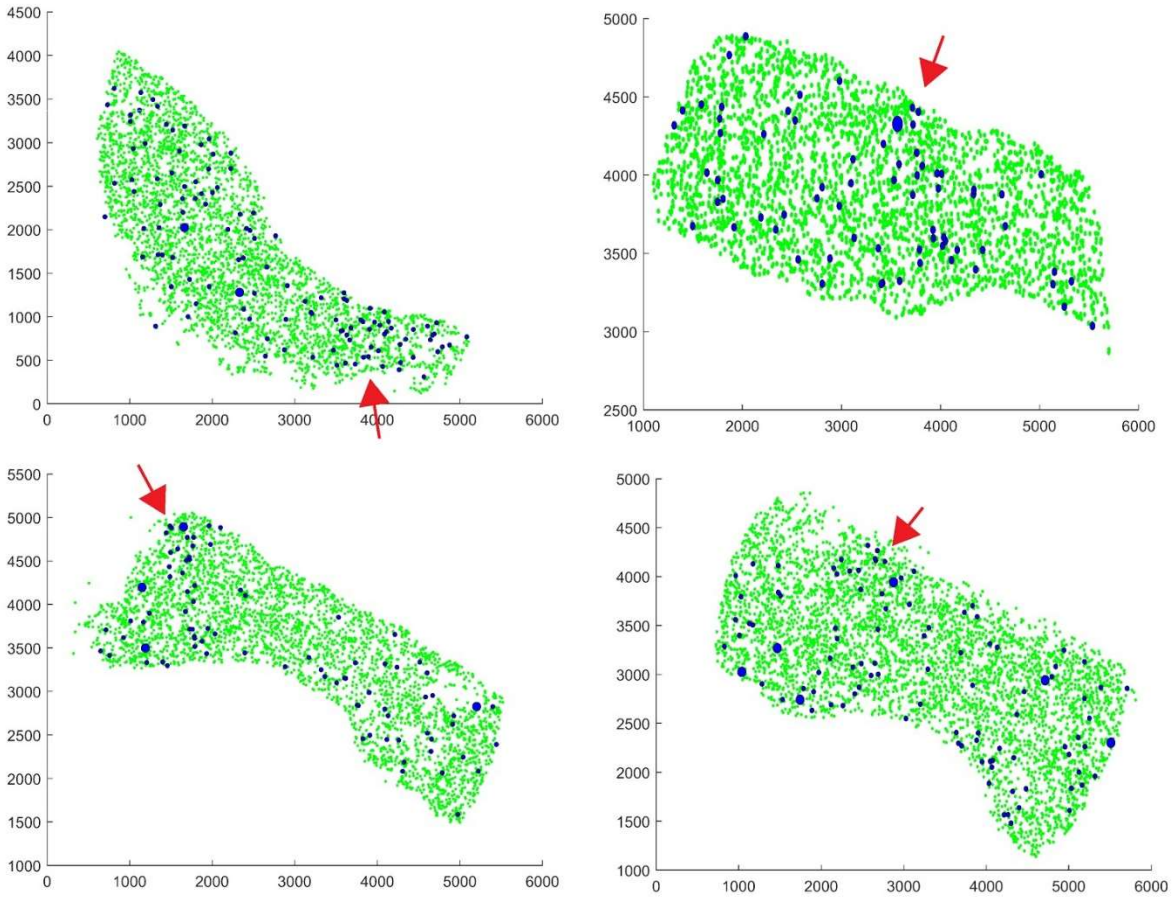

Figure 10-15: Beads expressing *Sox4* and *Sox10* are shown in blue for four pucks from the 2-week injury timepoint. The radius of blue beads is proportional to the total counts of *Sox4* and *Sox10*. The injury site is indicated with a red arrow.

Supplementary Video 1:
A 3D volume rendering of CA1, CA2/3 and dentate gyrus as shown in Figure 4-2. Scale bars: 500 µm.

Table 10-1: Oligonucleotides used in this study. Note r prior to base indicates RNA. + indicates LNA.

Name	Sequence
Truseq5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Smart PCR primer	AAGCAGTGGTATCAACGCAGAGT
Truseq_PCR_handle	CTACACGACGCTCTTCCGATCT
Template Switch Oligo (TSO)	AAGCTGGTATCAACGCAGAGTGAATrG+GrG

Truseq	/5Phos/AGATCGGAAGAGCGTCGTGTAG
Truseq -1	/5Phos/GATCGGAAGAGCGTCGTCTAG
Truseq -2	/5Phos/ATCGGAAGAGCGTCGTGTAG
TruSeq-3	/5Phos/TCGGAAGAGCGTCGTGTAG
TruSeq-4	/5Phos/CGGAAGAGCGTCGTGTAG
UP	/5Phos/TCTCGGGAACGCTGAAGA
UP-1	/5Phos/CTCGGGAACGCTGAAGA
UP-2	/5Phos/TCGGGAACGCTGAAGA
UP-3	/5Phos/CGGGAACGCTGAAGA
UP-4	/5Phos/GGGAACGCTGAAGA

Table 10-2: Gene lists referenced throughout the paper, by figure. All figures without "S" refer to Chapter 4, whereas all figures with "S" refer to this chapter.

Fig. S11C	
Genes enriched posterior of the primary fissure in the cerebellum	Gprin3, Cemip, Syndig1l, Mab21l2
Fig. 2	
Fig. 2E CA3/Hilum (plotted restricted to beads assigned by NMFreg to atlas cluster 6)	Satb1, Scg2, Nap1l5, Fxyd6, C1ql3, Necab, Slc35f1, Nrsn1, Calb2
Fig. 2E CA2 (plotted restricted to beads assigned by NMFreg to atlas cluster 6)	Adcy1, Pcp4, Rgs14
Fig. 2E Subiculum (plotted on all beads)	Rxfp1, Fn1, Lxn, Nr4a2
Fig. 2E CA1 (plotted restricted to beads assigned by NMFreg to atlas cluster 5)	Tenm3, Lypd1
Fig. 2E DG (plotted restricted to beads assigned by NMFreg to atlas cluster 4)	Mef2c
Fig. 2E Neurogenesis	All beads assigned to atlas cluster 13.
Fig. 3	
Fig. 3C Aldoc metagene	Aldoc, Kctd12, and Car7
Fig. 3C Cck metagene	Cck, Stmn4, Kcng4, and Atp6ap1l
Fig. 3D H2-D1 metagene	H2-D1, Cops7a, and Kmt2c
Fig. 3D Hspb1 metagene	Prkci and Hspb1
Fig. S11F Gnai1 metagene	Gnai1, Nefh, Plcb4, Rgs8, Homer3, Scg2, Scn4b, and Gm14033
Fig. S11F B3Galt5 metagene	B3galt5, Gdf10, Tmem248, Mpped2, and Dpf3
Plcb4-associated ATPases and sodium channels	Atp1a3, Atp1b1, Atp2b2, Atp6ap1l, Kcnab1, Kcnc3, Kcng4, Kcnma1. Note that *Kcng4* is associated with increased firing rate in fast motor neurons (*334*), suggesting that its expression contributes to the faster spiking measured in Zebrin II-negative Purkinje neurons (*164*, *335*), while the calcium-dependent channel *Kcnma1* is known to regulate the timing of dendritic calcium burst spiking in Purkinje cells (*336*), suggesting that it contributes to differences in bursting activity previously observed between lobules III-V and X (*337*).
Example genes expressed only in lobule X	Prkci, Prkcd, Hpsb1
Example genes that are expressed everywhere except in lobule X	H2-D1, Cops7a, Kmt2c
669 Candidate Significant Genes	1110001J03Rik 1700020I14Rik 1810037I17Rik 2210016L21Rik 2900093K20Rik AW047730 Abr Acin1 Actb Actr1a Actr3 Actr3b Acyp1 Adam11 Adam23 Add3 Aig1 Akap6 Akap9 Aldh5a1 Aldoc Alkbh7 Ank2 Ankrd12 Anks1b Ap1s1 Ap2a2 Aplp1 Apod Apoe App Appbp2 Ar Araf Arap2 Arfip2 Arhgap20 Arhgap5 Arl2 Arl4a Arpc4 Ascc1 Atp1a2 Atp1a3 Atp1b1 Atp1b2 Atp2b1 Atp2b2 Atp5c1 Atp5d Atp5h Atp5l Atp5o Atp6ap1l Atpif1 Atxn2 Atxn7l3b B230118H07Rik B2m Bag1 Baiap2 Bex2 Bhlhe41 Bloc1s6 Bola3 Brd7 Brd8 Brwd1 Bst2 Btbd17 Bzw1 Bzw2 Cacng2 Calb1 Calm2 Camk4 Capza2 Car2 Car7 Car8 Cbr1 Cbx6 Ccar1 Ccdc115 Ccdc50 Ccdc85b Ccdc88a Cck Cct6a Cd47 Cd63 Cd81 Cdc37l1 Cdc42ep4 Cdk5 Cdkal1 Cds2 Celf4 Cep126 Cept1

	Cerk Cers4 Cggbp1 Chd9 Chga Chn1 Cisd3 Cit Ckap5 Clasp2 Cmtm5 Cnbp Cnot6l Cnp Col18a1 Commd7 Comt Copa Cops3 Cops4 Cops7a Cox14 Cox7a2l Cox8a Cpne2 Cpne9 Cr1l Creg1 Cript Cryab Csnk2a1 Cspg5 Cst3 Ctr9 Cttn Cttnbp2 Cux2 Cystm1 Cyth3 D10Jhu81e Dab1 Dagla Dap Dars Dbi Dclk1 Dcun1d5 Ddx1 Ddx42 Dgcr6 Dgkz Dnaja1 Dnajb2 Dner Dpm3 Dpp10 Dpysl2 Dstn Dtna Dync2li1 Ebf1 Echs1 Eci2 Ednrb Eif1ax Eif3a Eif3d Eif3f Eif4a1 Elmod1 Epb4.1l1 Epc1 Epha5 Ergic2 Erh Ermn Erp29 Etfa Evl Fabp3 Fabp5 Fabp7 Fam107a Fam174a Fam21 Fam98b Fbxl15 Fbxo3 Fbxo9 Fdps Fdx1 Fem1c Fgfr3 Fkbp1a Fkbp3 Fkbp8 Fth1 Fxyd7 Gabra1 Galnt11 Garnl3 Gas5 Gatm Gcsh Ggt7 Glul Gm14033 Gm27199 Gm5083 Gna13 Gnai1 Gnao1 Gnb2 Gng13 Gnl3l Golga4 Golph3 Got1 Gpatch11 Gpbp1 Gpm6b Gpr37l1 Gria1 Gria2 Gria4 Grid2 Grik1 Gsk3b Gstm1 Gtf2b Gtf2i Gucy1b3 Guk1 H2-D1 Hccs Hcfc1r1 Hdgf Hdlbp Hexa Hgsnat Higd2a Hint1 Hlf Hnrnpc Homer3 Hopx Hpcal1 Hprt Hsbp1 Hsd17b12 Hsf1 Hspa12a Hspa14 Hspa4l Hspe1 Hsph1 Hypk Icmt Id4 Ide Ifi27 Ifit3 Ifit3b Ifitm3 Ift57 Ilf2 Iltifb Ina Inpp5a Isca1 Itm2b Itm2c Itpr1 Jkamp Jrkl Kat6a Kcnab1 Kcnc1 Kcnc3 Kcnd2 Kcng4 Kcnma1 Kcnmb4 Kctd12 Khsrp Kif21a Kif3c Kif5c Kitl Klc1 Klhdc2 Kmt2c Krt25 Lamtor5 Lap3 Lars2 Ldha Lgals3bp Lhx1 Lhx1os Lin7a Lpcat4 Lpgat1 Lrrc49 Lsamp Luc7l3 Luzp2 Lztfl1 Macf1 Macrod1 Magoh Malat1 Map1a Map2k1 Map3k12 Mapk8ip2 Mapre2 Mapt March6 Mbnl2 Mbp Med8 Mef2a Meg3 Megf9 Mgst3 Mif Mipep Mir6236 Mkrn1 Mlec Mllt6 Mobp Morf4l2 Morn2 Mplkip Mrpl16 Mrpl35 Mrpl45 Mrps2 Mrps31 Msi1 Msi2 Msl3 Mt1 Mt2 Mt3 Mtdh Mtfmt Mtss1 Myo5a N6amt2 Nae1 Napg Nat8l Ncoa7 Ncor2 Ndufa11 Ndufa13 Ndufa2 Ndufa3 Ndufa4 Ndufa9 Ndufb2 Ndufb3 Ndufb4 Ndufb5 Ndufb8 Ndufb9 Ndufc1 Ndufc2 Ndufv1 Nefh Nefl Nefm Nnat Nomo1 Nop10 Npas3 Npc2 Npepps Nptx1 Npy Nr2c2 Nrsn1 Nrxn1 Nrxn2 Nsg1 Nt5c Ntrk2 Ntsr2 Nucks1 Oaz1 Oaz2 Ogfrl1 Olfm1 Omg Opa1 Opcml Opn3 Osbpl6 Ostc Pabpc1 Paip1 Pak1 Park7 Patz1 Pax6 Pbrm1 Pbx1 Pcdh17 Pcmt1 Pcp2 Pcp4 Pdcl Pde5a Pdhb Pdia3 Pdlim2 Pex13 Phip Pi4k2a Picalm Pigk Pigs Pisd Pitpnc1 Pja2 Plcb4 Plekhb1 Plekhb2 Plekhd1 Plp1 Pltp Pmm1 Pnn Pno1 Polb Polr2b Ppa1 Ppm1l Ppp1r11 Ppp1r12b Ppp1r17 Ppp2r2b Prdx1 Prdx3 Prdx5 Prdx6 Prex1 Prex2 Prkcd Prkcg Prkg1 Prkrir Prpf6 Psd2 Psma2 Psma3 Psmb10 Psmd8 Ptgds Ptpmt1 Ptpn11 Ptpn4 Ptprr Puf60 Pura Purb Pvalb Pxmp2 Qdpr Qk Rab24 Rabep1 Rabgap1l Rad23a Rad23b Ramp1 Ran Rasa2 Rasa3 Rbm5 Reep1 Rftn2 Rgs7bp Rgs8 Rims4 Riok2 Rit2 Rn18s-rs5 Rnf13 Rnf167 Rora Rpl14 Rpl18 Rpl34 Rpl38 Rpl41 Rps15a Rps21 Rps28 Rragc Rrp1 Rtfdc1 Rtn4 S100b Sac3d1 Saraf Scaf11 Sccpdh Scg2 Scn2a1 Scn4b Sdc3 Sdc4 Sdhc Senp2 Sep15 Sepp1 Sept11 Sept4 Sept7 Serbp1 Serinc1 Setd7 Sfxn4

	Sigmar1 Slc13a5 Slc1a2 Slc1a3 Slc1a6 Slc24a2 Slc25a18 Slc25a39 Slc25a5 Slc33a1 Slc35a5 Slc38a1 Slc4a3 Slc4a4 Slc5a1 Smarca4 Smarcc1 Smpd1 Snap25 Snap47 Snapc3 Sncb Snhg11 Snrk Snrpn Snx24 Socs7 Sox9 Sparc Sparcl1 Spcs2 Sphkap Spock1 Spock2 Spred1 Srp9 Srsf2 Steap2 Stip1 Stk17b Stmn1 Stmn2 Stmn3 Stmn4 Strn3 Stt3b Stub1 Suclg1 Supt6 Sycp1 Syt2 Syt4 Syt7 Tardbp Tbc1d15 Tceb3 Tcf25 Tex261 Thy1 Thyn1 Timm10b Timm17b Tinf2 Tiprl Tln1 Tmed3 Tmed7 Tmeff2 Tmem11 Tmem158 Tmem167 Tmem184c Tmem255a Tmem47 Tmem50a Tmem50b Tmem64 Tmf1 Tmsb4x Tnik Tnrc6b Tomm22 Tomm40l Tpi1 Trf Trim2 Trp53bp1 Trpc3 Tsfm Tshz2 Tspan13 Tspyl4 Tst Ttc14 Ttc3 Ttl Ttyh1 Tuba1a Tubb2a Tubb2b Tubb4a Tubb5 Tulp4 U2af2 Ubap2l Ubb Ube2q1 Ube3a Ubfd1 Ubl5 Ubl7 Ublcp1 Uchl3 Ufc1 Upf2 Uqcr11 Uqcrb Uqcrh Usp14 Usp3 Usp33 Vcpip1 Vimp Vps26b Vps41 Wbp5 Wbscr22 Wdr33 Wdr7 Wwp1 Xrcc4 Ylpm1 Ywhah Zbtb20 Zcrb1 Zfc3h1 Zfp512 Zfp608 Zfp87 Zfr Zic1 Zmat2
Plcb4-Associated Genes	Anks1b Atp1a3 Atp1b1 Atp2b2 Atp6ap1l Baiap2 Car8 Cck Cerk Chn1 Cops7a Garnl3 Gm14033 Gnai1 Golga4 Gria2 Grid2 H2-D1 Hdlbp Hnrnpc Homer3 Hpcal1 Hspa12a Icmt Ina Kcnab1 Kcnc3 Kcng4 Kcnma1 Kitl Kmt2c Lpgat1 Macf1 Mbnl2 Mef2a Msl3 Ndufb8 Nefh Nefm Nptx1 Pde5a Pja2 Plcb4 Pno1 Prdx5 Prkrir Qdpr Rabep1 Rgs7bp Rgs8 Riok2 Scg2 Scn4b Snhg11 Spock2 Stmn2 Stmn4 Strn3 Supt6 Thy1 Tmem50b Tmem64 Trim2 Tspan13 Ttc3 Vps26b Wdr7 Wwp1 Zbtb20
Aldoc-Associated genes	Actb Aldoc Apoe Atp1a2 Atp1b2 Atp5l Atpif1 B230118H07Rik B2m Car7 Cd63 Cd81 Cdc42ep4 Cox14 Cpne9 Cst3 Dbi Dpm3 Dtna Ednrb Fam107a Fam98b Fth1 Glul Gpm6b Gpr37l1 Gria1 Gstm1 Hint1 Hopx Kctd12 Kif5c Mt1 Mt2 Mt3 Ndufa3 Ndufb4 Nomo1 Park7 Pigs Prdx6 Rpl34 Rpl38 Rpl41 S100b Sepp1 Sept4 Slc1a3 Sox9 Sparc Sparcl1 Suclg1 Tmem47 Tmsb4x Trf Tuba1a Zcrb1
Genes with $p<0.001$ (Fisher exact test)in the ventral part of puck 180819_12 compared to the dorsal part, and with greater than 80% of their counts in the ventral region.	Th Cemip Gprin3 Mab21l2 Syndig1l Hbb
Genes with $p<0.001$ (Fisher exact test) in the nodulus-uvula region of puck 180819_12 (i.e. all genes appearing in Fig. S3B, except Kctd12 and Car7)	Aldoc Cacng4 Calm1 Calm2 Car8 Ccdc23 Cck Creg1 Cst3 Fabp7 Homer3 Hspb1 Idh3b Irs2 Malat1 Ngdn Plcb4 Prkcd Prkci Prpf31 Pvalb Rgs8 Slc1a6 Slc25a4 Sparc Stmn4 Ttr Uchl1 mt-Cytb mt-Rnr1 mt-Rnr2
Genes with $p<0.05$ (Fisher exact test) in the nodulus and $p<0.05$ (Fisher exact test) in the VI/VII region of puck 180819_12 (i.e., all genes appearing in Fig. S11E).	Actb Aldoc B3galt5 Calm1 Car8 Cck Cdk5rap2 Chmp4b Cops3 Dbi Dpf3 Efr3a Eif5a Etfa Gad1 Gdf10 Gnai1 Gstm1 Homer3 Idh3g Itm2c Mpped2 Mybpc1 Nefh Nsg1 Plcb4 Ppp1r17 Pvalb Rabep1 Rgs8 Rims2 Rpl13 Sfxn1 Slc1a3 Sox9 Spock2 Timp4 Tmem248 Ttr Ufc1 Wbp2 Ywhah mt-Cytb mt-Rnr1 mt-Rnr2
Fig. 4	

Genes correlating with Vim, Ctsd, and Gfap at the 3 day timepoint.	Camk2n1 Ctsd H2-T22 Hexb Lcn2 Lgals1 Mthfd1 Slc16a11 Pvrl3 Ttr Ctss Dbi Dhrs1 Fabp7 Gfap Mgp Mrps6 Mt2 Nupr1 Pea15a Pold4 Sdc4 Smc4 Trim30a Tspo Vim Vip B2m C1qc Fam124a Fth1 Gcnt2 Gzf1 Ifi27l2a Ifitm3 Myo6 Rpl22 Serpina3n Tnfaip8 Uimc1 Usp12 Vamp8 Xaf1 Ccdc115 Igfbp2 Igfbp7 Ubap2 Eif2ak2 2010111I01Rik Ccnd1 Cnot6l Efcab14 Gbp7 Maged2 Med17 Nfkbia Pabpc1 Rgs8 Rpl10a Smc2 Ugt8a Dclk3 Rnase4 Wnt7b Plp1 Trf Irf9 Rhoc S100a16 S100a6 Srgn Actb Apod Arpc1b Bcas1 Car2 Cldn11 Cnp Cplx3 Enpp2 Ermn Fam46a Gjc3 Grb14 Id1 Id3 Ifi27 Ifit1 Ifit3 Igfbp5 Irgm1 Isg15 Itgam Itm2b Lrp4 Lta4h Mag Mal Malat1 Mbp Mgst1 Mobp Mt1 Nipbl Psmb8 Pvrl1 Rhog Siglech Tppp3 Traf7 Fgfbp3 Creld2 Kcnip2 Msl3l2 Nfkb1 Nkd1 Stat3 Abca1 Aif1 Apbb1ip C1qa C1qb Calb1 Clic1 Cpne6 Crip1 Ctsb Cx3cr1 Cyba Dcps Fcer1g Ftl1 Fyb Gm14295 Grn H2-D1 H2-K1 Hba-a1 Hba-a2 Hbb-bs Hbb-bt Heg1 Hpgd Lcorl Lgals9 Ly86 Mpeg1 Msn Myl12a Myo1c Ncf1 Nes Nfe2l2 Nptxr Pkn1 Plek Ptbp3 Pycard Rn18s-rs5 S100a11 Slc44a2 Sparc Tle1 Tuba1c Tyrobp Uaca Vcan Xpnpep3 Igfn1 Lars2 Pdlim4 Prdx6 S100a13 Sept11 Sorbs1 Syt17 Tmem176b Aco1 Agtrap Bst2 Cald1 Cd63 Cd81 Chd1l Ctdspl Gbp3 Npas3 Ptpn13 Cd52 Ilk Pou2f2 Stat1 Ybx1 Ccnd2 Ctsz Nek6
Genes correlating with Vim, Ctsd, and Gfap at the 2 week timepoint.	1500015O10Rik 1700017B05Rik 1700047M11Rik 1810058I24Rik 2610015P09Rik 2810474O19Rik 3830403N18Rik 4632428N05Rik A2m AF251705 AW112010 Abca9 Abcb1a Abcd1 Abhd12 Abhd4 Abi3 Acads Acer3 Adam10 Adam17 Adamts1 Adamtsl4 Adap2 Add3 Adgre1 Aebp1 Afap1 Aff1 Agps Ahnak Ahr Aim2 Akap12 Akap13 Aldh16a1 Aldh1a1 Aldh2 Anapc7 Ang Angpt1 Ankrd13a Anxa2 Anxa3 Anxa4 Anxa5 Aplp1 Apobec1 Apobec3 Apoc1 Apoe Aqp4 Arap1 Arhgap17 Arhgap29 Arhgap30 Arhgdib Arrdc4 Arvcf As3mt Ascc2 Aspa Atf3 Atp1a2 Atp1b3 Atp6v0e Axl Bach1 Bcl2a1b Bfsp2 Bgn Bhlhe41 Bin1 Bin2 Blvrb Bmp2k Brd7 Bri3 Btg1 C3ar1 C4b Calr Capg Capns1 Carf Carhsp1 Casp8 Cav2 Ccdc13 Ccdc50 Ccdc74a Ccl3 Ccl4 Ccl5 Ccl6 Ccl9 Ccpg1os Cd14 Cd151 Cd164 Cd180 Cd302 Cd37 Cd44 Cd48 Cd53 Cd68 Cd74 Cd82 Cd83 Cd84 Cd86 Cd9 Cdc42ep4 Cdc42se1 Cdkn1c Cebpa Cebpg Cela1 Cenpb Cfh Cflar Cgnl1 Ch25h Chd4 Chst2 Clec5a Clec7a Clic4 Clmp Clu Cnn3 Cntrl Col12a1 Col1a1 Col1a2 Col27a1 Col3a1 Col4a2 Col5a1 Col6a1 Col9a3 Colec12 Colgalt1 Commd10 Coro1b Cotl1 Cpe Cped1 Cpne3 Cpq Cpt1a Cpxm1 Creg1 Crlf2 Crot Cryab Cryba4 Csf1 Csf1r Csf2rb Csrp1 Cst3 Cst7 Cstb Ctdsp2 Ctnna1 Ctnnb1 Ctsa Ctsc Ctsh Ctsk Ctsl Cttnbp2nl Cxcl14 Cxcl16 Cyb5r3 Cybb Cyfip1 Cyp4f14 Cyth3 Cyth4

	Dab2 Dcn Ddah2 Ddr1 Diap2 Dio2 Dnase2a Dnm2 Dock1 Dock10 Dpp7 Dtx3l E130114P18Rik Edem1 Edn3 Ednrb Eef1a1 Eef1d Eef2 Ehd4 Eif3a Elf1 Elk3 Emid1 Eml4 Emp3 Endod1 Entpd1 Epas1 Epb4.1l2 Erbb2ip Erp44 Eya3 Ezr F11r Fabp5 Fam107a Fam114a1 Fam114a2 Fam46c Fblim1 Fbln1 Fbn1 Fcgr1 Fcgr2b Fcgr3 Fcho2 Fcrls Fermt3 Fgfr1 Fkbp7 Fli1 Flt1 Fmnl2 Fn1 Fnbp1 Fnip2 Foxc1 Foxo4 Frmd4a Fstl1 Fuca1 Fxyd1 Fxyd5 Gabarap Galnt10 Gatm Gbp2 Gcn1l1 Ghdc Gjb2 Gltp Glul Gm13139 Gm2a Gm973 Gna12 Gnai2 Gnb2l1 Gng12 Gng5 Gngt2 Gns Golim4 Golm1 Gpm6b Gpnmb Gpr183 Gpr34 Gpr37 Gpt Gpt2 Gpx1 Gsap Gsn Gstm1 Gstp1 Gucd1 Gusb Gyg H2-Aa H2-Ab1 H2-DMa H2-Eb1 H2-T23 H3f3b Hbegf Hdlbp Hes6 Hexa Hist1h1c Hist1h2bc Hk2 Hmha1 Hmox1 Hpgds Hrsp12 Hsd17b11 Hsd3b7 Hsp90b1 Hspb6 Hspb8 Hvcn1 Ifi30 Ifi35 Ifih1 Ifit2 Ifit3b Ifitm2 Ifnar1 Ifnar2 Ifngr1 Igbp1 Igf1 Igf2 Igfbp3 Ikbkb Il10rb Il21r Il33 Il6st Inpp5d Inppl1 Ipo8 Iqce Iqgap1 Irf8 Islr Itga6 Itgav Itgb1 Itgb3bp Itgb5 Itih5 Kcnj10 Kctd12 Kctd5 Kdm5a Kif5b Klf2 Klhl36 Klhl5 Klk6 Krcc1 Lactb Lactb2 Lair1 Lamb1 Lamb2 Lamc1 Lamp1 Lamp2 Lap3 Laptm4a Laptm5 Lat2 Lats2 Lcp1 Lgals3 Lgals3bp Lgmn Lhfpl2 Lilrb4 Lima1 Limch1 Lipa Lmo2 Lpar1 Lpcat1 Lpl Lrp10 Lsp1 Lsr Ltbr Ly6e Lyn Lyz2 Maf Mafb Magoh Magt1 Maml2 Man2b1 Map4k4 Marcks Matn4 Mcl1 Mdk Metap2 Mfap1b Mlc1 Mmp14 Mob1a Mob3b Mob3c Mog Mrpl52 Ms4a6c Msx1 Mt3 Mtdh Myh9 Mylip Myo18a Myo1f Myo9b Myoc Myof Naglu Nagpa Nbl1 Ncf2 Nckap1l Ncl Ndrg1 Neat1 Nek7 Nek9 Nfe2l3 Nfia Nhlrc3 Npc2 Npm1 Nrp1 Nrp2 Ntpcr Oard1 Oat Olfml1 Olfml3 Olig1 Opalin P2rx4 P2ry12 P2ry13 P4hb Pacsin3 Padi2 Palld Parp3 Pbrm1 Pbx3 Pbxip1 Pdcl Pde3b Pdgfra Pdia3 Pdlim2 Pdlim5 Pdpn Pex19 Pfn1 Phkg1 Phldb1 Phldb2 Pla2g15 Pla2g16 Pla2g7 Pld4 Plekhb1 Plekhf2 Plgrkt Plin2 Pllp Plod3 Pltp Plvap Plxdc2 Plxnb2 Pmp22 Ppap2b Ppfibp2 Ppp1r14b Ppp1r18 Prdx1 Prex1 Prex2 Prkcd Psap Psen1 Psme2b Ptgds Ptma Ptn Ptp4a2 Ptpn1 Ptpn18 Ptpn6 Ptprb Ptprc Ptprz1 Ptrf Ptrh1 Qdpr Qk Rab3il1 Rac2 Rad9a Ramp2 Rarres2 Rasgrp3 Rassf2 Rassf4 Rbms1 Rcan3 Rcn3 Reep3 Rel Renbp Rest Rgl2 Rgs10 Rgs5 Rhoa Rhoj Rhoq Rlbp1 Rnaset2a Rnaset2b Rnf130 Rnf141 Rnf213 Rock1 Rpl13a Rpl18 Rpl18a Rpl23 Rpl26 Rpl32 Rpl35a Rpl37 Rpl37a Rpl39 Rplp0 Rplp1 Rplp2 Rps10 Rps11 Rps14 Rps15a Rps20 Rps24 Rps26 Rps27l Rps3 Rps5 Rps9 Rras Rrbp1 Rtp4 Rufy1 Runx1 S100a1 S100a10 S100a4 S100b Sall1 Samd9l Samhd1 Samsn1 Sat1 Scamp2 Scara3 Scarb2 Scd1 Scd2 Scpep1 Scrg1 Sdc3 Selplg Sepp1 Sept10 Serinc3 Serpinb9 Serpine2 Serpinf1

	Serpinh1 Sfrp4 Sgk1 Sgpl1 Sh3bp2 Sh3d19 Sh3glb1 Sh3pxd2a Sirpa Sirt2 Slain2 Slc11a1 Slc12a2 Slc14a1 Slc15a3 Slc16a1 Slc16a2 Slc1a2 Slc1a3 Slc25a10 Slc25a15 Slc25a18 Slc26a2 Slc29a3 Slc38a6 Slc39a1 Slc44a1 Slco2b1 Slfn5 Smarca5 Smg8 Smim3 Snhg18 Snx18 Snx5 Soat1 Sowahc Sox10 Sox12 Sox4 Sp1 Sp100 Sparcl1 Spata13 Spi1 Spp1 Spsb1 Sspn St3gal6 Stat2 Stat6 Stx2 Sulf1 Sult1a1 Susd6 Svil Tab2 Tagln2 Tap2 Tapbp Tcirg1 Tead1 Tec Tep1 Tgfb1 Tgfb2 Tgfb3 Tgfbr1 Tgfbr2 Tgif1 Thbd Thbs2 Thbs4 Timp1 Timp2 Timp3 Tlr3 Tm4sf1 Tmed10 Tmed3 Tmed5 Tmem119 Tmem123 Tmem150a Tmem170b Tmem176a Tmem18 Tmem47 Tmem86a Tmsb4x Tmtc2 Tnfaip8l2 Tnfrsf1a Tnni1 Toporsos Tpm2 Tpm3 Tpm4 Tpp1 Tpr Trem2 Trex1 Trim12a Trim25 Trim56 Trip11 Trp53i13 Tsc22d4 Tspan2 Tspan4 Ttc28 Ubald2 Ucp2 Unc93b1 Usp25 Ust Vamp5 Vasp Vat1 Vgll4 Vkorc1 Vps54 Vtn Wapal Wasf2 Wfdc17 Wipf1 Wls Wnk1 Wnt5a Wrn Wsb1 Wwtr1 Xlr Ybx3 Zbtb20 Zc3hav1 Zeb2 Zfhx3 Zfp36l1 Zfp703 Zic1 Zmiz1 Znfx1 Abca1 Actb Agtrap Aif1 Apbb1ip Apod Arpc1b B2m Bcas1 Bst2 C1qa C1qb C1qc Cald1 Car2 Ccnd1 Ccnd2 Cd52 Cd63 Cd81 Cldn11 Clic1 Cnp Crip1 Ctsb Ctsd Ctss Ctsz Cx3cr1 Cyba Dbi Dhrs1 Eif2ak2 Enpp2 Ermn Fabp7 Fam46a Fcer1g Fth1 Ftl1 Fyb Gbp3 Gcnt2 Gfap Grb14 Grn H2-D1 H2-K1 Hexb Id1 Id3 Ifi27 Ifi27l2a Ifit1 Ifit3 Ifitm3 Igfbp2 Igfbp5 Igfbp7 Itgam Itm2b Lcn2 Lgals1 Lgals9 Ly86 Mag Mal Malat1 Mbp Mgp Mgst1 Mobp Mpeg1 Mrps6 Msn Mt1 Mt2 Myl12a Myo6 Ncf1 Nek6 Nfe2l2 Nfkb1 Nfkbia Nupr1 Pabpc1 Pdlim4 Pea15a Plek Plp1 Pold4 Pou2f2 Prdx6 Psmb8 Ptbp3 Pycard Rhoc Rhog Rnase4 Rpl22 S100a11 S100a13 S100a16 S100a6 Sdc4 Serpina3n Siglech Sparc Stat1 Stat3 Tmem176b Trf Trim30a Tspo Ttr Tyrobp Uaca Vamp8 Vcan Vim Ybx1
Immediate early genes that were observed to be upregulated around the injury site at 3 days and 2 weeks	Fos, Arc, Npas4, Junb
Genes that correlate with Fos, Arc, Npas4, and Junb in the overlap analysis at the 2 week timepoint	Egr1, Egr4, Lmo4, Nr4a1, Slc16a13, Rgs4, Grin2b, C1ql3
Fig. 4K metagene	Fos, Arc, Npas4, Junb, Egr1, Egr4, Lmo4, Nr4a1, Slc16a13, Rgs4, Grin2b, C1ql3

Table 10-3: All figures without "S" refer to Chapter 4, whereas all figures with "S" refer to this chapter.

Figure	Pucks used
1C	180413_7 (coronal hippocampus)
1D	180430_1(coronal cerebellum), 180528_23 (kidney), 180803_8 (liver), 180430_3(coronal olfactory bulb)
2B	180430_6 (coronal cerebellum)
2C	180819_9 (sagittal cerebellum), 180819_10 (sagittal cerebellum), 180819_11 (sagittal cerebellum), 180819_12 (sagittal cerebellum), 180430_1 (coronal cerebellum), 180430_5 (coronal cerebellum), 180430_6 (coronal cerebellum)
2D	180528_20, 180528_22, 180531_13, 180531_16, 180531_17, 180531_18, 180531_19, 180531_22, 180531_23, 180602_15, 180602_16, 180602_17, 180602_18, 180602_20, 180602_21, 180602_22, 180602_23, 180602_24, 180611_1, 180611_2 (sagittal hippocampus)
3A	sagittal cerebellum: 180819_9, 180819_10, 180819_11, 180819_12, 180819_24, 180819_26, 180819_30, 180821_8, 180821_9, 180821_12. coronal cerebellum: 180430_1, 180430_5, 180430_6
3B-D	180819_12 (sagittal cerebellum)
4A	180819_3 (coronal cortex)
4B	180819_19 (sagittal cortex)
4C	180819_6 (sagittal cortex)
4D	180819_19 (sagittal cortex)
4E	180819_6 (sagittal cortex)
4F	All sagittal cortex: 180819_5, 180819_6, 180819_7 (AS, MP, MM); 180819_19, 180821_3 (ML)
4G-J	All sagittal cortex: 180819_19 (top), 180819_6 (bottom). See Methods for a list of pucks used to determine the list of genes used for GO analysis in Fig. 4G-J.
4K	All sagittal cortex: 180819_5 (top), 180819_6 (bottom)
S1D (left)	180413_7 (coronal hippocampus)
S1D (right)	180819_3, 180819_4, 180819_5, 180819_6, 180819_7, 180819_8, 180819_9, 180819_10, 180819_11, 180819_12 (NA, beads counted on surface)
S1E	180611_6 (sagittal hippocampus)
S2	180430_1 (coronal cerebellum), 180413_7(coronal hippocampus), 180528_23 (kidney), 180803_8(liver), 180430_3(olfactory bulb)
S3B,C	180821_27 (coronal human cerebellum)
S3D	180430_1 (coronal cerebellum),180430_5 (coronal cerebellum), 180430_6 (coronal cerebellum),180821_27 (human cerebellum coronal), 180821_28 (human cerebellum coronal)
S4B	180620_4, 180531_17, 180602_20, 180531_13, 180531_22 (sagittal hippocampus)
S4C	180620_4, 180531_17, 180602_20, 180531_13, 180531_22 (sagittal hippocampus)
S4D	180602_17,180602_20,180611_6 (sagittal hippocampus)
S5A	180602_20 (sagittal hippocampus)

S6A	180430_6 (coronal cerebellum)
S6B	180430_6 (coronal cerebellum), 180413_7 (coronal hippocampus),180528_23 (Kidney)
S7	180819_9 (sagittal cerebellum), 180819_10 (sagittal cerebellum), 180819_11 (sagittal cerebellum), 180819_12 (sagittal cerebellum), 180430_1 (coronal cerebellum), 180430_5 (coronal cerebellum), 180430_6 (coronal cerebellum)
S8A-D	180430_6 (coronal cerebellum)
S8E	180430_6 (coronal cerebellum)
S8F	180430_6 (coronal cerebellum)
S9A,B	180602_16, 180602_17, 180602_18, 180602_20, 180618_4, 180618_7, 180618_12, 180618_13, 180618_14, 180618_15 (sagittal hippocampus)
S9C,D	180528_20, 180528_22, 180531_13, 180531_16, 180531_17, 180531_18, 180531_19, 180531_22, 180531_23, 180602_15, 180602_16, 180602_17, 180602_18, 180602_20, 180602_21, 180602_22, 180602_23, 180602_24, 180611_1, 180611_2, 180611_3, 180611_4, 180611_5, 180611_6, 180611_7, 180611_8, 180611_9, 180611_10, 180611_11, 180611_12, 180611_13, 180611_14, 180611_16, 180615_1, 180615_3, 180615_4, 180615_5, 180615_6, 180615_7, 180615_8, 180615_10, 180615_11, 180615_12, 180615_14, 180615_16, 180615_17, 180615_18, 180615_20, 180615_21, 180615_22, 180618_3, 180618_4, 180618_7, 180618_12, 180618_13, 180618_14, 180618_15, 180618_16, 180618_18, 180618_20, 180618_21, 180618_24, 180620_1, 180620_3, 180620_4, 180620_5 (sagittal hippocampus)
S10	180430_6 (coronal cerebellum)
S11A	180430_6 (coronal cerebellum)
S11C,E-G	180819_12 (sagittal cerebellum)
S12	180819_12 (sagittal cerebellum)
S14A-C	180819_3 (coronal hippocampus)
S14D	180430_1, 180430_5, 180430_6 (sagittal cerebellum)
S15	180819_5 (top left), 180819_6 (bottom left), 180819_7 (top right), 180819_8 (bottom right)

Chapter 11
Appendices to Chapter 5

Appendix A

Due to stochasticity, noise, and context-dependence (e.g. sequence-dependence) of the NAAB-amino acid interactions, a measurement performed on the kth target will yield an approximation $\vec{w}$ to the reference affinity vector $\vec{v}_k$. If we assume that the distribution according to which these measurements occur is Gaussian, then we can obtain a simple criterion for determining whether two N terminal amino acids will be distinguishable on the basis of affinity measurements made using a particular set of NAABs. We denote by $\sigma_j^{(i)}$ the standard deviation of the measurements made with NAAB i against amino acid j. For each amino acid, we may define a sphere of radius ρ_j, centered on the vector $\vec{v}_j$, which surrounds that amino acid in affinity space. Here,

$$\boldsymbol{\rho_j = 3 \max_i \frac{\sigma_j^{(i)}}{K_j^{(i)}}} \tag{33}$$

where $K_j^{(i)}$ is the dissociation constant for the binding of the ith NAAB to the jth amino acid.

N-terminal amino acids will be identifiable with 99.9% certainty provided that there is no overlap in affinity-space between the j spheres of radius ρ_j. To determine whether there is such an overlap, we must consider the distance metric.

$$\boldsymbol{D \equiv \min_{i,j \neq i} \left\| \frac{\vec{v}_i - \vec{v}_j}{\vec{v}_i} \right\|} \tag{34}$$

where the division is applied element-wise. In order to assign affinity measurements to the correct reference affinity 99.9% of the time, it is sufficient (but not necessary) to have

$$\boldsymbol{\max_{i,j \neq i} (\rho_i + \rho_j) \leq D} \tag{35}$$

Using Eq. (33), it is then sufficient to have

$$\boldsymbol{6 \max_{i,k \neq i} \left(\frac{\sigma_k^{(i)}}{K_k^{(i)}} \right) \leq D} \tag{36}$$

For the specific case of the NAAB affinity matrix, we find that $D = 3.84$. Thus, in order to ensure that the amino acids can be correctly identified 99.9% of the time, we must have

$$\boldsymbol{\max_{i,k \neq i} \left(\frac{\sigma_k^{(i)}}{K_k^{(i)}} \right) \leq 0.64} \tag{37}$$

or, equivalently, the standard deviation of the k_D measurements must be no greater than 64% of the mean.

Appendix B

Under the assumption of Poissonian noise, the photon rates in the bound and unbound states are given by

$$\boldsymbol{\lambda_f = R\tau_{obs} n_{free}} \tag{38}$$

and

$$\boldsymbol{\lambda_b = R\tau_{obs}(n_{free} + 1)} \tag{39}$$

respectively. In order to be able to distinguish the bound state from the unbound state, it is clear that it is sufficient to have

$$\boldsymbol{\lambda_f + 3\sqrt{\lambda_f} \leq \lambda_b - 3\sqrt{\lambda_b}} \tag{40}$$

Because $\lambda_b > \lambda_f$, we may replace the standard deviation $\sqrt{\lambda_f}$ on the left-hand side by the standard deviation $\sqrt{\lambda_b}$, obtaining

$$\boldsymbol{\lambda_f \leq \lambda_b - 6\sqrt{\lambda_b}} \tag{41}$$

Hence,

$$\boldsymbol{R\tau_{obs} \geq 6\sqrt{R\,\tau_{obs}(n_{free} + 1)}} \tag{42}$$

We find the final requirement:

$$\boldsymbol{n_{free} \leq \frac{R\,\tau_{obs}}{36} - 1} \tag{43}$$

Rephrased as a condition on the concentration of the binder, we find

$$\boldsymbol{c \leq \frac{\frac{R\,\tau_{obs}}{36} - 1}{1000\,N_A V}} \tag{44}$$

or

$$\boldsymbol{R\tau_{obs} \geq 36\,(1 + n_{free})} \tag{45}$$

If $n_{free} \leq 1$, then the assumption of Poissonian noise is invalidated because the emission of successive photons is not independent (it depends on the presence of fluorophores in the observation field). The assumption of Poissonian noise may also be invalidated if the frame rate is comparable to the rate at which fluorophores enter and leave the observation field. In either case, to correctly simulate the noise, one must draw the number of free binders that enter the observation field during a given frame from a Poisson distribution with mean $n_{free}\tau_{obs}/\tau_{dwell}$,

where τ_{dwell} is the amount of time each binder spends in the observation field on average. The average dwell time of free binders in a region of thickness Δx may be calculated as

$$\boldsymbol{\tau_{\mathrm{dwell}} = \frac{(\Delta x)^2}{D}} \tag{46}$$

where D is the diffusion constant (*204*). For a small protein in water, we have $D \sim 10^{10}\ \mathrm{m^2 s^{-1}}$. Taking $\Delta x = 100\ \mathrm{nm}$, we find that free binders will dwell on average $\tau_{\mathrm{dwell}} = 100\ \mu\mathrm{s}$ within the imaging plane.

Once the number of binders entering the observation field during the frame has been determined, one must draw the length of time t that each binder remains in the frame from an exponential distribution with mean τ_{dwell}. Finally, for each binder, one must draw the number of photons emitted by that binder from a Poisson distribution with mean Rt. When the number of free binders is small, the resulting noise will differ significantly from Poisson noise due to the exponential distribution over dwell times. In our simulations, the long tail of the exponential distribution tends to significantly increase the difficulty of distinguishing transient binding and unbinding events, compared to simple Poisson noise (data not shown).

Appendix C

One advantage of occupancy measurements is that if k_{on} is known, then k_{off} may be determined even in the presence of photobleaching. To do so, we note that T_i and T_b are independent variables that depend on k_{off}, k_{on}, and N_q. In the above analysis, we assumed that N_q was infinite, so that quenching could be neglected. If N_q is finite, however, then the true expressions for T_i and T_b are given by

$$\boldsymbol{T_b = \frac{1}{k_{\mathrm{off}} + R/N_q}} \tag{47}$$

and

$$\boldsymbol{T_i = \underbrace{\left(\frac{1}{k_{\mathrm{off}}} - T_b\right)}_{\text{target occupied}} + \underbrace{\frac{1}{k_{\mathrm{on}} c}}_{\text{target unoccupied}}} \tag{48}$$

The first term in Eq. (48) is the average time the target spends occupied by a quenched fluorophore, while the second term is the average time the target spends unoccupied between unbinding and binding events. Hence, if k_{on} is known, then k_{off} and N_q may be determined from T_b and T_i.

Appendix D

In contrast to occupancy measurements, luminosity measurements are sensitive to error in the calibration of the measurement apparatus. Calibration error arises from a combination of systematic differences in the brightness of the on- and off-states, which may result if different NAABs have different numbers of fluorophores on average, and from systematic error in the

measurement of the brightnesses of the on- and off-states. Systematic variation in the brightnesses of the fluorophores can be overcome by calibrating the device prior to each measurement (as discussed below). In general, however, systematic error in the measurement of S and N significantly disrupts attempts to determine the absolute value of k_D due to divergences in the derivative of k_D as M approaches N. Hence, for weak binders in particular, infinitesimal changes in the calibration level can lead to divergent changes in the measured value of k_D. For this reason, if the goal of the measurement is to determine the absolute value of k_D, it is essential that the concentration be chosen such that the value of M to be measured lies close to S, i.e., such that the concentration c is close to or greater than k_D. If k_D is large or unknown, however, this requirement may not be achievable.

In our case, however, we are interested not in determining the absolute value of k_D, but rather in determining the identity of a target (N-terminal amino acid) from the binding affinities of many binders (NAABs). In this case, one may significantly reduce the effects of calibration error by using the reference values of k_D to calculate the expected photon rate E from the brightnesses of the on- and off-states, for each of the possible target identities. After having performed the measurement with all 17 binders, one is left with a vector $\vec{M}$ of the photon rates measured for each binder, and a set of vectors $\vec{E}_k$, the kth of which is the vector of photon rates that one would have expected to measure if the target were of type k. The identity of the target is then determined by minimizing the norm of $\vec{M} - \vec{E}_k$ over k. The key difference here is that because one compares the expected photon rates to the measured photon rates, one avoids the nonlinearities inherent in calculating the measured dissociation constant from the measured photon rate.

Appendix E

Figure 11-1 shows the full set of accuracy matrices determined by simulation for 100 random affinity matrices.

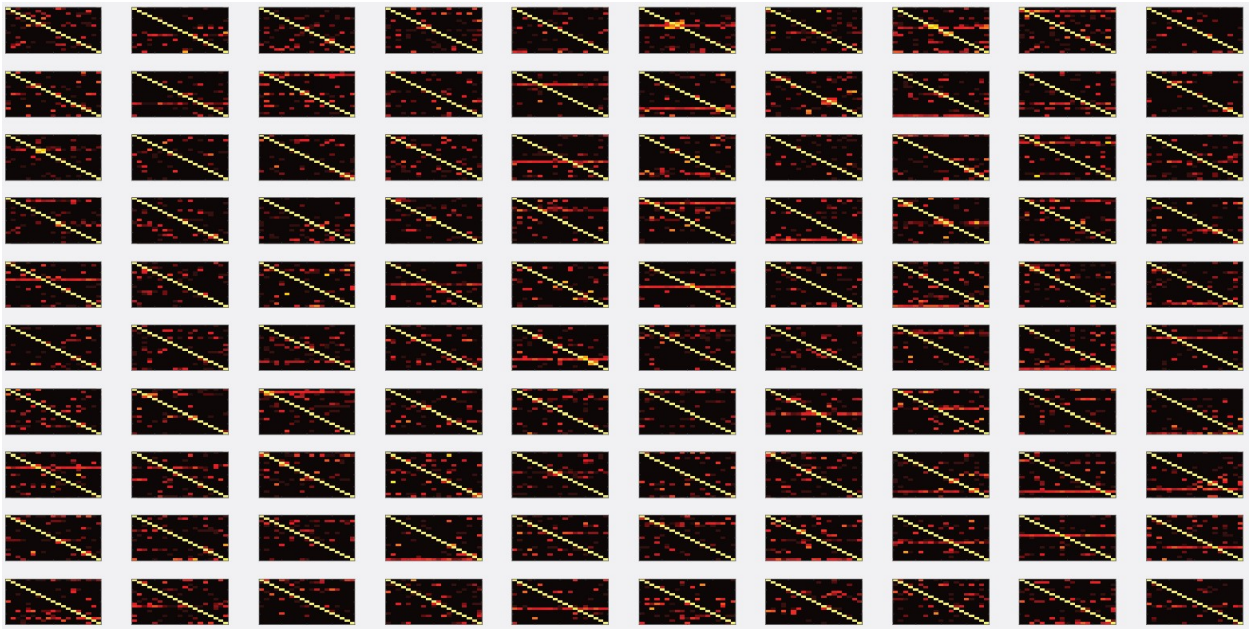

Figure 11-1: Accuracies for amino acid calling obtained for 100 random affinity matrices in simulations. 100 random affinity matrices were generated by randomly shuffling the entries of the NAAB affinity matrix. For each resulting matrix, we simulated 10000 amino acid calls, with 5% calibration error and 0.25% kinetic error. The resulting accuracy matrices are presented here. The scale and axes for each matrix are identical to those in Figure 5-4E.

Chapter 12
Supplementary Information to Chapter 6

Methods:

Cloning:

All plasmids were constructed either using restriction cloning using restriction enzymes from New England Biosciences and the NEB Quick Ligation kit (M2200L), or using the In-Fusion HD cloning enzyme mix (Clontech, 638911). Plasmids were grown in E.Cloni 10G Chemically Competent Cells (Lucigen, 60107-1) and were verified by Sanger sequencing (Eton biosciences). All plasmids are deposited on Addgene.

Due to high repetition present in the RNA editing templates, inserts for plasmids 76, 147, 148, 149, and 187 (see Table 12-2) were ordered as sense and antisense ultramer oligonucleotides, which were annealed to each other prior to cloning. Plasmid 76 was cloned by inserting RNA templates (A_Short, B_Short, C, D, E) into the 3' UTR of an iRFP transcript expressed under a UbC promoter in a second generation lentivirus backbone using SphI and ClaI. Subsequently, this plasmid was modified by the addition of a flavivirus xrRNA in the 5' UTR. Templates A_Short and B_Short were then extended by inserting another pair of annealed ultramers on the 5' side of A_Short and B_Short using SphI and MluI. The resulting templates are designated A and B. To generate plasmids 147, 148, 149, and 183 (as used in the paper), templates A and B were then moved into different backbones and different promoters by restriction cloning, or by Gibson assembly with PCR amplification of the repRNA template region. Template A is used throughout the paper, and Template B is shown in Figure 12-1 for comparison.

RNA Purification, Library Preparation, and Sequencing

All cell cultures were lysed with 600uL of buffer RLT Plus from the Qiagen RNEasy Plus Mini Kit (Qiagen, 74136), and were pipetted up and down vigorously to homogenize. RNA was then purified using the Qiagen RNEasy Plus Mini kit, following the instructions from the manufacturer. Subsequently, 11uL of purified RNA was reverse transcribed using Superscript IV (Thermofisher, 18090050) and a barcoded version of SGR-174 (see Table 12-2), following the protocol from the manufacturer. Reverse transcription reactions were then purified using Agencourt Ampure XP beads at a 1:1 dilution (Beckman-Coulter, A63881). Some portion of the eluent, typically 25%, was then PCRed using P5 and a barcoded version of SGR-176 (see Table 12-2) the Q5 Hot Start High Fidelity 2x Master Mix (NEB, M0492L) with the following settings: 30s of 98C denaturation;

then 25-30 cycles of 10s denaturation at 98C, 20s annealing at 70C, and then 25s extension at 72C. Neuron lysates were typically PCRed for 30 cycles, while HEK cell lysates were typically PCRed for 25 cycles. PCR reactions were then pooled and run on a gel, and a 400bp band was extracted using the NucleoSpin PCR Cleanup Kit (Macherey-Nagel, 740609.250). The concentration of DNA in the resulting eluent was determined via a Qubit 2 fluorometer (Thermofisher), and was then adjusted to 4nM for sequencing. The read structure is shown in Figure 12-6.

Sequencing was performed using NextSeq Mid Output 300 cycle kit (Illumina, FC-404-2004), Miseq 300 cycle v2 kits (MS-102-2002), or Miseq 600 cycle v3 kits (MS-102-3003), with at least 80bp read 1 and 185bp read 2, with 8bp index 1 and 15bp index 2.

HEK and 3T3 cell culture:

Except in the case of the single cell experiments, HEK293FT and 3T3 cells were plated in 24 well plates. Cells were grown in DMEM (Thermofisher, 10566016), supplemented with Pennicillin/Streptomycin (Thermofisher, 15140122) and 10% certified Tet-system approved FBS (Clontech, 631101). Transfections were performed using the TransIT-X2 system (Mirus, MIR 6000), following the manufacturer's instructions.

For doxycycline experiments, HEK and 3T3 cells in 24 well plates were transfected with 300ng of plasmid 147 or 148, 100ng of pCMV Tet3G from the Tet-on 3G system (Clontech, 631168), and 100ng of plasmids 116v1, 116v5, or 116v6. In the experiments for Figures 1, 2, 3, and S1, they were transfected with both 147 and 148, and received 150ng of each plasmid. At least 12 hours after transfection, cells were stimulated by adding doxycycline to a final concentration of 1ug/mL, followed by gentle mixing or swirling of the plate. Subsequently, transcription was halted by adding Actinomycin D to a final concentration of 1ug/mL in the same medium. After waiting for the experimental time period, cells were lysed using Buffer RLT Plus and libraries were prepared as described above.

For experiments using the Vivid promoter, 3T3s were transfected with 300ng of plasmid 149, 100ng of pCMV Tet3G, and 100ng of plasmid 116v5. For conditions in which cells were transfected with both plasmid 147 and plasmid 149, they received 150ng of each plasmid. For the experiments in Figure 12-3, cells were stimulated with a blue LED (Thor Labs, M455L2) with a total power of 200uW/cm^2. The LED was turned on for 1 hour, and was subsequently turned off. After the LED was turned off, the cells were wrapped in foil to prevent accidental light exposure. Cells were then lysed after the experimental time period.

HEK Cell Doxycycline Experiment

For the experiment in Figure 6-1E,F, cells were stimulated as above and were lysed at the following timepoints: 0 hours (i.e., immediately before adding dox), 0.5 hours after adding dox, 1 hour after adding dox (i.e., immediately before adding ActD), 2 hours after adding dox, 3 hours after adding dox, 4 hours after adding dox, 5 hours after adding dox, 6 hours after adding dox, 7 hours after adding dox, 8 hours after adding dox, 9 hours after adding dox, 10 hours after adding dox, 11 hours after adding dox, and 12 hours after adding dox. Each timepoint consisted of three replicates. On a separate occasion, we collected three replicates at 2.5 hours after adding dox and 4.5 hours after adding dox, and these timepoints functioned as our test timepoints in Fig. 2D,E.

Vivid Experiments:

For the experiment in Figure 12-3, we collected three replicates for each of the following timepoints: immediately prior to turning on the LED, 1 hour after turning on the LED (i.e., immediately prior to turning off the LED), 2 hours after turning on the LED, 3 hours after turning on the LED, 4 hours after turning on the LED, and 5 hours after turning on the LED.

Single Cell Experiments:

For all experiments involving single cells, HEK cell cultures were prepared, transfected with 100ng of pAAV-CAG-GFP (Addgene 37825), 200ng of plasmid 147, 100ng of plasmid 116v5, and 100ng of pCMV Tet3G, stimulated with doxycycline, and then silenced with actinomycin D as described above. Subsequently, at the designated timepoint (e.g., 8 hours or 4 hours after doxycycline was added to the culture medium), cells were treated with trypsin (Life Technologies, 25300054). Following trypsinization, cells were centrifuged at 850g, washed in cold PBS, and then resuspended in cold PBS. 96 well plates were prepared, with each well containing a solution of 0.2% Triton-X with 2U/uL RNAse inhibitor. Individual cells were sorted into the wells of this wellplate using a Moflo Astrios EQ flow cytometer. Following sorting, the wellplate was sealed, centrifuged, and then placed at -80C overnight.

For the analysis in Figure 6-4, cells in condition 2 received plasmid 147B1, while cells in condition 3 received plasmid 147B3. The two populations of cells were mixed following trypsinization and sorted together. By contrast, cells in condition 1 received plasmid 147B1, and were sorted separately from the others.

The single cell analysis was nominally conducted with cells from 4hr and 8hr timepoints. However, following trypsinization, cells remained in cold PBS for up to an hour and a half due to latencies in the sorting process. For this reason, we compared the estimates from the single cells to the estimates for populations of ~100,000 of the same cells (i.e., stored in cold PBS for the same amount of time) lysed immediately after sorting.

Library preparation for the single cells proceeded as follows. Plates containing single cells were thawed, and 7uL of nuclease free water was added to the single cells to bring the total volume up to 11uL. Subsequently, reverse transcription was performed using Superscript IV and the SGR-174 RT primers, as in the case of the bulk samples, with the following modifications. RT primers were distributed so that each cell at a given timepoint received an RT primer with a different barcode. In addition, for each timepoint, we performed two no-template RT reactions. Finally, after the 50C step in the Superscript IV protocol, we cooled the samples to 37C and added 20U of Exonuclease 1 (NEB, M0293S) to the reaction to remove excess primers. Samples then remained at 37C for 10 minutes, before proceeding to the 80C heat inactivation step. Following reverse transcription, the RT reactions for all cells and the two no-template controls at a given timepoint were pooled, cleaned with Ampure XP beads at a 1:1 dilution, and were then PCRed using the same protocol as for the bulk samples. Cells were pooled prior to PCR as a way of reducing the number of cycles necessary to achieve amplification. We excluded cells if they received fewer than 150 reads, or if the most common RNA barcode represented fewer than 80% of the total deduplicated reads, which would indicate index swapping between cells.

Neuron Culture Preparation and Transfection:

All procedures involving animals at MIT were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Massachusetts Institute of Technology Committee on Animal Care. Primary hippocampal neuron culture was prepared as previously described. Neuron cultures were transfected at 6-7 DIV using a commercial calcium-phosphate kit (Thermofisher, K278001), as previously described. Briefly, neurons were transfected with 600ng of pUC19, 200ng of plasmid 116v5, and 200ng of plasmid 187. Neurons were then incubated with calcium-phosphate precipitates for 30-60 minutes, followed by washing with MEM buffer at pH 6.7-6.8 to remove residual precipitates.

Neuron Culture Stimulation:

Neurons were stimulated at 14-15DIV. Neurons were placed in 1mL of plating medium (500mL MEM, 2.5g glucose, 50mg transferrin, 1.1g HEPES, 5mL 200mM L-Glutamine, 12.5mg insulin, 50mL HI FBS, 10mL B27 supplement). To stimulate the neurons, we added 250uL of 5x depolarization medium and agitated gently. Neurons were then left for one hour in an incubator. Subsequently, the medium was aspirated and neurons were washed twice in plating medium. They were then left in plating medium for a variable amount of time, before being lysed in 600uL of buffer RLT Plus.

Plating Medium:

1. 500mL MEM (Thermofisher, 51200-038)

2. 2.5g glucose (Sigma Aldrich, G7528-1KG)
3. 50mg transferrin (Sigma Aldrich, T1283-500mg)
4. 1.1g HEPES (Sigma Aldrich, H3375-500G)
5. 5mL 200mM L-Glutamine (Thermofisher, 25030-081)
6. 12.5mg insulin (Millipore, 407709)
7. 50mL HI FBS (VWR, 45000-736)
8. 10mL B27 Supplement (Thermofisher, 17504-044)

5x Depolarization Medium

1. 170mM KCl
2. 10mM HEPES pH 7.4
3. 1mM MgCl2
4. 2mM CaCl2

Neuron Inference Experiment:

Due to the limited availability of neuron culture at any given time, the data for Figure 6-5 was conducted in two separate experiments, which can be considered to be biological replicates. We collected the following timepoints: prior to stimulation (i.e., immediately before adding depolarization medium); 1 hour after stimulation (i.e., immediately before washing the neurons in fresh medium); 2 hours after stimulation; 3 hours after stimulation; 3.5 hours after stimulation; 4 hours after stimulation; 5 hours after stimulation; 5.5 hours after stimulation; 6 hours after stimulation; 7 hours after stimulation.

The breakdown of the data in Figure 6-5 by experiment is as follows. In the first experiment, we collected two samples prior to stimulation; three samples at 1 hour; three samples at 2 hours; three samples at 3 hours; three samples at 4 hours; and two samples at 5 hours. In the second experiment, we collected one sample at 2 hours, two samples at 3 hours, three samples at 3.5 hours, two samples at 4 hours, two samples at 5 hours, three samples at 5.5 hours, two samples at 6 hours, and two samples at seven hours.

Multiplexing:

Experiments for Figure 9-4 were conducted as follows. Three wells of 3T3 cells were transfected as described above with 100ng each of pCMV Tet3G, plasmid 133, plasmid 147B1, plasmid 149B3, and plasmid 116v5. Three wells were transfected with 100ng of pCMV Tet3G, 100ng of plasmid 116v5, and 100ng of plasmid 147B1, and 200ng of pAAV-CAG-GFP. Finally, three wells were transfected with 100ng of plasmid 133, 100ng of plasmid 149B3, 100ng of plasmid 116v5, and 200ng of pAAV-CAG-GFP. Subsequently, all 9 wells were irradiated with blue light as described above for 1 hour, and were the placed in darkness. 7 hours after placing the cells in darkness, cells

were stimulated with doxycycline as described above. After one hour in doxycycline, cells were lysed.

Alignment and Edit Counting:

The alignment and analysis pipeline for sequencing data is summarized in Figure 12-6. Analysis of sequencing data was performed using custom Matlab code. Briefly, in the case of single cell data, we first performed deduplication using a 9bp UMI on the RT primer (oligo SGR-174). Other datasets were not deduplicated. Reads were then filtered to ensure that they had the minimum necessary read length (67 bases on Read 1, and 184 bases on Read 2). Note that Read 1 was on the RT primer, so Read 1 reads the reverse complement of the RNA sequence. Thus, the expected mutation was A to G on Read 2, and T to C on Read 1. Alignment was performed using all bases that were not As on Read 2, or that were not Ts on Read 1. Reads were considered to be aligned to the template if 95% of the non-A (for Read 2) or non-T (for Read 1) bases matched the template. Furthermore, we required 90% of the bases that were expected to be As on Read 2 or Ts on Read 1 have Q scores greater than 27 (Figure 12-6); reads that failed to achieve this threshold were discarded.

Finally, except as stated in Figure 12-2, we required that all reads have at least one edit in Read 1 and at least one edit in Read 2 for analysis (Figure 12-6D). We implemented this requirement because it appeared to eliminate a number of artifacts that we occasionally observed in our data: for example, each well would sometimes have different (large) numbers of RNAs with zero edits or one edit, which would confound attempts to infer timing from the mean editing rate, as in Figures 5 and Figure 12-3. As a consequence of this requirement, all of the histograms of edits per RNA presented in this paper appear not to show any RNAs with fewer than ~12 edits. There are ~12 bases in template A, all of which are on Read 2, that are edited much more quickly than any bases on Read 1. These are of the form UAG, and all form bulges in the RNA secondary structure, which is thought to encourage editing by ADAR. Exclusion of RNAs with zero edits on Read 1 or Read 2 limits the analysis to RNAs that are already fully edited at all 12 of those As, thus causing all RNAs to have at least 12 edits.

Linear Interpolation:

In Figure 6-5 and Figure 12-3, the timepoints associated with the c-fos neural activity and with the vivid promoter were determined by linear interpolation, as follows. We first calculated the mean number of edits per RNA for all replicates, and determined the mean across replicates for each timepoint (plotted in Figure 6-5B and Figure 12-3B, designated M_t). Then, to perform the estimate, for each replicate R from timepoint t we identified the two timepoints t_1 and t_2 such

that t != t_1,t_2 and such that the mean m_R of replicate R obeyed $M_{t1} < m_R < M_{t2}$. The time estimate for replicate R is then determined as

$$t_R = \frac{m_R - M_{t1}}{M_{t2} - M_{t1}}(t_2 - t_1) + t_1$$

Exponential Model:

The exponential model in Figure 6-2 was implemented using custom code in Python, as follows. For each editable position i on the template, we assume the likelihood of base i being edited follows an exponential distribution with parameter λ_i, to be estimated from the data. Assuming an instantaneous pulse of transcriptional activity at time t=0, the fraction of edited bases for position i, y_i, can be modelled as the CDF of the exponential distribution:

$$y_i(t) = 1 - e^{-\lambda_i t}$$

To more accurately capture the experimental setup, we model y_i as an underlying process which is exponential, but with start time uniformly distributed in [0, t_{stop}], where t=0 represents when doxycycline is added to the cells and t_{stop} is the time at which actinomycin D was added to the cells. Specifically, we fit a function of the form

$$y_i(t) = \begin{cases} 1 - \dfrac{1 - e^{-\lambda_i t}}{\lambda_i t} & \text{if} \quad t \leq t_{stop} \\ 1 - \dfrac{e^{-\lambda_i (t - t_{stop})} - e^{-\lambda_i t}}{\lambda_i t_{stop}} & \text{if} \quad t > t_{stop} \end{cases}$$

where t_{stop} was 1hr and λ_i was fit to the data using non-linear least squares. This function was fit for times $t \geq$ 1.5hr, since the editing distributions for earlier timepoints are strongly affected by populations of RNA present prior to doxycycline addition (for example, the mean editing rate in Fig. 1F decreases from t=0 to t=1). For the analysis in Fig. 3, analysis was then performed using only those adenosines for which the R^2 of the resulting fit was greater than 0.9. We model the total number of edits to the RNA with a Poisson binomial distribution with N trials where N is the total number of editable positions and success probabilities given by $y_i(t)$ for each position i. The probability of having n edits at time t is given by

$$p(n,t) = \sum_{\mathrm{A:\ sum}(A)=n} \prod_{k:\ \mathrm{A_k}=1} y_k(t) \prod_{j:\ \mathrm{A_j}=0} 1 - y_j(t)$$

Here, A is a binary vector with each entry corresponding to a specific adenosine in the repRNA editing region. A_k=1 if adenosine k has been edited to inosine, and sum(A) counts the total number of edits in A. Time estimates using the exponential model were then made by minimizing the Kullback-Leibler divergence between $p(n,t)$ and the empirical distribution $q(n)$ over t. $p(n,t)$ was calculated in practice via a dynamic programming approach.

For Figure 12-2A-C, the exponential model was calculated using the data from a single replicate of the HEK doxycycline experiment. The distributions in Figure 6-2C show the number of edits per RNA calculated across all bases with R^2 greater than 0.9 for that replicate, and the Poisson binomial model in Figure 6-2C likewise included the same bases. By contrast, for Figure 12-2D-E, bases were only retained if they had R^2 greater than 0.9 in all three replicates from the HEK doxycycline experiment. For this reason, the apparent numbers of edits per RNA are lower in Figure 9-2D-E than in Figure 12-2C.

Gradient Descent:

The gradient descent in Figure 6-3 and Figure 6-4 was implemented using custom code in Matlab. Briefly, the gradient descent algorithm was given an RNA editing distribution, which could either be an empirical distribution (Figure 6-3B single induction timepoints; Figure 6-3F,G; Figure 6-4A-C) or a simulated distribution (Figure 6-3B except single-induction; Figure 6-3C-E; Figure 6-4D,E). Simulated distributions were convex combinations of the editing histograms for a single replicate from the HEK doxycycline experiment. The gradient descent algorithm was also given a set of "basis vector" histograms, which were obtained by combining the data at each timepoint from all three replicates from the HEK doxycycline experiment. The gradient descent was then initialized by drawing a set of weights from a Dirichlet distribution with all parameters set to unity. The gradient descent minimized the mean squared error (L2 norm) between the input distribution and the convex combination of the basis vectors given by the weights. For each simulated distribution, we performed the gradient descent 1000 times and took the solution that minimized the L2 norm. For the analysis in Figure 12-6, we generated 1000 simulated distributions from a Dirichlet distribution with all parameters set to unity.

Accuracy Metrics:

For the single- and double-induction samples in Figure 6-3B and Figure 6-4D, temporal resolution is calculated by multiplying the distance of each timepoint away from the expected timepoint by the weight assigned to that timepoint, and summing. Thus, for the 3-hour single-induction pulse, if the decoder assigned weights of 0.5 to the 3-hour timepoint and 0.5 to the 5-hour timepoint, the resulting resolution would be 0.5*1 + 0.5*3=2 hours. The accuracy of the decoder is measured in

three ways throughout the manuscript. For the double-induction timepoints, we summed over all timepoints greater than 3 hours, after renormalizing the weights so that the sum of the weights assigned to timepoints above 3 hours equaled 1.

For the square waves in Figure 6-3B,F,G, and for the arbitrary transcriptional program experiments in Figure 12-5, we calculated the accuracy as the sum of the absolute values of the differences between the assigned and expected weights, divided by 2 to avoid double-counting. Thus, if we expected one timepoint to get 100% of the total weight, and that timepoint instead got 80% of the total weight, then the resulting accuracy would be 80%.

In Figure 6-4B, the accuracy is calculated as the mean absolute difference between the single cell estimates and the estimate for the bulk distribution. We calculate the accuracy in this way for the single cells because the ground truth transcriptional program is not known. The single cells stay on ice for up to an hour during processing, and we have not measured the editing kinetics during that time.

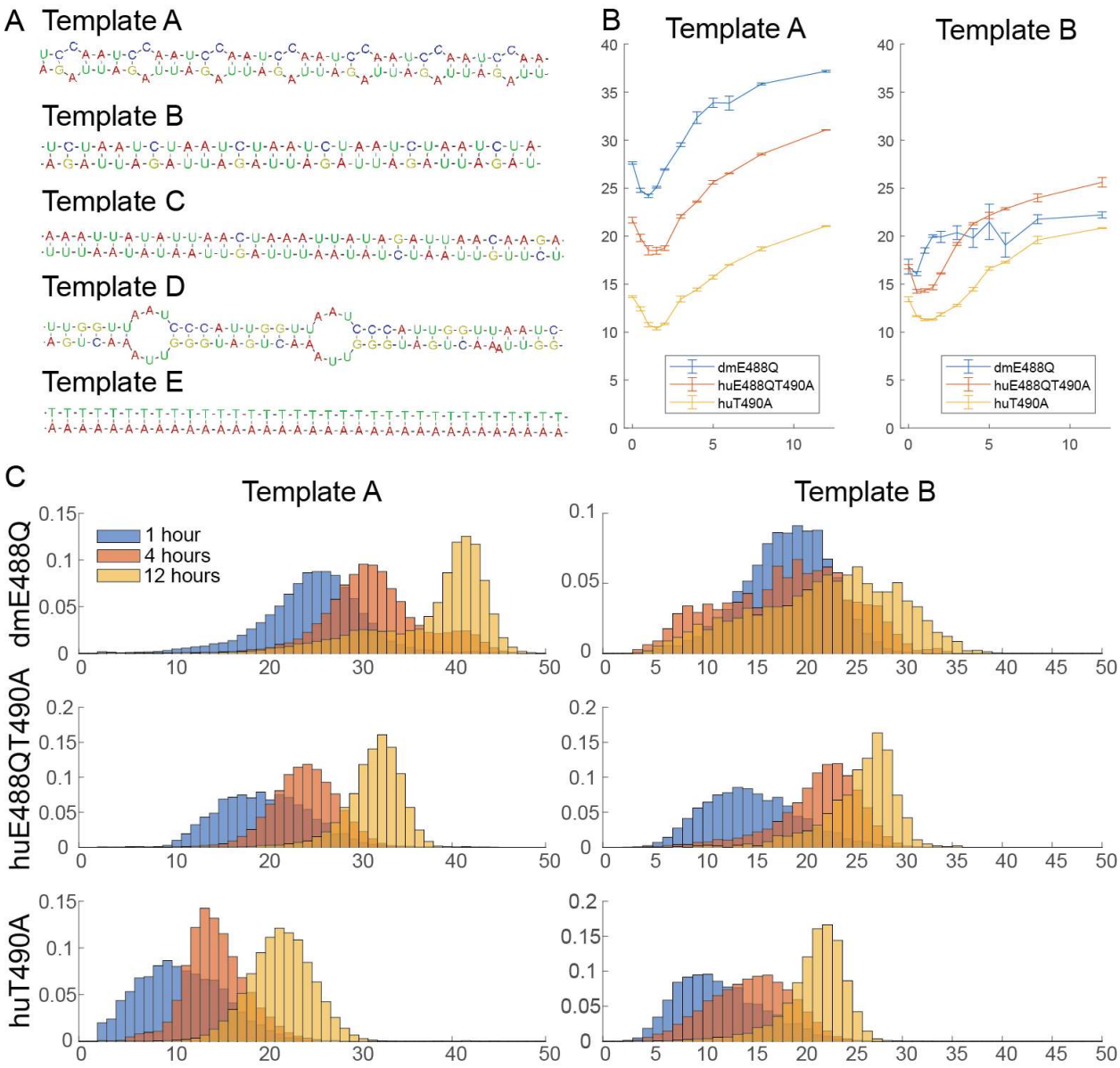

Figure 12-1: **(A)** Five editing templates that were tested are shown. Of these, only templates A and B showed robust temporal editing that seemed appropriate for the construction of the tickertape system. Notes on the templates are in Table 12-3. **(B)** The mean number of edits per RNA for several different timepoints is shown for three different ADAR variants, and for templates A and B. The protocol used here is identical to that in Figure 6-1E. Some combinations, such as dmE488Q with template A, may show greater temporal resolution at short timescales. **(C)** Example editing histograms are shown for three different timepoints, for each combination of the three enzymes and two templates in (B).

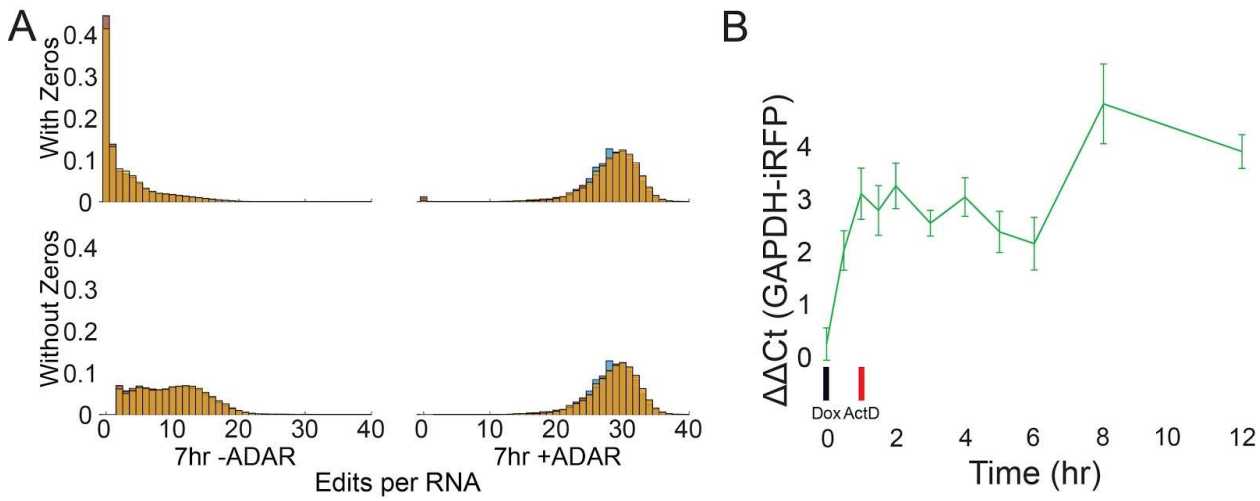

Figure 12-2: (A) Cells were induced with doxycycline, followed by actinomycin D 1 hour later, and then lysis 7 hours after actinomycin D. All editing histograms are normalized to sum to 1. Top left: the editing histogram for cells that were not transfected with ADAR, without removing RNAs with no edits on read 1 or read 2 (i.e., "with zeros"). Top right: The editing histogram for cells that were transfected with ADAR, without removing RNAs with no edits on read 1 or read 2. Bottom: Same as top, but only considering RNAs with at least one edit on both Read 1 and Read 2 (i.e., "without zeros," see Methods). **(B)** The qPCR for the iRFP transcript, normalized to GAPDH, is shown as a function of time during the experiment in Figure 6-1E. Values are normalized to the pre-doxycycline timepoint. Error bars show standard deviation (N=3).

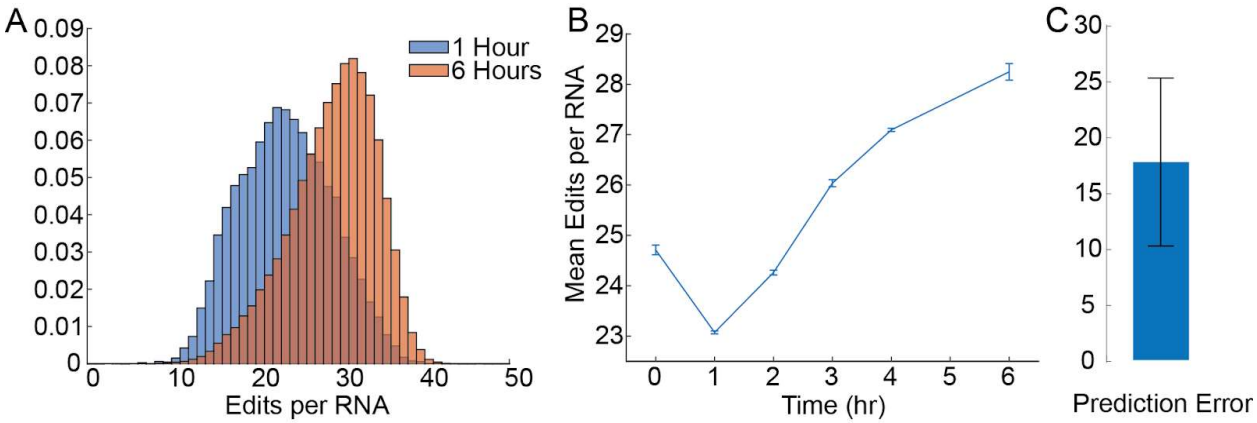

Figure 12-3: The Poisson binomial approach is the preferred approach for this form of estimation because it accounts for the exponential nonlinearity inherent in Poisson processes. However, we also found that a simple linear interpolation of the mean yields accurate estimations in many cases. In the case of the TRE tickertape, the mean interpolation estimated the 2.5hr and 4.5hr timepoints as 2.53hr ± 0.08hr and 4.38hr ± 0.02hr (mean ± s.d., N=3 replicates), with errors of 5min ± 0.3min and 7.5min ± 1.1min (mean ± s.d., N=3 replicates), respectively. We performed similar experiments in 3T3 cells using repRNAs expressed under a light-inducible Vivid promoter (*268*), induced with blue light for one hour. We estimated the timing of light induction by interpolation of the mean number of edits per RNA, and yielded an accuracy of 17.7 ± 7.5 minutes (mean ± s.d., N=9 samples total across three timepoints). **(A)** Editing histograms are shown for 3T3 cells transfected with repRNAs expressed under the Vivid promoter, and stimulated for 1 hour (see Methods). In blue is the editing histogram for cells lysed one hour after stimulation began (i.e., immediately after it ended), and in orange is the histogram for cells lysed 6 hours after stimulation began. All editing histograms are normalized to sum to 1. **(B)** The mean number of edits per RNA is shown for the timepoints generated. Time indicates number of hours since the beginning of stimulation (the first timepoint is pre-stimulation). Error bars are standard deviation (N=3). **(C)** The absolute prediction error, in minutes, is shown, averaged over all replicates for the 2hr, 3hr, and 4hr timepoints. The prediction was performed by mean interpolation, analogously to Figure 6-5D. Error bar is standard deviation (N=9, 3 replicates at each of 3 timepoints).

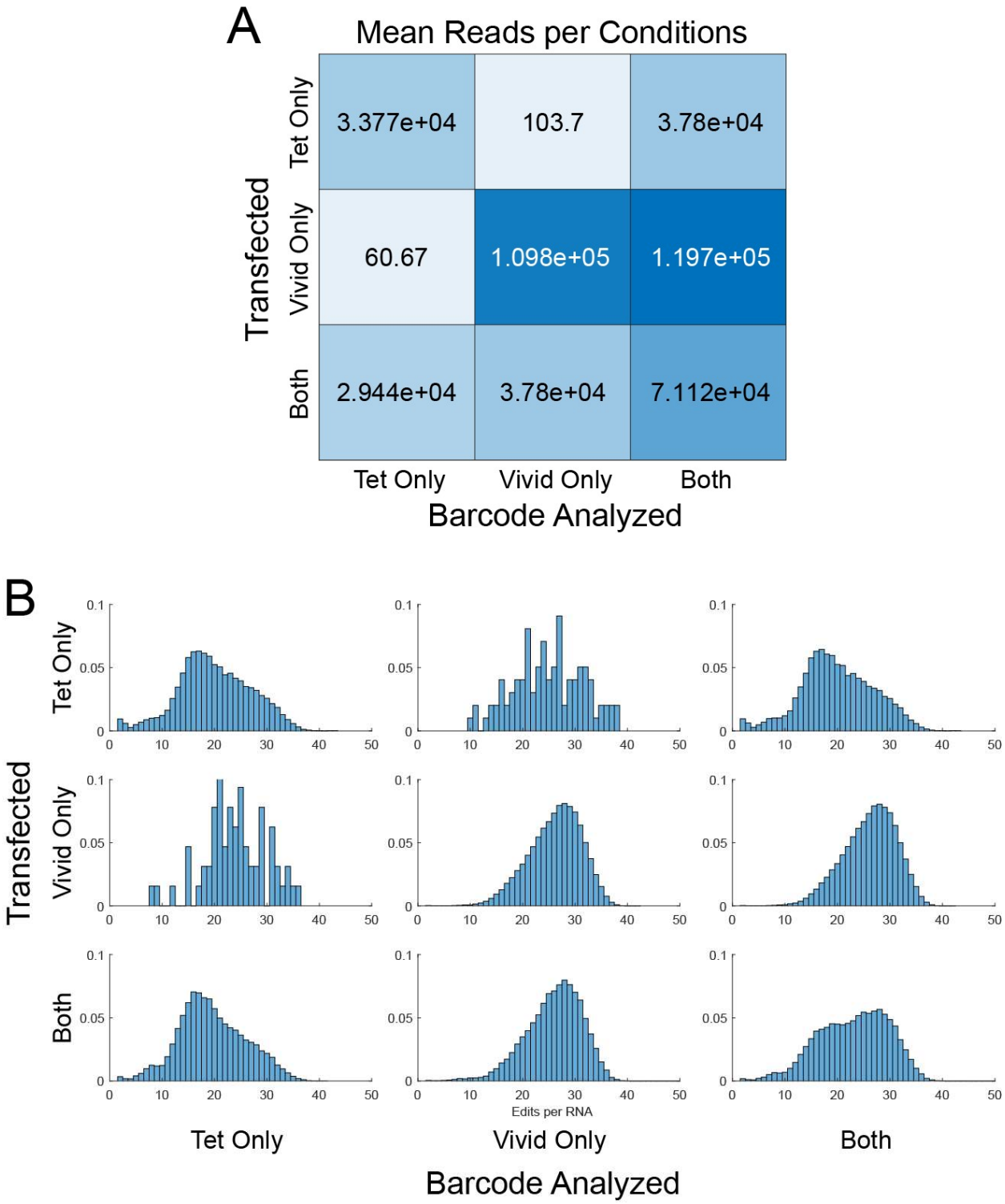

Figure 12-4: The fact that tickertape works with multiple promoters raises the possibility of recording the activity of multiple promoters simultaneously in a single cell population, and we validated that this is possible using barcoded repRNAs responsive to the Tet and Vivid promoters. All editing histograms are normalized to sum to 1. **(A)** For cells transfected with a barcoded TRE-responsive repRNA construct, a barcoded Vivid-responsive repRNA construct, or both, the number of reads for the TRE-responsive

repRNA, Vivid-responsive repRNA, or both are shown. When only one repRNA is transfected, only one barcode is detected in significant numbers, confirming that there is minimal crossover between repRNA barcodes. Note that the third column is not the sum of the first and second columns, because it includes barcodes that did not perfectly align to either the Tet or Vivid repRNA barcodes. **(B)** To further confirm the possibility of multiplexing using barcoded repRNAs, we analyzed the editing histograms for cells that were transfected with a barcoded TRE-responsive repRNA construct, a barcoded Vivid-responsive repRNA construct, or both. The editing histograms for the Vivid-responsive and TRE-responsive repRNAs do not seem to change when the other repRNA is also present, again suggesting that there is minimal cross-talk between barcoded repRNA constructs. All editing histograms are normalized to sum to 1.

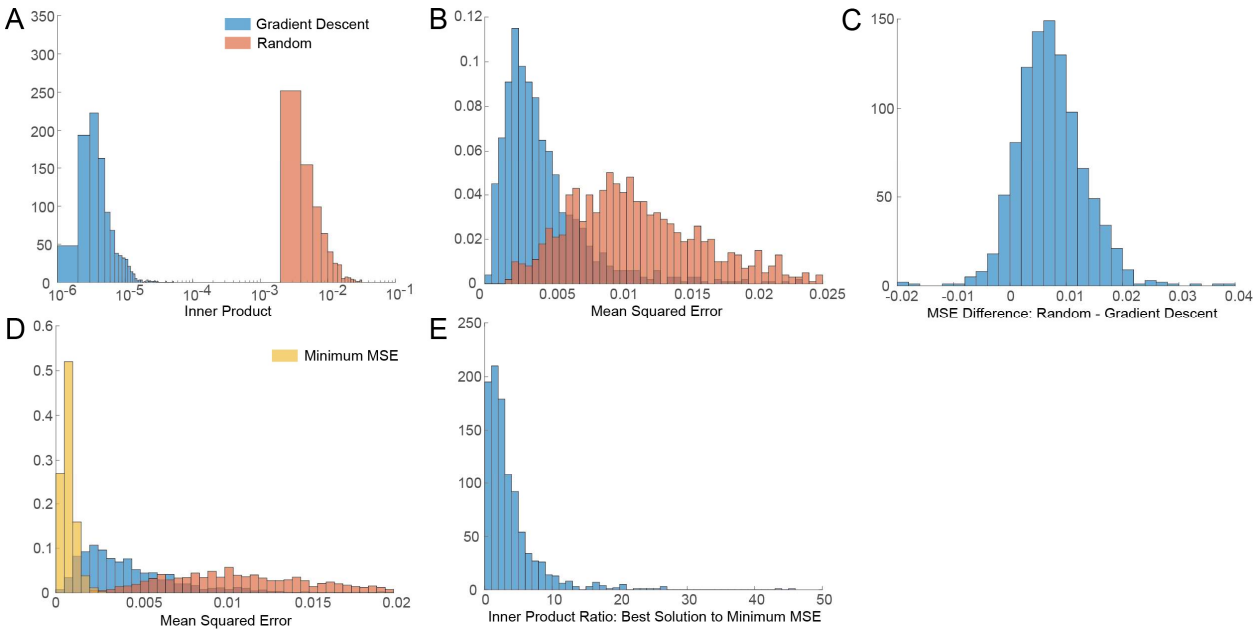

Figure 12-5 : For 1000 randomly generated weight vectors ("simulated vectors"), we used gradient descent to find the approximation ("approximated vectors") that minimized the L2 norm ("inner product") between the RNA editing distribution corresponding to the simulated vectors ("simulated distributions") and the RNA editing distribution corresponding to the approximated vectors ("approximated distributions"). We refer to the L2 norm between the distributions as the inner product to distinguish it from the L2 norm between the vectors, which we refer to as the mean squared error (MSE). **(A)** The inner product between simulated distributions and approximated distributions is shown in blue. By contrast, the inner product between simulated distributions and other random distributions is shown in orange. **(B)** The mean squared error between the simulated vectors and approximated vectors is shown in blue. By contrast, the inner product between the simulated distributions and other random distributions is shown in orange. Note that a substantial number of random weight vectors have lower mean squared error than the approximated vectors. This is possible because the noise in the basis distribution set used to generate the approximated distributions from the approximated vectors is different from the noise in the basis distribution set used to generate the simulated distributions from the simulated vectors, so the minimum of inner product between the simulated and approximated distributions is not always the same as the minimum of the MSE between the simulated and approximated vectors. **(C)** Another visualization of (B). For each simulated vector, we calculated both an approximated vector and a random vector. The difference in MSE between the approximated and random vectors is shown. Negative values correspond to test vectors for which the associated random vector was a better approximation to the simulated vector than the approximated vector. **(D)** Blue and orange bars are the same as in (B). Yellow bars correspond to the minimum MSE among all of the solutions found by gradient descent for a given test vector, indicating that the inner product minima found by the gradient descent are not in general minima of the MSE. **(E)** The difference in the inner product between the solutions with the minimum MSE found by gradient descent, and the solutions with the minimum inner product, as a fraction of the minimum inner product. The solutions with the minimum MSE discovered by gradient descent often have inner products several fold higher than the solution with the minimum inner product.

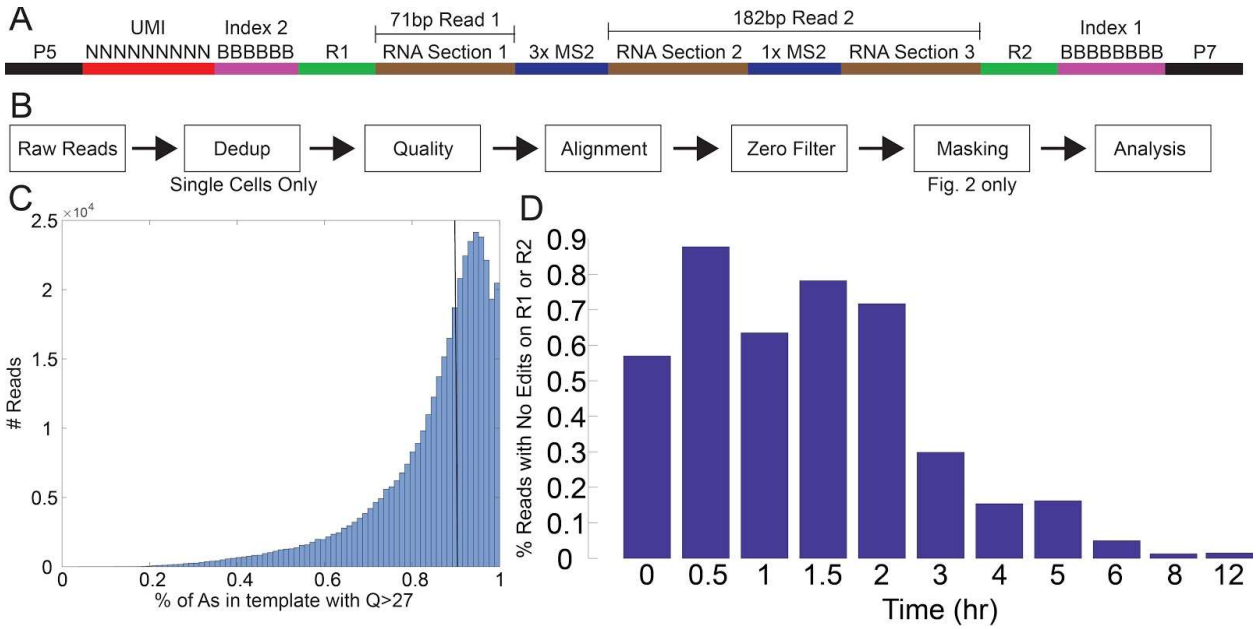

Figure 12-6: **(A)** The read structure of the repRNA is shown. **(B)** A schematic of the analysis pipeline is shown. See Methods. **(C)** For one replicate from the experiment in Figure 6-1E a histogram of the number of reads with a given percentage of As with Q score >27 is shown. This includes all sites that are As on the repRNA template, i.e., it also counts Gs that are read at positions that are A on the template. The black line indicates the 90% cutoff, which was applied to all analysis. **(D)** For one replicate from the experiment in Figure 6-1E, the percentage of reads having no edits in either R1 or R2 is shown as a function of time. These reads were excluded from analysis, except where otherwise stated in Figure 12-2.

Table 12-1: List of plasmids used in this study. This list excludes pCMV Tet3G, which is available commercially from Clontech.

Num	Name	Description	Used in
116v1	pAAV-Ef1a-MCP-dmADARE488Q	Fusion of MS2 coat protein to Drosophila ADAR E488Q, under Ef1a promoter, with WPRE	Supp. Fig 1B,C
116v5	pAAV-Ef1a-MCP-huADARE488QT490A	As with 116v1, but Human ADAR2 E488QT490A	All Figures
116v6	pAAV-Ef1a-MCP-huADART490A	As with 116v1, but Human ADAR2 T490A	Supp. Fig. 1B,C
133	pcDNA3.1-GAVPO	GAVPO (VIVID transactivator) expressed under the CMV promoter in the pcDNA3.1 backbone.	Supp. Fig. 3,4
147B1	pTRE3G-iRFP-B1-repRNA_A	repRNA Template A inserted into the 3' UTR of iRFP between a bActin Zipcode element and a WPRE element, in the pTRE3G backbone, with RNA barcode TGC. Also includes a xrRNA element in the 5' UTR.	Fig.1,2,3,4,Supp. Fig. 1,2,4.
148B1	pTRE3G-iRFP-B1-repRNA_B	Same as 147B1, but with RNA Template B.	Supp. Fig. 1
149B1	pLenti-5xUASG-iRFP-B1-repRNA-A	repRNA Template A inserted into the 3' UTR of iRFP between a bActin Zipcode element and a WPRE element, in a second generation lentiviral backbone with the Vivid promoter, with RNA barcode TGC. Also includes a xrRNA element in the 5' UTR.	Supp. Fig. 3
149B3	pLenti-5xUASG-iRFP-B3-repRNA-A	Same as 149B1, but with RNA barcode CTG.	Supp. Fig.4.
187	pTRE3G-c-fos-iRFP-B3-repRNA-A	Same as 147B1, with the TRE promoter removed and replaced with a c-Fos promoter from pAAV-cFos-EYFP (Addgene 47907), and with RNA barcode CTG.	Fig.5

Table 12-2: List of oligos used in this study.

Name	Description	Sequence
SGR-174B-1	Barcoded RT Primer with 3bp barcode	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNN CCT GCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-2	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNN GAG GCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-3	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNN TTA GCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-4	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNN AGC GCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-5	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNN AAT GCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-6	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNN CAA GCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-7	Barcoded RT primer with 6 base barcode	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNAGTGTCGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-8	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNTATCCGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-9	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNCATTTGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-10	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNATGCTAGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-11	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNCCGTGGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-12	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNATGAGTGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-13	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNCGAGCAGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-14	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNCGCGGCGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-15	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNACTTATGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-16	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNTGCATGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-17	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNAGTAGGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-18	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNGTTGACGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-19	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNTATCACGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-20	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNCCCTAGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-21	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNGCCCGTGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG

SGR-174B-22	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNTTCCCGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-23	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNCATATAGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-24	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNAACGCCGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-25	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNAGGTTGGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-174B-26	“”	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNTCAATAGCG AGG CCC GCATCTTTCACAAATTTTGTAATCCAGAGG
SGR-175	Custom Read 1	GCG AGG CCC GCA TCT TTC ACA AAT TTT GTA ATC CAG AGG
SGR-175-RC	Custom Index 2	CCTCTGGATTACAAAATTTGTGAAAGATGCGGGCCTCGC
SGR-176	Barcoded PCR primer	CAAGCAGAAGACGGCATACGAGAT ACTGGTCA AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-2	“”	CAAGCAGAAGACGGCATACGAGAT GTGTTCGT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-3	“”	CAAGCAGAAGACGGCATACGAGAT TAACTGTT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-4	“”	CAAGCAGAAGACGGCATACGAGAT GATTGGTG AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-5	“”	CAAGCAGAAGACGGCATACGAGAT GGAGAGAG AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-6	“”	CAAGCAGAAGACGGCATACGAGAT TGAGCGAT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-7	“”	CAAGCAGAAGACGGCATACGAGAT CCTCCGTT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-8	“”	CAAGCAGAAGACGGCATACGAGAT AACATATT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-9	“”	CAAGCAGAAGACGGCATACGAGAT CTTACGTA AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-10	“”	CAAGCAGAAGACGGCATACGAGAT TGACGTAG AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-11	“”	CAAGCAGAAGACGGCATACGAGAT CTATGTAT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-12	“”	CAAGCAGAAGACGGCATACGAGAT TTTGCAGA AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-13	“”	CAAGCAGAAGACGGCATACGAGAT GGTAGCGA AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-14	“”	CAAGCAGAAGACGGCATACGAGAT ACGGGTTT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-15	“”	CAAGCAGAAGACGGCATACGAGAT TAAACCTC AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-16	“”	CAAGCAGAAGACGGCATACGAGAT GAGAACTG AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG

SGR-176-15	“”	CAAGCAGAAGACGGCATACGAGAT GGTTTGAT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-18	“”	CAAGCAGAAGACGGCATACGAGAT TAGATTAT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-19	“”	CAAGCAGAAGACGGCATACGAGAT AAGGTTAG AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-176-20	“”	CAAGCAGAAGACGGCATACGAGAT CCGAAAAT AAG TTA CTA TCG AAATGCCCTGAGTCCACCCCGG
SGR-177	Custom Read 2	AAG TTA CTA TCG AAA TGC CCT GAG TCC ACC CCG G
SGR-177-RC	Custom Index 1	CCGGGGTGGACTCAGGGCATTTCGATAGTAACTT

Table 12-3: List of RNA editing templates used in this study. The following sequences are the sequences that were analyzed for RNA editing. Notes are supplied as a courtesy to follow-on studies, and no representations are made as to their accuracy or reproducibility.

	Sequence	Notes
A_Short	AGTACGCGTTAGATTAGATTAGATTAGATTAGAT TAGATTAGAAAAATTAATACGTACACCATCAGG GTACGTCTCAGACACCATCAGGGTCTGTCTGGTA CAGCATCAGCGTACCATATATTTTTTCCAATCCA ATCCAATCCAATCCAATCCAATCCAAATAGATCC TAATCA	
A	TTAGATTAGATTAGATTAGATTAGATTAGATTAG AAAAATTAATATACGTACACCATCAGGGTACGTC ATATATTTTTTCCAATCCAATCCAATCCAATCCA ATCCAATCCAATACGCGTTAGATTAGATTAGATT AGATTAGATTAGATTAGAAAAATTAATACGTAC ACCATCAGGGTACGTCTCAGACACCATCAGGGTC TGTCTGGTACAGCATCAGCGTACCATATATTTTT TCCAATCCAATCCAATCCAATCCAATCCAATCCA AATAGATCCTAATCA	
B_Short	AGTACGCGTTAGATTAGATTAGATTAGATTAGAT TAGATTAGAAAAATTAATACGTACACCATCAGG GTACGTCTCAGACACCATCAGGGTCTGTCTGGTA CAGCATCAGCGTACCATATATTTTTTCTAATCTA ATCTAATCTAATCTAATCTAATCTAAATAGATCC TAATCA	
B	TTAGATTAGATTAGATTAGATTAGATTAGATTAG AAAAATTAATATACGTACACCATCAGGGTACGTC ATATATTTTTTCTAATCTAATCTAATCTAATCTAA TCTAATCTAAACGCGTTAGATTAGATTAGATTAG ATTAGATTAGATTAGAAAAATTAATACGTACACC ATCAGGGTACGTCTCAGACACCATCAGGGTCTGT CTGGTACAGCATCAGCGTACCATATATTTTTTCT AATCTAATCTAATCTAATCTAATCTAATCTAAAT AGATCCTAATCA	
C	AGTACGCGTTAAATTATATTAACTAAATTATAGA TTAACAAGAATATTAAATACGTACACCATCAGG GTACGTCTCAGACACCATCAGGGTCTGTCTGGTA CAGCATCAGCGTACCTATTTAATATTCTTGTTAA TCTATAATTTAGTTAATATAATTTAAATAGATCC TAATCA	This template shows significant background editing by endogenous ADAR enzymes, even in the absence of trans-expression of ADAR. It also showed extremely rapid editing on a timescale of single minutes in the presence of blue light, when MCP-Cry2 and CIBN-dmADARE488Q were co-expressed.
D	AGTACGCGATTGGTTAATCCCATTGGTTAATCCC ATTGGTTAATCCCTTAATACGTACACCATCAGGG TACGTCTCAGACACCATCAGGGTCTGTCTGGTAC AGCATCAGCGTACCATATATGGGTTAAACTGATG	Editing on this template showed significant sensitivity to the identity of the N-terminal fusion. MCP-ADAR was able to edit this

	GGTTAAACTGATGGGTTAAACTGATATAGATCCT AATCA	template, whereas other ADAR enzymes, like a CIBN-ADAR fusion, were unable.
E	AGTACGCGAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAACGTACACCATC AGGGTACGTCTCAGACACCATCAGGGTCTGTCTG GTACAGCATCAGCGTACCTTTTTTTTTTTTTTTTT TTTTTTTTTTTTTTTTTTTTTTTTTTTTATAGATCCT AATCA	This template was always severely underrepresented in sequencing, either due to difficulties with expression, amplification, or alignment.

Chapter 13
Appendices to Chapter 7

Appendix 1: Simulations of recombination cassette diversity

For 4-value cassettes (two inversion units) we expect to obtain each possible recombination outcome with equal probability in the limit of many recombination operations, since both units have an equal probability of excision. For 8-value cassettes (four inversion units), this is no longer the case, because there are more ways to excise the internal cassettes. In Figure 13-1, we simulated 10^6 independent 8-value cassettes each undergoing a number of recombinations chosen from a Poisson distribution with mean 3, 4, 5, 6, 7 or 20. For > 10 recombinations per cassette we see convergence to the equilibrium distribution (infinite recombinations per cassette) with a Shannon entropy of 2.9 bits, as compared to 3 bits for a uniform distribution. (Similar simulations show that for a 4-value, 2-inversion-unit cassette, we achieve a nearly uniform distribution of values once > 4 recombinations occur per cassette.)

Other factors (such as the distance between recombination sites) may also introduce bias into the recombination process, which may decrease the entropy provided by recombination. For this reason, we have assumed that 8-value cassettes generate 2.8 bits of information, even in the long-time limit. However, it is important to emphasize that the Shannon entropy is relatively robust against deviations from uniformity, and that the expected error rate increases only linearly with the barcode degeneracy, so significant deviations from uniformity would be required to affect the error rate in a major way.

The appropriateness of the Shannon entropy for evaluating the number of effective barcodes derives from its interpretation as a measure of compressibility. If a probability distribution over B bits has a Shannon entropy of D bits, then any sufficiently-long sequence of values drawn from the B-bit non-uniform distribution over barcodes can be mapped without loss of information onto a sequence of values drawn from a uniform D-bit probability distribution of the same length. The axon tracing problem encountered here for an 8-value cassette involves disambiguating a neuron of interest from a sequence of neurons with barcodes drawn from a non-uniform probability distribution over $3C$ bits, with $2.8C$ bits of entropy. Thus, for the purpose of the average barcode degeneracy (discussed further in Appendix 4 of this chapter), we may equivalently behave as though the barcodes encountered during axon tracing were drawn from a uniform probability distribution over $2.8C$ bits.

Appendix 2: Possible methods for increasing the achievable genetic diversity

In order to increase the amount of achievable genetic diversity for a given number of orthogonal recombination sites, it is necessary either to increase the number of values per cassette or to

implement a system that allows for the same recombination site to be used on multiple different cassettes. We consider one strategy in each category.

Exponential scaling by eliminating excisions

In the standard strategy, a cassette coding for m values can produce one of m proteins (via m RNAs), but if a system with exponential scaling could be found, a cassette coding for m values could produce $2m$ proteins. One approach to achieve exponential scaling would be to eliminate the excision events altogether, for example by using Rci recombinase (*289*, *338*), a recombinase which inverts but only very rarely excises, as proposed by Zador. If the only available operation were flipping, then a cassette with 4 flippable value registers (requiring 4 epitopes) would be capable of producing $2^4 = 16$ possible molecular strings of those epitopes. If a cassette could be produced with 8 flippable value registers (requiring 8 epitopes), it would be capable of producing $2^8 = 256$ possible strings of epitopes. In this case, dedicating all available epitope/fluorophore pairs to values, C cassettes could produce $\binom{256}{C}$ phenotypes, in which case 6 cassettes would be sufficient to barcode any animal.

Temporal Multiplexing

Rather than limiting the multiplexing of cassettes to the number of orthogonal recombination sites available, one possibility is to use temporal multiplexing, re-using recombination sites across successive temporal cycles.

For example, to double the effective number of sites, one could use inducible DNA binding proteins (*339–345*), such as CRISPRs – i.e., programmable gRNAs + nuclease-deficient Cas9 (dCas9) – first blocking recombination at a first copy of each site and then blocking recombination at a second copy. For any given LoxP site, two cassettes using that site could be inserted into the genome, including LoxP-overlapping gRNA binding sites containing unique flanking sequences shared by all sites that are to recombine in a given temporal cycle. Initially, the gRNA corresponding to the first of the two cassettes would be expressed, blocking access to this first cassette by Cre recombinase. Upon induction of Cre, Cre would access the second of the two cassettes, permitting recombination there. Next, production of the second gRNA would be induced, blocking access of Cre to the second cassette. After a sufficient delay, induction of the first gRNA would be removed, allowing Cre to access the first cassette. This would prevent crosstalk between the two cassettes, while allowing recombination to occur at each one individually. Alternatively, these systems could also be made to progress autonomously by triggering the production of the gRNAs for the Nth phase only after recombination in the $(N-1)$th phase is completed.

Conservatively, at least six underlying orthogonal cassettes could be achieved using only published LoxP and Frt sites, although more may be available (*184*). Twelve cassettes could be achieved using 3 Cre sites, 3 Flp sites sites and a manually-induced dCas9 temporal multiplexing strategy to double the number of effective sites from 6 to 12. Finally, 18 cassettes could be achieved using a total of 9 underlying LoxP, Frt and other orthogonal sites, plus a two-step temporal multiplexing approach.

Note that it may also be possible to re-use identical recombination sites across multiple cassettes (i.e., to give up orthogonality between cassettes), although at the risk of inter-cassette crosstalk. For example, BrainBow has often used multiple genome-integrated copies of the same cassette to drive analog color addition, e.g., roughly 16 copies of the same BrainBow cassette (*291*).

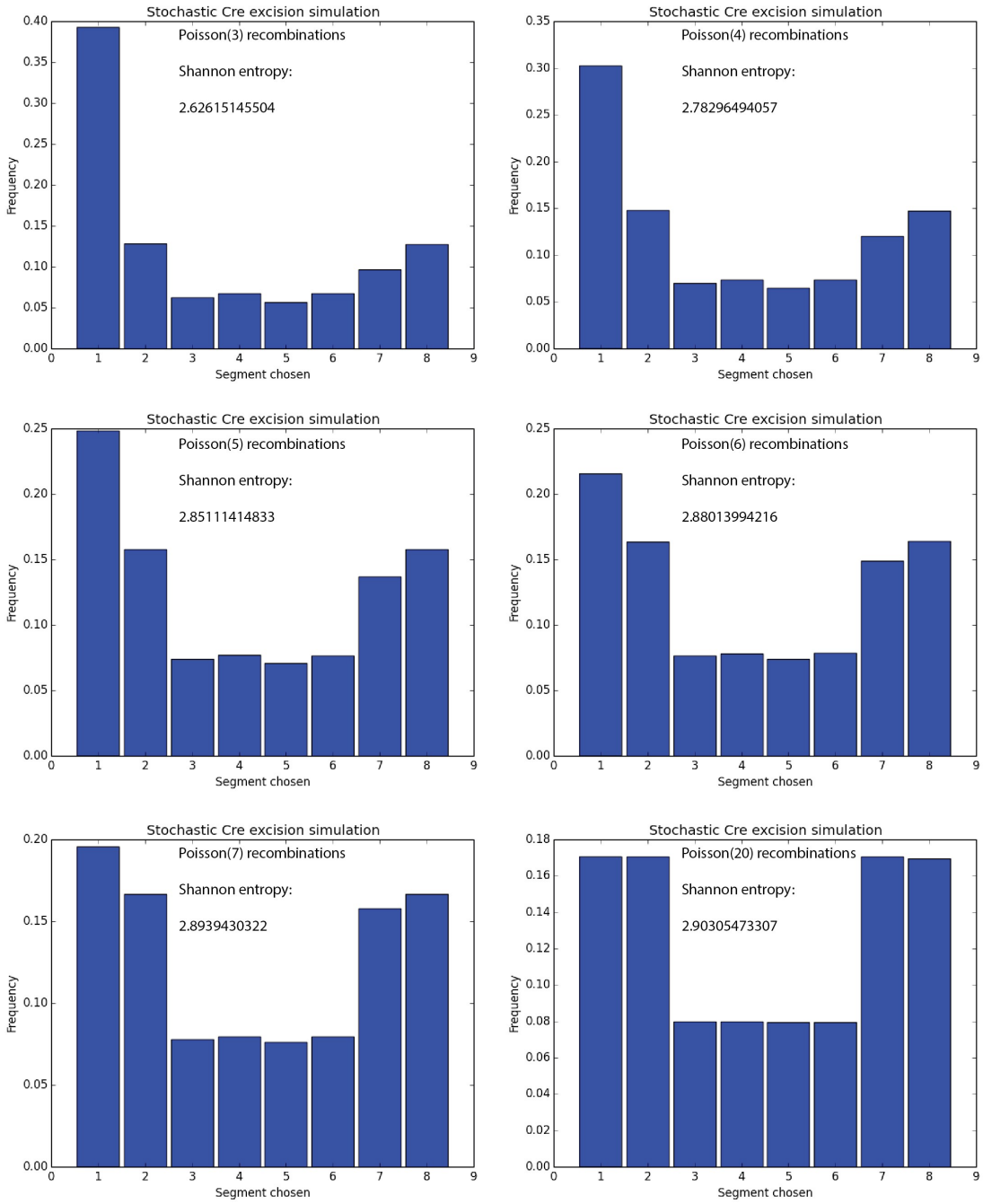

Figure 13-1: Diversity generation from an 8-value cassette for different amounts of recombinase activity. Simulations of 10^6 cassettes undergoing recombination are shown for the cases in which the number of recombinations performed is drawn from a Poisson distribution with an average of 3 (top left), 4 (top right), 5 (middle left), 6 (middle right), 7 (bottom left), and 20 (bottom right). After determining the number of recombinations to be performed, the actual recombination events were determined by randomly choosing an available pair of recombination sites and performing the corresponding operation. In the limit of many recombination events, the resulting distribution has 2.9 bits of entropy; in the most pessimistic case of 3 recombination events, the distribution has 2.6 bits of entropy.

Appendix 3: Possible fluorescent imaging scheme with >10 spectrally orthogonal colors

Based on Figure 13-2, we anticipate that it would be possible to perform multicolor imaging with up to 12 channels on a customized microscope with 8 Coherent lasers and a Zeiss 34-channel QUASAR detection unit. The 12 fluorophores depicted in Figure 13-2 have been chosen for minimal spectral overlap, although some optimization may be required, especially in the range of the 405nm laser line4. Likewise, we anticipate that 10-color imaging would be possible on a 6-laser system, although this would require a relatively unusual laser line at 350nm (such as the Obis LG). Alternatively, it would be possible to achieve 10 color fluorescent imaging on a more standard 6 laser system with laser lines at 405 nm, 445 nm, 505 nm, 561 nm, 647 nm, and 685 nm, if a long Stokes-shift dye could be found with an excitation maximum at 560 nm.

Note that in order to have cross-talk between two fluorophores, they must overlap in both their excitation spectra and in their emission spectra. Thus, although Atto 430LS is excited significantly by the Obis LG, Obis 405LX, and Obis 445LX, it can be spectrally distinguished from all other fluorophores excited by those lasers by virtue of its long emission wavelength.

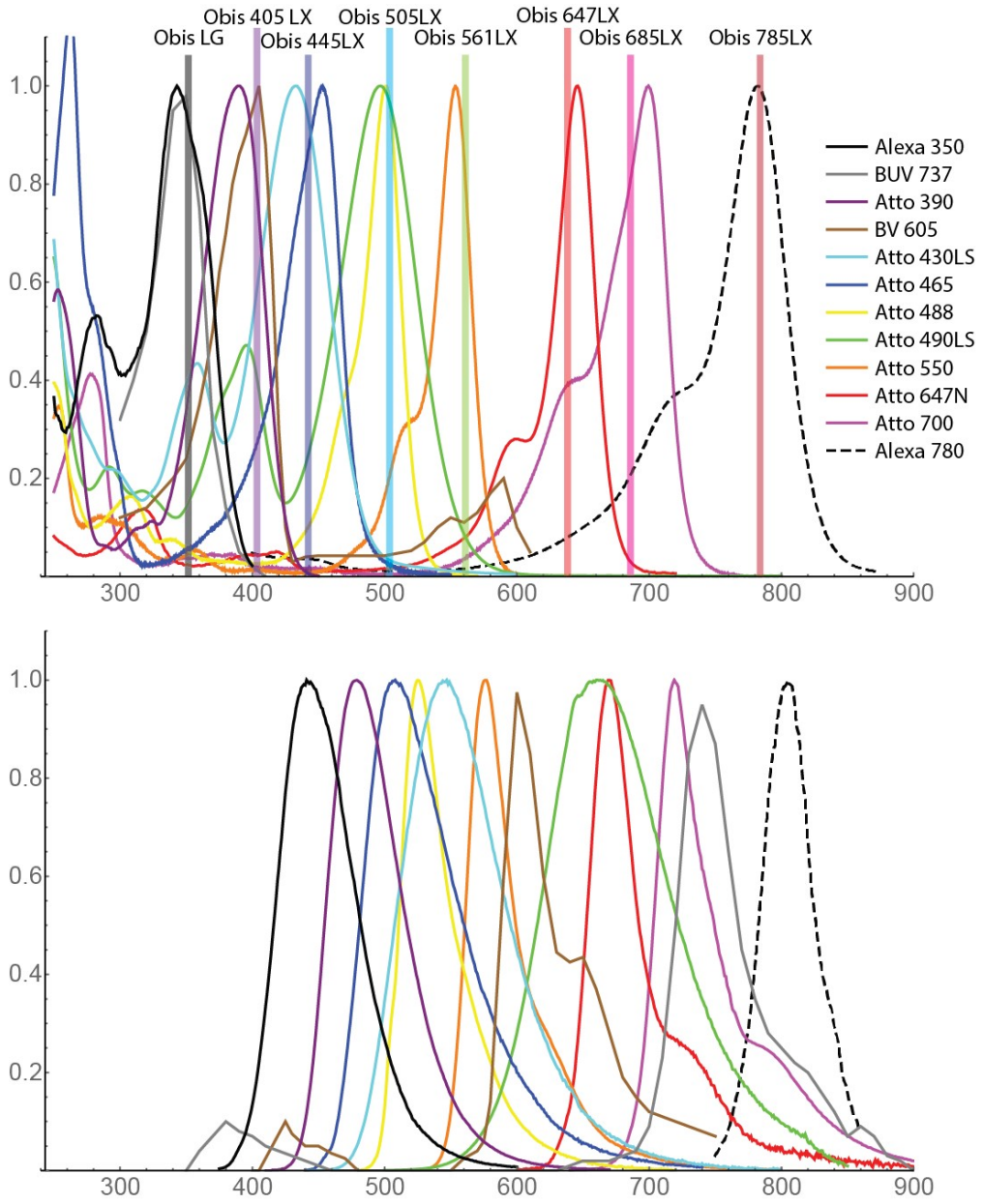

Figure 13-2: An example 12-color strategy with 8 lasers. Using an 8-laser system from Coherent and a Zeiss QUASAR detection unit, we anticipate that it would be possible to detect up to 12 orthogonal colors at the single molecule level. A set of 12 prospective fluorophores are shown along with the laser lines that would be used. In the present system, we would expect significant cross-talk between BV605 and Atto 430LS and between Atto 390 and Atto 425 due to the excitation of Atto 425 and Atto 430LS by the Obis 405LX laser. This problem could be resolved if a laser were available with a line at 390nm rather than 405 nm. There also appears to be overlap between Atto 490LS and BV605 due to non-negligible absorption of Atto 490LS at 400 nm, but in preliminary experiments we were able to distinguish between these two dyes at the single molecule level. Depending on the severity of the overlap, we expect that this system would allow imaging in at least 9 orthogonal channels, and potentially in 12 orthogonal channels.

Appendix 4: Axon tracing vs. unique barcoding

In the ideal case, the number of barcodes generated by a phenotyping scheme is large enough that every cell in the brain receives a unique barcode with high probability (a generalized instance of the "birthday problem") (*346*). In order for this to occur with high likelihood, we need $P(m,n) = n! \times \text{Binomial}(m,n)/m^n$, the probability of no two identical barcodes chosen from a barcode pool of size m when sampling n neurons, to be close to 1. In this case, assuming the readout can be done faithfully, there is no chance of making errors in morphological tracing of neural projections, since the color code can be used to error-correct any tracing error.

However, if color-based barcoding is combined with automated segmentation algorithms, it may be possible to achieve acceptably low error rates even if multiple neurons share the same barcode. Based on the performance of automated segmentation algorithms as currently applied to EM connectomics, we can crudely estimate the tracing error rate as a function of the barcode degeneracy.

The probability of making an error in tracing is a function of the degeneracy of the barcode attached to the neuron, i.e., the total number of neurons in the brain that share the barcode. For a neuron with degeneracy M – i.e., that neuron's barcode appears in a total of M neurons across a brain of size N neurons – the probability that any given neuron it encounters has the same barcode can be estimated as $(M-1)/N$. Thus, for each tracing error that a given automated tracing algorithm would make in a grayscale image, we can reasonably expect that the probability that the same algorithm would make the same error in a color-coded image is $(M-1)/N$.

If the number of errors made per unit length in a grayscale image is ϵ, then as a function of the degeneracy M, the probability that an error is made at least once in tracing a projection of length L is

$$\boldsymbol{P_E(M,L) = 1 - \left(1 - \frac{M-L}{N}\right)^{\epsilon L} \approx \frac{\epsilon L(M-1)}{N}} \tag{49}$$

The approximation holds assuming that $M \ll N$. Thus, if there are R projections per neuron, the expected number of incorrectly traced projections (i.e., projections with at least one error) in the entire brain is given by

$$\boldsymbol{\langle E \rangle = \epsilon \langle L \rangle R(\langle M \rangle - 1)} \tag{50}$$

In a recent EM connectomics study in mouse cortex (*299*), the authors found that there were on average 200 profiles per cubic micron, and that currently available automated tracing algorithms made an average of 7 errors per cubic micron. Given this, we can take ϵ to be on the order of 7 errors per 200 microns of axonal length, or 35 errors per millimeter of axonal length (1 error per $29\ \mu\text{m}$), for current automated tracing algorithms as per (*299*). For axons, we take $\langle L \rangle \sim 1\ \text{mm}$

and $R \sim 1$, in which case we find that the expected number of axon tracing errors across a brain is, very roughly,

$$\langle E \rangle_{\mathbf{axon}} = 35(\langle M \rangle - 1) \tag{51}$$

On the other hand, if we take $R \sim 100$ for dendrites and $L \sim 100\ \mu\mathrm{m}$, we have, in the entire brain, a rough estimate of

$$\langle E \rangle_{\mathbf{axon}} = 350(\langle M \rangle - 1) \tag{52}$$

Remarkably, these numbers are independent of the number of neurons in the brain, which can be understood by the fact that, as the number of neurons increases, the probability that neurons with the same barcode encounter each other decreases for fixed projection number and length, in our crude error model.

Appendix 5: Peptide vs. RNA implementations

We have chosen to use peptide epitopes here to achieve high labeling densities, since peptides can be expressed to high levels. However, RNA implementations of similar address-value barcoding schemes are also possible, and RNA readout is simplified by the ability to use hybridization probes – which can be made extremely specific (*347*) and orthogonal (*348*) – and/or direct in-situ sequencing (*290*).

FISH-based RNA ColorCodes: Instead of probing protein epitopes with antibodies, we can use the identical genetic diversification strategy (top of Figure 7-2) but probe the resulting RNAs with FISH probes. This can be done using a single round of FISH probing with K spectrally distinct fluorescently tagged hybridization probes, much as proposed for immuno-labeling of peptide epitopes above, or alternatively it can be done across multiple wash cycles of sequential FISH (*10*, *297*) using a smaller number of colors (e.g., only 2 or 4 fluorescent colors imaged per wash cycle).

FISSEQ-based RNA ColorCodes: Fluorescent in-situ sequencing (FISSEQ) can also be employed to read out such RNA barcodes. The address will then become a FISSEQ-sequencable short RNA string, unique for each cassette, while the possible values will be short RNA strings shared across cassettes with a given value stochastically chosen for each cassette upon recombination. The main constraint here is that the currently-demonstrated FISSEQ read length is roughly 30 base-pairs, whereas individual LoxP sites themselves take up 34 base-pairs. We can solve this using a design like the following: (PanNeuronal Promoter) (Primer Binding Site 1) (Address) (First LoxP site) (Values), where the possible values resulting from recombination are (Primer Binding Site 2)(SeqA) and (Primer Binding Site 2)(SeqB), i.e., all possible RNAs contain (Primer Binding Site 2) directly adjacent to the chosen value (either SeqA or SeqB). The proposal is to run FISSEQ with two targeted sequencing primers, first one and then the other. Primer 1 sequences from (Primer Binding Site 1) into the (Address). Primer 2 sequences from (Primer Binding Site 2) into

one of (SeqA) or (SeqB) depending on which recombination event occurred in the genome-integrated cassette. Note that we can readily achieve a huge address space here, since sequencing even 3 address bases gives $4^3 = 64$ addresses, and therefore the number of values per cassette can be very limited, e.g., there are only two possible values in the above design. An alternative approach would be to leverage the single-base resolution of FISSEQ by making the Value strings also encode an Address for their parent cassettes. Then FISSEQ from only one primer would be needed.

Appendix 6: Directed vs. random spatial separation

It may be possible to increase the effective number of "addresses" for a fixed number of fluorophores by directing certain labels to certain sub-cellular compartments which can be spatially resolved from one another at the high spatial resolutions considered here. Compartments such as the membrane, microtubules, actin filaments, spectrin structures (*349*), mitochondria or simply the intracellular space (cytosol) are worth considering as potential independent domains to enhance multiplexing capacity. The phenotype is read out by observing the presence or absence of each of the labels from each of the structures. In this way, with s structures and a labeling scheme that would generate b bits of information in the absence of structure-based spatial multiplexing, we can raise the number of bits per ROI to $s \times b$. The $15\ \text{nm}$ resolution of next-generation ExM should have sufficient spatial resolution to resolve such distinct subcellular structures and attribute the presence or absence of each label from each structure. Note that many of these components form precise intracellular geometric structures in the axon and in dendrites, e.g., actin rings (*349*), which could be useful in identifying and spatially resolving them. Moreover, there are well-known mouse lines with e.g. GFP-labeled actin filaments and microtubules, not to mention genetically encoded membrane and cytosolic labels. Of course, increasing the effective address space in this way requires an increase in the generated genotypic diversity, e.g., more orthogonal recombination sites.

Application of directed spatial multiplexing to Drosophila/Zebrafish: A particularly interesting limiting case of this approach for Drosophila or larval Zebrafish uses just two resolvable spatial compartments, e.g., membrane vs. microtubules, and devotes all fluorophores to values rather than addresses, as shown in Figure 13-3. This scheme requires 10 epitope-fluorophore pairs and 10 orthogonal recombinase sites applied to 4-value cassettes, and should thus be realizable with existing technologies. Because 4-value cassettes are expected to yield 2 bits of information per cassette, this approach yields $2^{20} = 1{,}048{,}576$ effective cell labels. Using this strategy, we would expect that in a Zebrafish or Drosophila brain with roughly 100k neurons, approximately 9% of available barcodes will be used, with 91% of neurons receiving a unique barcode, and 8.6% of neurons receiving a barcode that appears twice. See Appendix 4 of this chapter for a discussion of error rates in the scenario where not all neurons receive unique barcodes.

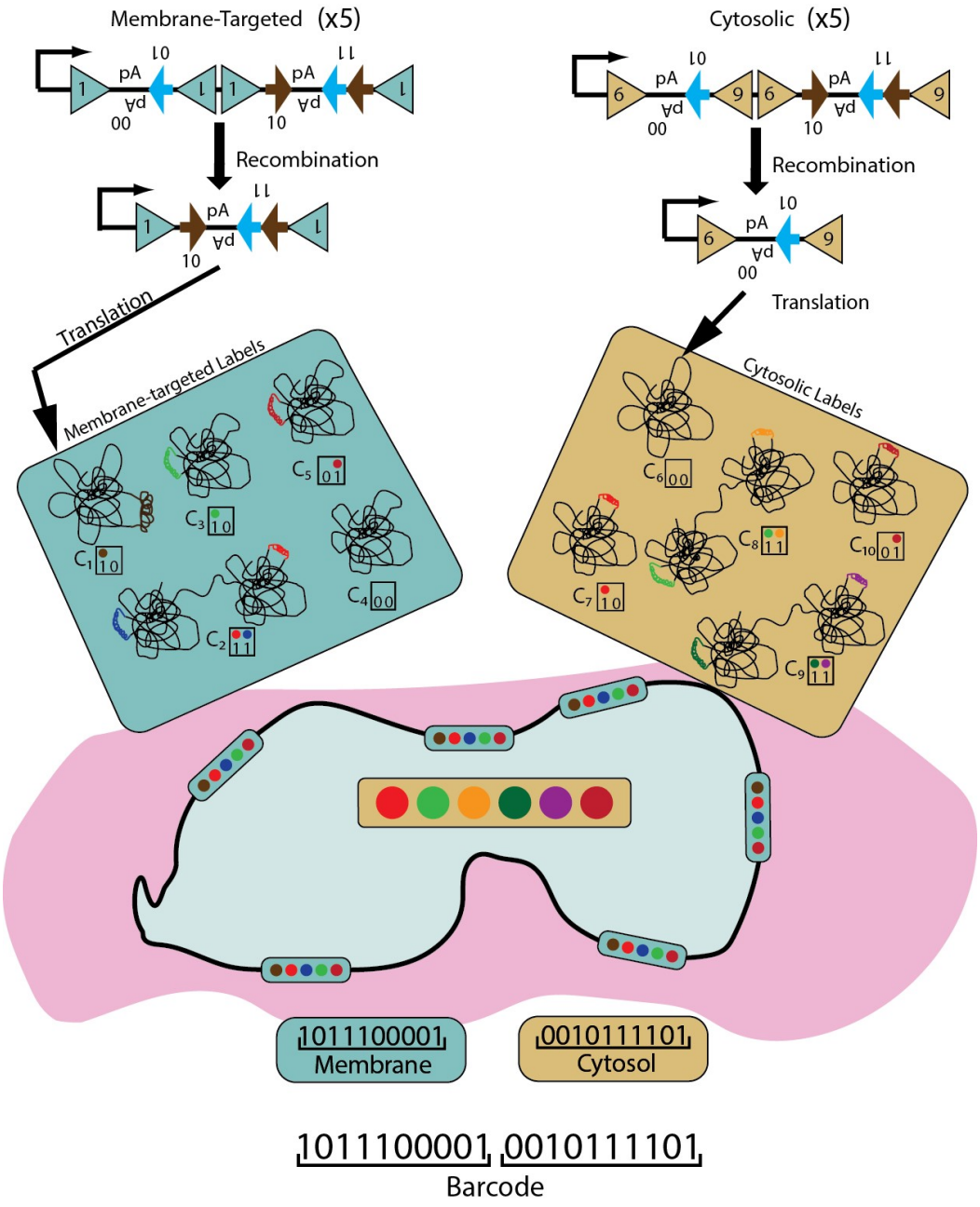

Figure 13-3: Structural Barcoding for Whole Drosophila or Zebrafish Brains. This method uses 10 orthogonal recombination sites, 10 orthogonal fluorophores, 10 epitopes/antibodies and two spatially resolvable cellular compartments (e.g., membrane and cytosol or membrane and microtubules) to uniquely barcode the Zebrafish or Drosophila brain. Each recombination site can generate four equi-probable epitope displays by producing either a scaffold without any epitopes, a scaffold with one of two epitopes, or a fusion of two scaffolds, each with an epitope. Cassettes differ only in the protein's target (e.g., membrane or cytosol/microtubules) and in the epitopes displayed on the protein. The membrane and cytosol/microtubules both receive 5 proteins, leading to a readout diversity of $(4^5)^2 = 1048576$ barcodes. This is sufficient to uniquely label 90% of neurons in the Zebrafish or Drosophila brains. If this strategy were expanded to have 12 orthogonal fluorophores and 12 epitopes/antibodies, it would produce 16.8 million barcodes.

The cassette design follows a standard BrainBow design with two successive inversion units. Five cassettes each produce one of four proteins that are directed to the membrane, while the remaining five cassettes each produce one of four proteins that are targeted to some other spatially-resolvable non-membrane compartment such as the microtubule cytoskeleton. The proteins produced are either scaffolding proteins labeled with a single epitope (when the corresponding bit string is 01 or 10); scaffolding proteins without epitopes (when the corresponding bit string is 00); or a fusion of two scaffolding proteins, each with a single epitope, connected by a flexible linker (when the corresponding bit string is 11). Each protein produced in this way thus encodes 2 bits of information. Readout is performed by primary antibody staining against the epitopes, followed by staining with fluorescently labeled secondary antibodies. The sample can be imaged using confocal microscopy with 20x expansion. We anticipate that it would be possible to achieve 10 color fluorescent imaging on a 6 laser system (see Appendix 3, Chapter 13).

Chapter 14 Bibliography

1. H. S. Seung, U. Sümbül, Neuronal Cell Types and Connectivity: Lessons from the Retina. *Neuron*. **83**, 1262–1272 (2014).

2. H. Zeng, J. R. Sanes, Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Publ. Gr.* **18**, 530–546 (2017).

3. C. Liu, P. S. Kaeser, Mechanisms and regulation of dopamine release. *Curr. Opin. Neurobiol.* **57**, 46–53 (2019).

4. S. Reardon, A giant neuron found wrapped around entire mouse brain. *Nature*. **543**, 14–15 (2017).

5. T. R. Insel, S. C. Landis, F. S. Collins, The NIH BRAIN Initiative. *Science*. **340**, 687–688 (2013).

6. A. P. Alivisatos *et al.*, A National Network of Neurotechnology Centers for the BRAIN Initiative. *Neuron*. **88**, 445–448 (2015).

7. L. A. Jorgenson *et al.*, The BRAIN Initiative: developing technology to catalyse neuroscience discovery. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140164–20140164 (2015).

8. C.-H. L. Eng *et al.*, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*. **568**, 235–239 (2019).

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

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

11. X. Wang *et al.*, Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science.* **5691**, 1–18 (2018).

12. E. Murray *et al.*, Simple, Scalable Proteomic Imaging for High-Dimensional Profiling of Intact Systems. *Cell*. **163**, 1500–1514 (2015).

13. K. D. Micheva, S. J. Smith, Array Tomography: A New Tool for Imaging the Molecular Architecture and Ultrastructure of Neural Circuits. *Neuron*. **55**, 25–36 (2007).

14. P. W. Tillberg *et al.*, Protein-retention expansion microscopy of cells and tissues labeled using standard fluorescent proteins and antibodies. *Nat. Biotechnol.* **34**, 987–992 (2016).

15. T. Ku *et al.*, Multiplexed and scalable super-resolution imaging of three-dimensional protein localization in size-adjustable tissues. *Nat. Biotechnol.* **34**, 973-981 (2016) doi:10.1038/nbt.3641.

16. J. M. Kebschull, P. Garcia, A. P. Reid, I. D. Peikon, F. Dinu, High-throughput mapping of single-neuron projections by sequencing of barcoded RNA. *Neuron*. **91**, 975–987 (2016).

17. Y. Han *et al.*, The logic of single-cell projections from visual cortex. *Nature*. **556**, 51–56 (2018).

18. L. Huang *et al.*, High-throughput mapping of mesoscale connectomes in individual mice. *BioRxiv*, 1–6 (2018).

19. J. M. Kebschull, A. M. Zador, Cellular barcoding: lineage tracing, screening and beyond. *Nat. Methods*. **15**, 871–879 (2018).

20. M. Ahrens, M. Orger, D. Robson, J. Li, P. Keller, Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nat. Methods*. **10** (2013), doi:10.1038/NMETH.2434.

21. G. J. Broussard, R. Liang, L. Tian, Monitoring activity in neural circuits with genetically encoded indicators. *Front. Mol. Neurosci.* **7** (2014), doi:10.3389/fnmol.2014.00097.

22. M. B. Bouchard *et al.*, Swept confocally-aligned planar excitation (SCAPE) microscopy for high-speed volumetric imaging of behaving organisms. *Nat. Photonics*. **9**, 113–119 (2015).

23. D. R. Hochbaum *et al.*, All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nat. Methods* (2014), doi:10.1038/nmeth.3000.

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

25. E. Z. Macosko *et al.*, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. **161**, 1202–1214 (2015).

26. E. Adrian, *The Basis of Sensation* (1928).

27. E. Fernández *et al.*, Acute human brain responses to intracortical microelectrode arrays: challenges and future prospects. *Front. Neuroeng.* **7**, 1–6 (2014).

28. D. Prodanov, J. Delbeke, Mechanical and biological interactions of implants with the brain and their impact on implant design. *Front. Neurosci.* **10** (2016), doi:10.3389/fnins.2016.00011.

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

30. A. H. Marblestone *et al.*, Physical principles for scalable neural recording. *Front. Comput. Neurosci.* **7**, 137 (2013).

31. S. G. Rodriques *et al.*, Multiplexed neural recording along a single optical fiber via optical reflectometry. *J. Biomed. Opt.* **21** (2016), doi:10.1117/1.JBO.21.5.057003.

32. J. J. Jun *et al.*, Fully Integrated Silicon Probes for High-Density Recording of Neural Activity. *Nature*. **551**, 232–236 (2017).

33. P. R. Patel *et al.*, Insertion of linear 8.4 μm diameter 16 channel carbon fiber electrode arrays for single unit recordings. *J. Neural Eng.* **12** (2015), doi:10.1088/1741-2560/12/4/046009.

34. N. A. Steinmetz, C. Koch, K. D. Harris, M. Carandini, Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. *Curr. Opin. Neurobiol.* **50**, 92–100 (2018).

35. L. Luan *et al.*, Ultraflexible nanoelectronic probes form reliable, glial scar–free neural integration. *Sci. Adv.* **3**, 1–10 (2017).

36. D. Oran *et al.*, 3D nanofabrication by volumetric deposition and controlled shrinkage of patterned scaffolds. *Science*. **1285**, 1281–1285 (2018).

37. F. Chen, P. W. Tillberg, E. S. Boyden, Expansion microscopy. *Science*. **347**, 543–8 (2015).

38. E. S. Lein *et al.*, Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. **445**, 168–176 (2007).

39. N. Fortelny, C. M. Overall, P. Pavlidis, G. V. C. Freue, Can we predict protein from mRNA levels ? *Nature*. **547**, 582–587 (2017).

40. G. Gut, M. D. Herrmann, L. Pelkmans, Multiplexed protein maps link subcellular organization to cellular states. *Science.* **7042** (2018), doi:10.1126/science.aar7042.

41. V. Marx, Mapping proteins with spatial proteomics. *Nat. Methods*. **12**, 815–819 (2015).

42. R. M. Levenson, A. D. Borowsky, M. Angelo, Immunohistochemistry and mass spectrometry for highly multiplexed cellular molecular imaging. *Lab. Investig.* **95**, 397–405 (2015).

43. M. Baker, Reproducibility crisis: Blame it on the antibodies. *Nature*. **521**, 274–276 (2015).

44. S. Rodriques, A. Marblestone, E. Boyden, A theoretical analysis of single molecule protein sequencing via weak binding spectra. *PLoS One*, 1–25 (2019).

45. K. P. Kording, Of Toasters and Molecular Ticker Tapes. *PLoS Comput. Biol.* **7**, e1002291 (2011).

46. T. H. Kim, Y. Zhang, J. C. Jung, J. Li, H. Zeng, Long-Term Optical Access to an Estimated One Million Neurons in the Live Mouse Cortex. *Cell Rep.* **17**, 3385–3394 (2016).

47. W. Zong *et al.*, Fast high-resolution miniature two-photon microscopy for brain imaging in freely behaving mice. *Nat. Methods*. **14**, 713–719 (2017).

48. Y. Li *et al.*, Neuronal Representation of Social Information in the Medial Amygdala of

Awake Behaving Mice. *Cell*, 1–15 (2017).

49. B. F. Grewe *et al.*, Neural ensemble dynamics underlying a long-term associative memory. *Nature*. **543**, 670–675 (2017).

50. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).

51. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, The Structure of the Nervous System of the Nematode Caenorhabditis elegans. *Philos. Trans. R. Soc. B Biol. Sci.* **314**, 1–340 (1986).

52. G. Yan *et al.*, Network control principles predict neuron function in the Caenorhabditis elegans connectome. *Nature*. **550**, 519–523 (2017).

53. T. Kitamura *et al.*, Island cells control temporal association memory. *Science.* **343**, 896–901 (2014).

54. E. Hoshi, L. Tremblay, J. Féger, P. L. Carras, P. L. Strick, The cerebellum communicates with the basal ganglia. *Nat. Neurosci.* **8**, 1491–1493 (2005).

55. M. F. Bear, W. Singer, Modulation of visual cortical plasticity by acetylcholine and noradrenaline. *Nature*. **320**, 172–176 (1986).

56. L. A. Glantz, D. A. Lewis, Decreased Dendritic Spine Density on Prefrontal Cortical Pyramidal Neurons in Schizophrenia. *Arch. Gen. Psychiatry*. **57**, 65 (2000).

57. S. A. Irwin, R. Galvez, W. T. Greenough, Dendritic Spine Structural Anomalies in Fragile-X Mental Retardation Syndrome. *Cereb. Cortex*. **10**, 1038–1044 (2000).

58. T. Kim *et al.*, Human LilrB2 is a beta-amyloid receptor and its murine homolog PirB regulates synaptic plasticity in an Alzheimer's model. *Science.* **341**, 1399–1404 (2013).

59. A. Bhattacharya, U. Aghayeva, E. G. Berghoff, O. Hobert, Plasticity of the Electrical Connectome of C. elegans. *Cell*, 1174–1189 (2019).

60. M. Helmstaedter, Cellular-resolution connectomics: challenges of dense neural circuit reconstruction. *Nat. Methods*. **10**, 501–7 (2013).

61. J. Kornfeld, W. Denk, Progress and remaining challenges in high-throughput volume electron microscopy. *Curr. Opin. Neurobiol.* (2018), doi:10.1016/j.conb.2018.04.030.

62. M. Januszewski *et al.*, High-precision automated reconstruction of neurons with flood-filling networks. *Nat. Methods* (2018), doi:10.1038/s41592-018-0049-4.

63. J. L. Morgan, D. R. Berger, A. W. Wetzel, J. W. Lichtman, The Fuzzy Logic of Network Connectivity in Mouse Visual Thalamus. *Cell*. **165**, 192–206 (2016).

64. A. S. Chiang *et al.*, Three-dimensional reconstruction of brain-wide wiring networks in

drosophila at single-cell resolution. *Curr. Biol.* **21**, 1–11 (2011).

65. W. Denk, H. Horstmann, Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biol.* **2** (2004), doi:10.1371/journal.pbio.0020329.

66. D. Grant *et al.*, Whole-brain serial-section electron microscopy in larval zebrafish. *Nature*. **545**, 345–349 (2017).

67. K. Eichler *et al.*, The complete connectome of a learning and memory centre in an insect brain. *Nature*. **548**, 175–182 (2017).

68. Z. Zheng *et al.*, A Complete Electron Microscopy Volume of the Brain of Adult Drosophila melanogaster. *Cell*. **174**, 1–14 (2018).

69. K. L. Briggman, M. Helmstaedter, W. Denk, Wiring specificity in the direction-selectivity circuit of the retina. *Nature*. **471**, 183–8 (2011).

70. A. Narayanan Kasthuri *et al.*, Saturated Reconstruction of a Volume of Neocortex. *Cell*. **162**, 648–661 (2015).

71. A. A. Wanner, C. Genoud, T. Masudi, L. Siksou, R. W. Friedrich, Dense EM-based reconstruction of the interglomerular projectome in the zebrafish olfactory bulb. *Nat. Neurosci.* **19** (2016), doi:10.1038/nn.4290.

72. J. Kornfeld *et al.*, EM connectomics reveals axonal target variation in a sequence-generating network. *Elife*. **6**, 1–20 (2017).

73. J. Scholvin *et al.*, Guiding, Modulating, and Emitting Light on Silicon-Challenges and Opportunities. *Micromachines*. **9**, 1–21 (2018).

74. G. Dimitriadis, J. P. Neto, A. Aarts, A. Alexandru, M. Ballini, Why not record from every channel with a CMOS scanning probe ? *BioRxiv* (2018).

75. R. Gesteland, B. Howland, J. Lettvin, W. Pitts, Comments on Microelectrodes. *Proc. IRE*. **47**, 1856–1862 (1959).

76. G. Buzsáki *et al.*, Tools for Probing Local Circuits: High-Density Silicon Probes Combined with Optogenetics. *Neuron*. **86**, 92–105 (2015).

77. M. Hilbert, P. Lopez, The world's technological capacity to store, communicate and compute information. *Science*. **332**, 60–65 (2011).

78. M. R. Bower *et al.*, Intravenous recording of intracranial, broadband EEG. *J. Neurosci. Methods*. **214**, 21–26 (2013).

79. R. R. Llinás, K. D. Walton, M. Nakao, I. Hunter, P. A. Anquetil, Neuro-vascular central nervous recording/stimulating system: Using nanotechnology probes. *J. Nanoparticle Res.* **7**, 111–127 (2005).

80. J. Du, T. J. Blanche, R. R. Harrison, H. A. Lester, S. C. Masmanidis, Multiplexed, High Density Electrophysiology with Nanofabricated Neural Probes. *PLoS One*. **6**, e26204 (2011).

81. J. Jarzynski, R. P. De Paula, R. P. DePaula, E. Udd, Eds. (International Society for Optics and Photonics, 1987; http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.937498), vol. 718, p. 48.

82. W. V. Sorin, J. P. Dakin, A. D. Kersey, Eds. (International Society for Optics and Photonics, 1993; http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1001278), vol. 1797, pp. 109–118.

83. L. Palmieri, Distributed polarimetric measurements for optical fiber sensing. *Opt. Fiber Technol.* **19**, 720–728 (2013).

84. X. Bao, L. Chen, X. Bao, L. Chen, Recent Progress in Distributed Fiber Optic Sensors. *Sensors*. **12**, 8601–8639 (2012).

85. A. Liu *et al.*, A high-speed silicon optical modulator based on a metal–oxide–semiconductor capacitor. *Nature*. **427**, 615–618 (2004).

86. A. Liu *et al.*, High-speed optical modulation based on carrier depletion in a silicon waveguide. *Opt. Express*. **15**, 660 (2007).

87. Q. Xu, B. Schmidt, S. Pradhan, M. Lipson, Micrometre-scale silicon electro-optic modulator. *Nature*. **435**, 325–327 (2005).

88. R. Soref, The Past, Present, and Future of Silicon Photonics. *IEEE J. Sel. Top. Quantum Electron.* **12**, 1678–1687 (2006).

89. B. Jalali, S. Fathpour, Silicon Photonics. *J. Light. Technol.* **24**, 4600–4615 (2006).

90. A. Taflove, S. C. Hagness, *Computational electrodynamics : the finite-difference time-domain method* (Artech House, 2005; http://cds.cern.ch/record/1698084).

91. A. F. Oskooi *et al.*, Meep: A flexible free-software package for electromagnetic simulations by the FDTD method. *Comput. Phys. Commun.* **181**, 687–702 (2010).

92. B. J. Soller, D. K. Gifford, M. S. Wolfe, M. E. Froggatt, High resolution optical frequency domain reflectometry for characterization of components and assemblies. *Opt. Express*. **13**, 666 (2005).

93. B. Vuong *et al.*, in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, 2011; http://ieeexplore.ieee.org/document/6091515/), pp. 6134–6137.

94. M. Lipson, Guiding, Modulating, and Emitting Light on Silicon-Challenges and

Opportunities. *J. Light. Technol.* **23**, 4222 (2005).

95. C. T. Phare, Y.-H. Daniel Lee, J. Cardenas, M. Lipson, Graphene electro-optic modulator with 30 GHz bandwidth. *Nat. Photonics*. **9**, 511–514 (2015).

96. R. A. Soref, B. R. Bennett, M. A. Mentzer, S. Sriram, Eds. (International Society for Optics and Photonics, 1987; http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.937193), vol. 704, p. 32.

97. K. Suzuki, K. Kijima, Optical Band Gap of Barium Titanate Nanoparticles Prepared by RF-plasma Chemical Vapor Deposition. *Jpn. J. Appl. Phys.* **44**, 2081–2082 (2005).

98. S. H. Wemple, M. Didomenico, I. Camlibel, Dielectric and optical properties of melt-grown BaTiO3. *J. Phys. Chem. Solids*. **29**, 1797–1803 (1968).

99. M. D. Scott, J. O. Williams, R. C. Goodfellow, Excitonic luminescence of epitaxial CdS grown on InP(111)A. *Thin Solid Films*. **83**, L151–L154 (1981).

100. R. Soref, B. Bennett, Electrooptical effects in silicon. *IEEE J. Quantum Electron.* **23**, 123–129 (1987).

101. L. R. Dalton, P. A. Sullivan, D. H. Bale, Electric Field Poled Organic Electro-optic Materials: State of the Art and Future Prospects. *Chem. Rev.* **110**, 25–55 (2010).

102. T.-Z. Shen, S.-H. Hong, J.-K. Song, Electro-optical switching of graphene oxide liquid crystals with an extremely large Kerr coefficient. *Nat. Mater.* **13**, 394–399 (2014).

103. C. Uwe, S. Koichi, "Ultra-Sensitive Measurement of Sheath Electric Fields by Laser-Induced Fluorescence-Dip Spectroscopy," (available at http://www.jspf.or.jp/Journal/PDF_JSPF/jspf2007_03/jspf2007_03-215.pdf).

104. M. J. Adams, S. Ritchie, M. J. Robertson, Optimum overlap of electric and optical fields in semiconductor waveguide devices. *Appl. Phys. Lett.* **48**, 820–822 (1986).

105. T. Imai, M. Sasaura, N. Nakamura, K. Fujiura, Crystal Growth and Electro-optic Properties of KTa(1−x)NbxO3. *NTT Tech. Rev.* **5**, 1–8 (2007).

106. B. Van Zeghbroeck, *Principles of Semiconductor Devices* (Colorado University, Boulder, CO, 2004; http://ecee.colorado.edu/~bart/book/).

107. P. H. Paul, G. Kychakoff, Fiber-optic evanescent field absorption sensor. *Appl. Phys. Lett.* **51**, 12–14 (1987).

108. B. R. Bennett, R. A. Soref, J. A. Del Alamo, Carrier-induced change in refractive index of InP, GaAs and InGaAsP. *IEEE J. Quantum Electron.* **26**, 113–122 (1990).

109. W. Franks, I. Schenker, P. Schmutz, A. Hierlemann, Impedance Characterization and Modeling of Electrodes for Biomedical Applications. *IEEE Trans. Biomed. Eng.* **52**, 1295

(2005).

110. S. K. Arfin, thesis, Massachusetts Institute of Technology (2006).

111. R. W. de Boer, A. van Oosterom, Electrical properties of platinum electrodes: Impedance measurements and time-domain analysis. *Med. Biol. Eng. Comput.* **16**, 1–10 (1978).

112. J. Scholvin *et al.*, Close-Packed Silicon Microelectrodes for Scalable Spatially Oversampled Neural Recording. *IEEE Trans. Biomed. Eng.* **63**, 120–130 (2016).

113. S. F. Cogan, Neural Stimulation and Recording Electrodes. *Annu. Rev. Biomed. Eng.* **10**, 275–309 (2008).

114. A. Branwood, J. D. Hurd, R. H. Tredgold, Dielectric breakdown in barium titanate. *Br. J. Appl. Phys.* **13**, 528–528 (1962).

115. G. Lazzi, Thermal effects of bioimplants. *IEEE Eng. Med. Biol. Mag.* **24**, 75–81 (2005).

116. A. Cutolo, M. Iodice, P. Spirito, L. Zeni, Silicon electro-optic modulator based on a three terminal device integrated in a low-loss single-mode SOI waveguide. *J. Light. Technol.* **15**, 505–518 (1997).

117. M. Bugajski, W. Lewandowski, Concentration-dependent absorption and photoluminescence of *n* -type InP. *J. Appl. Phys.* **57**, 521–530 (1985).

118. G. Arlt, D. Hennings, G. de With, Dielectric properties of fine-grained barium titanate ceramics. *J. Appl. Phys.* **58**, 1619–1625 (1985).

119. J. Ihlefeld *et al.*, Copper Compatible Barium Titanate Thin Films for Embedded Passives. *J. Electroceramics*. **14**, 95–102 (2005).

120. S. Guillemet-Fritsch, T. Lebey, M. Boulos, B. Durand, Dielectric properties of CaCu3Ti4O12 based multiphased ceramics. *J. Eur. Ceram. Soc.* **26**, 1245–1257 (2006).

121. C. Pecharromán, F. Esteban-Betegón, J. F. Bartolomé, S. López-Esteban, J. S. Moya, New Percolative BaTiO3–Ni Composites with a High and Frequency-Independent Dielectric Constant ($\varepsilon r \approx 80000$). *Adv. Mater.* **13**, 1541 (2001).

122. T. B. Adams, D. C. Sinclair, A. R. West, Giant Barrier Layer Capacitance Effects in CaCu3Ti4O12 Ceramics. *Adv. Mater.* **14**, 1321–1323 (2002).

123. R. Landauer, D. R. Young, M. E. Drougard, Polarization Reversal in the Barium Titanate Hysteresis Loop. *J. Appl. Phys.* **27**, 752–758 (1956).

124. J. P. Kinney *et al.*, A direct-to-drive neural data acquisition system. *Front. Neural Circuits*. **9**, 46 (2015).

125. P. Blinder *et al.*, The cortical angiome: an interconnected vascular network with noncolumnar patterns of blood flow. *Nat. Neurosci.* **16**, 889–897 (2013).

126. F. Meng, Y. Ding, Sub-Micrometer-Thick All-Solid-State Supercapacitors with High Power and Energy Densities. *Adv. Mater.* **23**, 4098–4102 (2011).

127. P. J. Dean, G. Kaminsky, R. B. Zetterstrom, Intrinsic Optical Absorption of Gallium Phosphide between 2.33 and 3.12 eV. *J. Appl. Phys.* **38**, 3551–3556 (1967).

128. P. J. Dean, D. G. Thomas, Intrinsic Absorption-Edge Spectrum of Gallium Phosphide. *Phys. Rev.* **150**, 690–703 (1966).

129. R. Chein, J. Chuang, Experimental microchannel heat sink performance studies using nanofluids. *Int. J. Therm. Sci.* **46**, 57–66 (2007).

130. R. Chein, G. Huang, Analysis of microchannel heat sink performance using nanofluids. *Appl. Therm. Eng.* **25**, 3104–3114 (2005).

131. M. A. Scott, Z. D. Wissner-Gross, M. F. Yanik, Ultra-rapid laser protein micropatterning: screening for directed polarization of single neurons. *Lab Chip*. **12**, 2265 (2012).

132. M. A. Skylar-Scott, M.-C. Liu, Y. Wu, M. F. Yanik, G. von Freymann, W. V. Schoenfeld, R. C. Rumpf, Eds. (International Society for Optics and Photonics, 2017), vol. 10115.

133. M. A. Skylar-Scott, M.-C. Liu, Y. Wu, A. Dixit, M. F. Yanik, Guided Homing of Cells in Multi-Photon Microfabricated Bioscaffolds. *Adv. Healthc. Mater.* **5**, 1233–1243 (2016).

134. M. Deubel *et al.*, Direct laser writing of three-dimensional photonic-crystal templates for telecommunications. *Nat. Mater.* **3**, 444–447 (2004).

135. C. M. Soukoulis, M. Wegener, Past achievements and future challenges in the development of three-dimensional photonic metamaterials. *Nat. Photonics*. **5**, 523–530 (2011).

136. C. A. Ross, K. K. Berggren, J. Y. Cheng, Y. S. Jung, J.-B. Chang, Three-Dimensional Nanofabrication by Block Copolymer Self-Assembly. *Adv. Mater.* **26**, 4386–4396 (2014).

137. J.-B. Chang *et al.*, Design rules for self-assembled block copolymer patterns using tiled templates. *Nat. Commun.* **5**, 3305 (2014).

138. I. Wathuthanthri, Y. Liu, K. Du, W. Xu, C.-H. Choi, Simple Holographic Patterning for High-Aspect-Ratio Three-Dimensional Nanostructures with Large Coverage Area. *Adv. Funct. Mater.* **23**, 608–618 (2013).

139. S. Matsui, T. Kaito, J. Fujita, M. Komuro, K. Kanda, Three-dimensional nanostructure fabrication by focused-ion-beam chemical vapor deposition. *J. Vac. Sci. Technol. B*. **3181**, 3–7 (2013).

140. S. Kawata, H. B. Sun, T. Tanaka, K. Takada, Finer features for functional microdevices. *Nature*. **412**, 697–698 (2001).

141. L. R. Meza, S. Das, J. R. Greer, Strong, lightweight, and recoverable three-dimensional ceramic nanolattices. *Science*. **345**, 1322–6 (2014).

142. A. Vyatskikh *et al.*, Additive manufacturing of 3D nano-architected metals. *Nat. Commun.* **9**, 593 (2018).

143. Y. Y. Cao, N. Takeyasu, T. Tanaka, X. M. Duan, S. Kawata, 3D metallic nanostructure fabrication by surfactant-assisted multiphoton-induced reduction. *Small.* **5**, 1144–1148 (2009).

144. M. Hegde *et al.*, 3D Printing All-Aromatic Polyimides using Mask-Projection Stereolithography: Processing the Nonprocessable. *Adv. Mater.* **29**, 1701240 (2017).

145. X.-M. Zhao, Y. Xia, O. J. A. Schueller, D. Qin, G. M. Whitesides, Fabrication of microstructures using shrinkable polystyrene films. *Sensors Actuators A Phys.* **65**, 209–217 (1998).

146. J. Bauer, A. Schroer, R. Schwaiger, O. Kraft, Approaching theoretical strength in glassy carbon nanolattices. *Nat. Mater.* **15**, 438–443 (2016).

147. D. L. Holmes, N. C. Stellwagen, Estimation of polyacrylamide gel pore size from Ferguson plots of linear DNA fragments. II. Comparison of gels with different crosslinker concentrations, added agarose and added linear polyacrylamide. *Electrophoresis*. **12**, 612–619 (1991).

148. F. Ilmain, T. Tanaka, E. Kokufuta, Volume transition in a gel driven by hydrogen bonding. *Nature*. **349**, 400–401 (1991).

149. Y. Hirokawa, T. Tanaka, in *AIP Conference Proceedings* (American Institute of Physics, 1984; http://aip.scitation.org/doi/abs/10.1063/1.34300), vol. 107, pp. 203–208.

150. A. Suzuki, T. Tanaka, Phase transition in polymer gels induced by visible light. *Nature*. **346**, 345–347 (1990).

151. C. A. DeForest, K. S. Anseth, Cytocompatible click-based hydrogels with dynamically tunable properties through orthogonal photoconjugation and photocleavage reactions. *Nat. Chem.* **3**, 925–931 (2011).

152. C. A. DeForest, K. S. Anseth, Photoreversible Patterning of Biomolecules within Click-Based Hydrogels. *Angew. Chemie*. **124**, 1852–1855 (2012).

153. A. Miura *et al.*, Fabrication of gold clusters photoreduced in gold-dendrimer complex nanoparticles. *Opt. Mater. Express*. **7** (2017), doi:10.1364/OME.7.002224.

154. K. Esumi, A. Suzuki, N. Aihara, K. Usui, K. Torigoe, Preparation of Gold Colloids with UV Irradiation Using Dendrimers as Stabilizer. *Langmuir*. **14**, 3157–3159 (1998).

155. P. W. K. Rothemund, Folding DNA to create nanoscale shapes and patterns. *Nature*. **440**, 297–302 (2006).

156. P. L. Stahl *et al.*, Visualization and analysis of gene expression in tissue sections by spatial

transcriptomics. *Science*. **353**, 78–82 (2016).

157. J. H. Lee *et al.*, Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–58 (2015).

158. K. L. Gunderson *et al.*, Decoding Randomly Ordered DNA Arrays. *Genome Res.* **14**, 870–877 (2004).

159. K. B. Halpern *et al.*, Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*. **542**, 352–356 (2017).

160. M. J. West, New stereological methods for counting neurons. *Neurobiol. Aging*. **14**, 275–285 (1993).

161. L. O. Trussell, D. Oertel, (Springer, Cham, 2018; http://link.springer.com/10.1007/978-3-319-71798-2_4), pp. 73–99.

162. M. T. Roberts, L. O. Trussell, Molecular Layer Inhibitory Interneurons Provide Feedforward and Lateral Inhibition in the Dorsal Cochlear Nucleus. *J. Neurophysiol.* **104**, 2462–2473 (2010).

163. C. Gravel, R. Hawkes, Parasagittal organization of the rat cerebellar cortex: Direct comparison of purkinje cell compartments and the organization of the spinocerebellar projection. *J. Comp. Neurol.* **291**, 79–102 (1990).

164. J. Xiao *et al.*, Systematic regional variations in Purkinje cell spiking patterns. *PLoS One*. **9**, e105633 (2014).

165. N. H. Barmack, Z. Qian, J. Yoshimura, Regional and cellular distribution of protein kinase C in rat cerebellar Purkinje cells. *J. Comp. Neurol.* **427**, 235–54 (2000).

166. N. L. Cerminara, E. J. Lang, R. V Sillitoe, R. Apps, Redefining the cerebellar cortex as an assembly of non-uniform Purkinje cell microcircuits. *Nat. Rev. Neurosci.* **16**, 79 (2015).

167. A. Demilly, S. L. Reeber, S. A. Gebre, R. V. Sillitoe, Neurofilament Heavy Chain Expression Reveals a Unique Parasagittal Stripe Topography in the Mouse Cerebellum. *The Cerebellum*. **10**, 409–421 (2011).

168. G. Brochu, L. Maler, R. Hawkes, Zebrin II: A polypeptide antigen expressed selectively by purkinje cells reveals compartments in rat and fish cerebellum. *J. Comp. Neurol.* **291**, 538–552 (1990).

169. K. Maekawa, J. I. Simpson, Climbing Fiber Responses Evoked in Vestibulocerebellum of Rabbit From Visual System. *J. Neurophysiol.*, 649–666 (1972).

170. D. R. W. Wylie, M. R. Brown, I. R. Winship, N. A. Crowder, K. G. Todd, Zonal organization of the vestibulocerebellum in pigeons (Columba livia): III. Projections of the translation zones of the ventral uvula and nodulus. *J. Comp. Neurol.* **465**, 179–194 (2003).

171. C. Chung *et al.*, Heat Shock Protein Beta-1 Modifies Anterior to Posterior Purkinje Cell Vulnerability in a Mouse Model of Niemann-Pick Type C Disease. *PLoS Genet.* **12**, 1–22 (2016).

172. C. J. Stoodley, J. D. Schmahmann, Evidence for topographic organization in the cerebellum of motor control versus cognitive and affective processing. *Cortex*. **46**, 831–44 (2010).

173. R. A. Carter *et al.*, A Single-Cell Transcriptional Atlas of the Developing Murine Cerebellum. *Curr. Biol.* **28**, 2910–2920. (2018).

174. P. D. Storer, K. J. Jones, Ribosomal RNA transcriptional activation and processing in hamster rubrospinal motoneurons: Effects of axotomy and testosterone treatment. *J. Comp. Neurol.* **458**, 326–333 (2003).

175. K. L. Adams, V. Gallo, The diversity and disparity of the glial scar. *Nat. Neurosci.* (2017), doi:10.1038/s41593-017-0033-9.

176. A. M. Kenney, J. D. Kocsis, Peripheral axotomy induces long-term c-Jun amino-terminal kinase-1 activation and activator protein-1 binding activity by c-Jun and junD in adult rat dorsal root ganglia In vivo. *J. Neurosci.* **18**, 1318–28 (1998).

177. G. A. Robinson, Immediate early gene expression in axotomized and regenerating retinal ganglion cells of the adult rat. *Mol. Brain Res.* **24**, 43–54 (1994).

178. J. Honkaniemi, S. M. Sagar, I. Pyykönen, K. J. Hicks, F. R. Sharp, Focal brain injury induces multiple immediate early genes encoding zinc finger transcription factors. *Mol. Brain Res.* **28**, 157–163 (1995).

179. R. D. Smith, X. Cheng, J. E. Bruce, S. A. Hofstadler, G. A. Anderson, Trapping, detection and reaction of very large single molecular ions by mass spectrometry. *Nature*. **369** (1994).

180. J. J. Havranek, B. Borgo, Molecules and methods for iterative polypeptide analysis and processing (2013), (available at https://patents.google.com/patent/US20140273004A1/en).

181. J. Swaminathan, A. A. Boulgakov, E. M. Marcotte, A Theoretical Justification for Single Molecule Peptide Sequencing. *PLOS Comput. Biol.* **11**, e1004080 (2015).

182. J. van Ginkel *et al.*, Single-molecule peptide fingerprinting. *Proc. Natl. Acad. Sci.* **115**, 3338–3343 (2018).

183. C. Joo, M. Fareh, V. Narry Kim, Bringing single-molecule spectroscopy to macromolecular protein complexes. *Trends Biochem. Sci.* **38**, 30–37 (2013).

184. J. Swaminathan *et al.*, Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* **36**, 1076–1082 (2018).

185. J. Shendure, E. L. Aiden, The expanding scope of DNA sequencing. *Nat. Biotechnol.* **30**, 1084–1094 (2012).

186. J. Shendure, R. D. Mitra, C. Varma, G. M. Church, Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).

187. D. R. Bentley *et al.*, Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. **456**, 53–59 (2008).

188. S. Brenner *et al.*, Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).

189. R. D. Mitra, J. Shendure, J. Olejnik, Edyta-Krzymanska-Olejnik, G. M. Church, Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem.* **320**, 55–65 (2003).

190. M. Levene, J. Korlach, S. Turner, Zero-mode waveguides for single-molecule analysis at high concentrations. *Science.* **299**, 682–685 (2003).

191. I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake, Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci.* **100**, 3960–3964 (2003).

192. C. W. Fuller *et al.*, Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. *Proc. Natl. Acad. Sci.* **113**, 5233–5238 (2016).

193. D. R. Garalde *et al.*, Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods*. **15**, 201–206 (2018).

194. J. Nivala, D. B. Marks, M. Akeson, Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* **31**, 247–250 (2013).

195. M. Kolmogorov, E. Kennedy, Z. Dong, G. Timp, P. A. Pevzner, Single-molecule protein identification by sub-nanopore sensors. *PLOS Comput. Biol.* **13**, e1005356 (2017).

196. G. Sampath, A digital approach to protein identification and quantity estimation using tandem nanopores, peptidases, and database search. *bioRxiv*, 24158 (2015).

197. Y. Yao, M. Docter, J. van Ginkel, D. de Ridder, C. Joo, Single-molecule protein sequencing through fingerprinting: computational assessment. *Phys. Biol.* **12**, 55003 (2015).

198. B. Borgo, thesis, WUSTL (2014).

199. B. Borgo, J. J. Havranek, Computer-aided design of a catalyst for Edman degradation utilizing substrate-assisted catalysis. *Protein Sci.* **24**, 571–579 (2015).

200. L. A. Tessler *et al.*, Nanogel surface coatings for improved single-molecule imaging substrates. *J. R. Soc. Interface*. **8**, 1400–8 (2011).

201. B. Borgo, J. J. Havranek, Motif-directed redesign of enzyme specificity. *Protein Sci.* **23**, 312–320 (2014).

202. R. D. Mitra, L. A. Tessler, Single molecule protein screening (2010).

203. A. Sharonov, R. M. Hochstrasser, Wide-field subdiffraction imaging by accumulating binding of diffusion probes. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 18911–18916 (2006).

204. R. Jungmann *et al.*, Single-Molecule Kinetics and Super-Resolution Microscopy by Fluorescence Imaging of Transient Binding on DNA Origami. *Nano Lett.* **10**, 4756–4761 (2010).

205. L. A. Tessler, R. D. Mitra, Sensitive single-molecule protein quantification and protein complex detection in a microarray format. *Proteomics*. **11**, 4731–4735 (2011).

206. S. D. Chandradoss *et al.*, Surface Passivation for Single-molecule Protein Studies. *J. Vis. Exp.* (2014), doi:10.3791/50549.

207. P. R. Selvin, T. Ha, *Single-molecule techniques : a laboratory manual* (Cold Spring Harbor Laboratory Press, 2008; https://www.cshlpress.com/default.tpl?cart=15544866181876953 41&fromlink=T&linkaction=full&linksortby=oop_title&--eqSKUdatarq=642).

208. J. Groll, M. Moeller, in *Methods in enzymology* (2010; http://www.ncbi.nlm.nih.gov/pubmed/20580956), vol. 472, pp. 1–18.

209. I. J. Finkelstein, E. C. Greene, in *Methods in molecular biology (Clifton, N.J.)* (2011; http://www.ncbi.nlm.nih.gov/pubmed/21660710), vol. 745, pp. 447–461.

210. H. Pan, Y. Xia, M. Qin, Y. Cao, W. Wang, A simple procedure to improve the surface passivation for single molecule fluorescence studies. *Phys. Biol.* **12**, 45006 (2015).

211. P. EDMAN, A method for the determination of amino acid sequence in peptides. *Arch. Biochem.* **22**, 475 (1949).

212. R. A. Laursen, Solid-phase Edman degradation. An automatic peptide sequencer. *Eur. J. Biochem.* **20**, 89–102 (1971).

213. J. Vogelsang *et al.*, A Reducing and Oxidizing System Minimizes Photobleaching and Blinking of Fluorescent Dyes. *Angew. Chemie Int. Ed.* **47**, 5465–5469 (2008).

214. F. Pennacchietti, T. J. Gould, S. T. Hess, The Role of Probe Photophysics in Localization-Based Superresolution Microscopy. *Biophys. J.* **113**, 2037–2054 (2017).

215. G. S. Schlau-Cohen, Q. Wang, J. Southall, R. J. Cogdell, W. E. Moerner, Single-molecule spectroscopy reveals photosynthetic LH2 complexes switch between emissive states. *Proc. Natl. Acad. Sci.* **110**, 10899–10903 (2013).

216. A. M. van Oijen, Single-molecule approaches to characterizing kinetics of biomolecular interactions. *Curr. Opin. Biotechnol.* **22**, 75–80 (2011).

217. J. Foote, H. N. Eisen, Kinetic and affinity limits on antibodies produced during immune responses. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 1254–6 (1995).

218. N. Nemoto, E. Miyamoto-Sato, H. Yanagawa, Fluorescence labeling of the C-terminus of proteins with a puromycin analogue in cell-free translation systems. *FEBS Lett.* **462**, 43–6 (1999).

219. G. Xu, S. B. Y. Shin, S. R. Jaffrey, Chemoenzymatic Labeling of Protein C-Termini for Positive Selection of C-Terminal Peptides. *ACS Chem. Biol.* **6**, 1015–1020 (2011).

220. Y. Lu, B. Lee, R. W. King, D. Finley, M. W. Kirschner, Substrate degradation by the proteasome: a single-molecule kinetic analysis. *Science*. **348**, 1250834 (2015).

221. G. Dempsey, J. Vaughan, K. Chen, M. Bates, X. Zhuang, Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat. Methods*. **8**, 1027–1036 (2011).

222. Y. Cui, Q. Wei, H. Park, C. M. Lieber, Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science*. **293**, 1289–1292 (2001).

223. F. Patolsky *et al.*, Electrical detection of single viruses. *Proc. Natl. Acad. Sci.* **101**, 14017–14022 (2004).

224. E. Stern *et al.*, Label-free immunodetection with CMOS-compatible semiconducting nanowires. *Nature*. **445**, 519–522 (2007).

225. A. Kim *et al.*, Ultrasensitive, label-free, and real-time immunodetection using silicon field-effect transistors. *Appl. Phys. Lett.* **91**, 103901 (2007).

226. J. M. Rothberg *et al.*, An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. **475**, 348–352 (2011).

227. D. L. Bellin, S. B. Warren, J. K. Rosenstein, K. L. Shepard, in *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings* (IEEE, 2014; http://ieeexplore.ieee.org/document/6981766/), pp. 476–479.

228. S. Sorgenfrei *et al.*, Label-free single-molecule detection of DNA-hybridization kinetics with a carbon nanotube field-effect transistor. *Nat. Nanotechnol.* **6**, 126–132 (2011).

229. N. Lu *et al.*, Label-Free and Rapid Electrical Detection of hTSH with CMOS-Compatible Silicon Nanowire Transistor Arrays. *ACS Appl. Mater. Interfaces*. **6**, 20378–20384 (2014).

230. N. Guo *et al.*, CMOS Time-Resolved, Contact, and Multispectral Fluorescence Imaging for DNA Molecular Diagnostics. *Sensors*. **14**, 20602–20619 (2014).

231. K. P. Kording, Of Toasters and Molecular Ticker Tapes. *PLoS Comput. Biol.* **7** (2011), doi:10.1371/journal.pcbi.1002291.

232. B. M. Zamft *et al.*, Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PLoS One*. **7** (2012), doi:10.1371/journal.pone.0043876.

233. A. M. de Paz *et al.*, High-resolution mapping of DNA polymerase fidelity using nucleotide

imbalances and next-generation sequencing. *Nucleic Acids Res.* **46**, e78–e78 (2018).

234. S. D. Perli *et al.*, Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science*. **353**, 339–342 (2016).

235. F. Farzadfard, N. Gharaei, Y. Higashikuni, G. Jung, Single-Nucleotide-Resolution Computing and Memory in Living Cells. *bioRxiv* (2018).

236. F. Farzadfard, T. K. Lu, Genomically encoded analog memory with precise in vivo dna writing in living cell populations. *Science*. **346** (2014), doi:10.1126/science.1256272.

237. R. Kalhor *et al.*, Rapidly evolving homing CRISPR barcodes. *Nat. Methods*. **14**, 195–200 (2017).

238. R. U. Sheth, S. S. Yim, F. L. Wu, H. H. Wang, Multiplex recording of cellular events over time on CRISPR biological tape. *Science*. **358** (2017), doi:10.1126/science.aao0958.

239. W. Tang, D. R. Liu, Rewritable multi-event analog recording in bacterial and mammalian cells. *Science*. **360** (2018).

240. A. Shur, R. M. Murray, Proof of concept continuous event logging in living cells. *bioRxiv* (2018).

241. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2016).

242. S. L. Shipman *et al.*, Molecular recordings by directed CRISPR spacer acquisition. *Science*. **353** (2016), doi:10.1126/science.aaf1175.

243. F. Schmidt, M. Y. Cherepkova, R. J. Platt, Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature*. **562**, 380–385 (2018).

244. C. P. Lapointe, D. Wilinski, H. A. J. Saunders, M. Wickens, Protein-RNA networks revealed through covalent RNA marks. *Nat. Methods*. **12**, 1163–1170 (2015).

245. C. P. Lapointe, M. Wickens, The Nucleic Acid-binding Domain and Translational Repression Activity of a Xenopus Terminal Uridylyl. *J. Biol. Chem.* **288**, 20723–20733 (2013).

246. J. E. Kwak, L. Wang, S. Ballantyne, J. Kimble, M. Wickens, Mammalian GLD-2 homologs are poly (A) polymerases. *Proc. Natl. Acad. Sci.* **101** (2004).

247. L. Wang, C. R. Eckmann, L. C. Kadyk, M. Wickens, J. Kimble, A regulatory cytoplasmic poly(A) polymerase in Caenorhabditis elegans. *Nature*. **419**, 312–316 (2002).

248. I. Aphasizheva *et al.*, Novel TUTase associates with an editosome-like complex in mitochondria of Trypanosoma brucei Novel TUTase associates with an editosome-like complex in mitochondria of Trypanosoma brucei. *RNA*, 1322–1337 (2009).

249. P. Munoz-Tello, L. Rajappa, S. Coquille, S. Thore, Polyuridylation in eukaryotes: A 3′-end modification regulating RNA life. *Biomed Res. Int.* (2015), , doi:10.1155/2015/968127.

250. M. Morgan *et al.*, mRNA 3' uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nat. Publ. Gr.* **548**, 347–351 (2017).

251. I. Aphasizheva, R. Aphasizhev, L. Simpson, RNA-editing terminal uridylyl transferase 1. Identification of functional domains by mutational analysis. *J. Biol. Chem.* **279**, 24123–24130 (2004).

252. J. Lim *et al.*, Uridylation by TUT4 and TUT7 Marks mRNA for Degradation. *Cell.* **159**, 1365–1376 (2014).

253. D. B. T. Cox, J. S. Gootenberg, O. O. Abudayyeh, B. Franklin, M. J. Kellner, RNA editing with CRISPR-Cas13. *Science.* **180**, 1–15 (2017).

254. T.-W. Chen *et al.*, Ultra-sensitive fluorescent proteins for imaging neuronal activity. *Nature*. **499**, 295 (2013).

255. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science*. **297**, 1183–6 (2002).

256. M. Muhar, S. L. Ameres, J. Zuber, SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science.* **2793**, 1–10 (2018).

257. V. A. Herzog *et al.*, Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods*. **14** (2017), doi:10.1038/nmeth.4435.

258. J. A. Schofield, E. E. Duffy, L. Kiefer, M. C. Sullivan, M. D. Simon, TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods* (2018), doi:10.1038/nmeth.4582.

259. G. La Manno *et al.*, RNA velocity of single cells. *Nature*, 206052 (2018).

260. P. E. Hardin, J. C. Hall, M. Rosbash, Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels. *Nature*. **343**, 536–540 (1990).

261. M. W. Young, S. A. Kay, Time zones: a comparative genetics of circadian clocks. *Nat. Rev. Genet.* **2**, 702–715 (2001).

262. K. D. Piatkevich *et al.*, A robotic multidimensional directed evolution approach applied to fluorescent voltage reporters. *Nat. Chem. Biol.* **14** (2018), doi:10.1038/s41589-018-0004-9.

263. M. M. Matthews *et al.*, Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. *Nat. Struct. Mol. Biol.* **23**, 426–433 (2016).

264. A. Kuttan, B. L. Bass, Mechanistic insights into editing-site specificity of ADARs. *Proc. Natl. Acad. Sci.* **109**, 3295–3304 (2012).

265. T. Eifler, S. Pokharel, P. A. Beal, RNA-seq analysis identifies a novel set of editing substrates for human ADAR2 present in saccharomyces cerevisiae. *Biochemistry.* **52**, 7857–7869 (2013).

266. E. Bertrand *et al.*, Localization of ASH1 mRNA Particles in Living Yeast. *Mol. Cell.* **2**, 437–445 (1998).

267. R. P. Perry, D. E. Kelley, Inhibition of RNA synthesis by actinomycin D: Characteristic dose-response of different RNA species. *J. Cell. Physiol.* **76**, 127–139 (1970).

268. X. Wang, X. Chen, Y. Yang, Spatiotemporal control of gene expression by a light-switchable transgene system. *Nat. Methods.* **9**, 266–271 (2012).

269. Z. Ma, Z. Du, X. Chen, X. Wang, Y. Yang, Fine tuning the LightOn light-switchable transgene expression system. *Biochem. Biophys. Res. Commun.* **440**, 419–423 (2013).

270. S. Tay *et al.*, Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature.* **466**, 267–71 (2010).

271. N. Nandagopal *et al.*, Dynamic Ligand Discrimination in the Notch Signaling Pathway. *Cell.* **172**, 869–880.e19 (2018).

272. A. H. Marblestone *et al.*, Physical principles for scalable neural recording. *Front. Comput. Neurosci.* **7**, 137 (2013).

273. A. E. West *et al.*, Calcium regulation of neuronal gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 11024–31 (2001).

274. A. E. West, E. C. Griffith, M. E. Greenberg, Regulation of transcription factors by neuronal activity. *Nat. Rev. Neurosci.* **3**, 921–931 (2002).

275. V. M. Rivera *et al.*, A humanized system for pharmacologic control of gene expression. *Nat. Med.* **2**, 1028–32 (1996).

276. D. Erhart *et al.*, Chemical Development of Intracellular Protein Heterodimerizers. *Chem. Biol.* **20**, 549–557 (2013).

277. A. H. Marblestone *et al.*, Rosetta Brains: A Strategy for Molecularly-Annotated Connectomics (2014) (available at http://arxiv.org/abs/1404.5103).

278. E. G. Chapman *et al.*, The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production. *Science.* **344**, 307–310 (2014).

279. B. M. Akiyama *et al.*, Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science.* **3963** (2016), doi:10.1126/science.aah3963.

280. L. Wei *et al.*, Super-multiplex vibrational imaging. *Nature.* **544**, 465–470 (2017).

281. S. Viswanathan *et al.*, High-performance probes for light and electron microscopy. *Nat. Methods*. **12**, 568–576 (2015).

282. W. Pei *et al.*, Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*. **548**, 456–460 (2017).

283. K. Y. Chan *et al.*, Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nat. Neurosci.* **20**, 1172–1179 (2017).

284. B. E. Deverman *et al.*, Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nat. Biotechnol.* **34**, 204–209 (2016).

285. R. C. Challis *et al.*, Systemic AAV vectors for widespread and targeted gene delivery in rodents. *Nat. Protoc.* **14**, 379–414 (2019).

286. R. Kalhor *et al.*, Developmental barcoding of whole mouse via homing CRISPR. *Science.* **9804**, 1–19 (2018).

287. Y. Goltsev *et al.*, Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*. **174**, 968–981 (2018).

288. J.-B. Chang *et al.*, Iterative expansion microscopy. *Nat. Methods*. **14**, 593–599 (2017).

289. A. M. Zador *et al.*, Sequencing the Connectome. *PLoS Biol.* **10** (2012), doi:10.1371/journal.pbio.1001411.

290. J. H. J. H. Lee *et al.*, Highly Multiplexed Subcellular RNA Sequencing in Situ. *Science.* **343**, 1360–1363 (2014).

291. J. Livet *et al.*, Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*. **450**, 56–62 (2007).

292. D. Cai, K. B. Cohen, T. Luo, J. W. Lichtman, J. R. Sanes, Improved tools for the Brainbow toolbox. *Nat. Methods*. **10**, 540–547 (2013).

293. A. Nern, B. D. Pfeiffer, K. Svoboda, G. M. Rubin, Multiple new site-specific recombinases for use in manipulating animal genomes. *Proc. Natl. Acad. Sci.* **108**, 14198–14203 (2011).

294. S. Turan, J. Kuehle, A. Schambach, C. Baum, J. Bode, Multiplexing RMCE: Versatile Extensions of the Flp-Recombinase-Mediated Cassette-Exchange Technology. *J. Mol. Biol.* **402**, 52–69 (2010).

295. S. D. Colloms *et al.*, Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination. *Nucleic Acids Res.* **42**, e23 (2014).

296. K. D. Micheva, S. J. Smith, Array Tomography: A New Tool for Imaging the Molecular Architecture and Ultrastructure of Neural Circuits. *Neuron*. **55**, 25–36 (2007).

297. 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).

298. A. Nern, B. D. Pfeiffer, G. M. Rubin, Optimized tools for multicolor stochastic labeling reveal diverse stereotyped cell arrangements in the fly visual system. *Proc. Natl. Acad. Sci.* **112**, 2967–2976 (2015).

299. N. Kasthuri *et al.*, Saturated Reconstruction of a Volume of Neocortex. *Cell.* **162**, 648–661 (2015).

300. S. Ramon y Cajal, *Advice for a Young Investigator* (1897).

301. L. Wu, D. Wang, J. A. Evans, Large teams develop and small teams disrupt science and technology. *Nature* (2019).

302. "The Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative 2.0" (2019), (available at https://www.braininitiative.nih.gov/strategic-planning/acd-working-group/brain-research-through-advancing-innovative-neurotechnologies).

303. J. L. Morgan, J. W. Lichtman, Why not connectomics? *Nat. Methods.* **10**, 494–500 (2013).

304. J. W. Lichtman, H. Pfister, N. Shavit, The big data challenges of connectomics. *Nat. Neurosci.* **17**, 1448–1454 (2014).

305. The Comptroller General of the United States, "Nanomanufacturing: Emergence and Implications for U.S. Competitiveness, the Environment, and Human Health" (2014), (available at https://www.gao.gov/assets/670/660591.pdf).

306. M. G. Lawrence *et al.*, Evaluating climate geoengineering proposals in the context of the Paris Agreement temperature goals. *Nat. Commun.* **9** (2018), doi:10.1038/s41467-018-05938-3.

307. E. Singer, Mapping the Brain to Build Better Machines. *Quanta Mag.* (2016).

308. D. Castelvecchi, Hunt for gravitational waves to resume after massive upgrade. *Nature.* **525**, 301–302 (2015).

309. A. Knapp, How Much Does It Cost To Find A Higgs Boson? *Forbes* (2012), (available at https://www.forbes.com/sites/alexknapp/2012/07/05/how-much-does-it-cost-to-find-a-higgs-boson/#7364f87b3948).

310. A. R. Jones, C. C. Overly, S. M. Sunkin, The Allen Brain Atlas: 5 years and beyond. *Nat. Rev. Neurosci.* **10**, 821–828 (2009).

311. E. A. Moore, A $55 million atlas of the human brain. *Cnet* (2011).

312. T.-W. Chen *et al.*, Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature.* **499**, 295–300 (2013).

313. G. M. Rubin, E. K. O. Shea, Looking back and looking forward at Janelia (2006).

314. M. Watzinger, T. A. Fackler, M. Nagler, M. Schnitzer, How Antitrust Enforcement Can Spur Innovation: Bell Labs and the 1956 Consent Decree (2017) (available at https://ssrn.com/abstract=2904315).

315. D. W. Braben, *Pioneering research: A risk worth taking* (John Wiley & Sons, Ltd, 2004).

316. I. D. Peikon *et al.*, Using high-throughput barcode sequencing to efficiently map connectomes. *Nucleic Acids Res.* **45** (2017), doi:10.1093/nar/gkx292.

317. J. C. Venter *et al.*, The Sequence of the Human Genome. *Science*. **291** (2001) (available at www.sciencemag.org/cgi/content/full/291/).

318. Pennisi E., Funders reassure genome sequencers. *Science.* **280**, 1185 (1998).

319. J. L. Weber, E. W. Myers, Human whole-genome shotgun sequencing. *Genome Res.* **7**, 401–9 (1997).

320. P. Green, Against a whole-genome shotgun. *Genome Res.* **7**, 410–7 (1997).

321. J. Segal, How Social Entrepreneurship is Changing Chicago (and the World). *Technori* (2012).

322. M. Berger, California Social Purpose Corporation: An Overview. *Nonprofit Law Blog* (2015).

323. B. Nauntofte, Inside View: Novo Nordisk Foundation. *NatureJobs* (2015), doi:10.1038/nj0470.

324. D. Coldewey, OpenAI shifts from non-profit to "capped-profit" to attract capital. *Techcrunch* (2019).

325. J. Buchanan, With $294M, Kamen Hopes to Bring Regenerative Medicine "Up to Scale." *Xconomy* (2017).

326. P. Cohan, MIT's $1.9 Trillion Money Machine. *Inc.* (2017).

327. E. B. Roberts, F. Murray, J. D. Kim, "Entrepreneurship and Innovation at MIT" (2015).

328. M. Farrell, Universities That Turn Research Into Revenue. *Forbes* (2008).

329. W. D. Valdivia, "University Start-Ups: Critical for Improving Technology Transfer" (2013).

330. D. G. Lowe, in *Proceedings of the Seventh IEEE International Conference on Computer Vision* (IEEE, 1999; http://ieeexplore.ieee.org/document/790410/), pp. 1150–1157 vol.2.

331. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. **14**, 417–419 (2017).

332. P. Thevenaz, U. E. Ruttimann, M. Unser, A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Process.* **7**, 27–41 (1998).

333. Q. Kong, M. P. Stockinger, Y. Chang, H. Tashiro, C. L. G. Lin, The presence of rRNA sequences in polyadenylated RNA and its potential functions. *Biotechnol. J.* **3**, 1041–1046 (2008).

334. D. Müller *et al.*, Dlk1 Promotes a Fast Motor Neuron Biophysical Signature Required for Peak Force Execution. *Science*. **343**, 1264–1266 (2014).

335. H. Zhou *et al.*, Cerebellar modules operate at different frequencies. *Elife*. **3**, 2536 (2014).

336. M. D. Womack, Dendritic Control of Spontaneous Bursting in Cerebellar Purkinje Cells. *J. Neurosci.* **24**, 3511–3521 (2004).

337. C. H. Kim *et al.*, Lobule-specific membrane excitability of cerebellar Purkinje cells. *J. Physiol.* **590**, 273–288 (2012).

338. I. D. Peikon, D. I. Gizatullina, A. M. Zador, In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic Acids Res.* **42**, e127–e127 (2014).

339. C. Guo, W. Yang, C. G. Lobe, A Cre recombinase transgene with mosaic, widespread tamoxifen-inducible action. *Genesis*. **32**, 8–18 (2002).

340. A. Jenett *et al.*, A GAL4-Driver Line Resource for Drosophila Neurobiology. *Cell Rep.* **2**, 991–1001 (2012).

341. H. Kakidani, M. Ptashne, GAL4 activates gene expression in mammalian cells. *Cell*. **52**, 161–7 (1988).

342. C. Kellendonk *et al.*, Regulation of Cre recombinase activity by the synthetic steroid RU 486. *Nucleic Acids Res.* **24**, 1404–11 (1996).

343. M. Lewandoski, Conditional control of gene expression in the mouse. *Nat. Rev. Genet.* **2**, 743–755 (2001).

344. N. Sharma, B. Moldt, T. Dalsgaard, T. G. Jensen, J. G. Mikkelsen, Regulated gene insertion by steroid-induced PhiC31 integrase. *Nucleic Acids Res.* **36**, e67 (2008).

345. H. Taniguchi *et al.*, A Resource of Cre Driver Lines for Genetic Targeting of GABAergic Neurons in Cerebral Cortex. *Neuron*. **71**, 995–1013 (2011).

346. A. H. Marblestone *et al.*, Conneconomics: The Economics of Dense, Large-Scale, High-Resolution Neural Connectomics. *bioRxiv*, 1214 (2014).

347. D. Y. Zhang, S. X. Chen, P. Yin, Optimizing the specificity of nucleic acid hybridization. *Nat. Chem.* **4**, 208–214 (2012).

348. H. M. T. Choi *et al.*, Programmable in situ amplification for multiplexed imaging of mRNA expression. *Nat. Biotechnol.* **28**, 1208–12 (2010).

349. K. Xu, G. Zhong, X. Zhuang, Actin, Spectrin, and Associated Proteins Form a Periodic

Cytoskeletal Structure in Axons. *Science.* **339**, 452–456 (2013).

350. R. M. Levenson, A. D. Borowsky, M. Angelo, Immunohistochemistry and mass spectrometry for highly multiplexed cellular molecular imaging. *Lab. Investig.* **95**, 397–405 (2015).

351. D. B. Burckel *et al.*, Micrometer-Scale Cubic Unit Cell 3D Metamaterial Layers. *Adv. Mater.* **22**, 5053–5057 (2010).