

Encoding of neural circuit information into DNA



Ian Peikon

Watson School of Biological Sciences

Cold Spring Harbor Laboratory

A thesis submitted for the degree of

Philosophiæ Doctor (PhD)

2014, September 26

-
1. Thesis Advisor: Anthony Zador

 2. Academic Advisor: Gurinder Atwal

 3. Thesis committee member: Bruce Stillman

 4. External thesis committee member: Cornelia Bargmann

 4. Thesis committee chair: Joshua Dubnau

Day of the defense:

Signature from head of PhD committee:

Abstract

Despite a few exceptional examples, our understanding of neural circuit computation remains limited. A full understanding of a given computational circuit requires knowledge of the function of the individual components and a wiring diagram. The brain is composed of extremely complex cellular networks, consisting of numerous cell-types - some of which likely remain unknown - connected by orders of magnitude more synapses. Our understanding is largely shrouded by our lack of tools for probing neuronal circuits with high-throughput. Therefore, we have developed methods of encoding information about neuronal circuits into DNA, a biologically relevant high-density storage medium. Advances in high-throughput DNA sequencing may enable the dissection of neural circuits with unprecedented depth.

To Tim Hanson - who taught me how to think.
And to Daniel Brown and Jack Peikon for their support.

Acknowledgements

During my time at CSHL I think I managed to ask pretty much every one of my colleagues for help at some point or another. I want to thank everyone at CSHL for continuing the great collaborative environment that exists here.

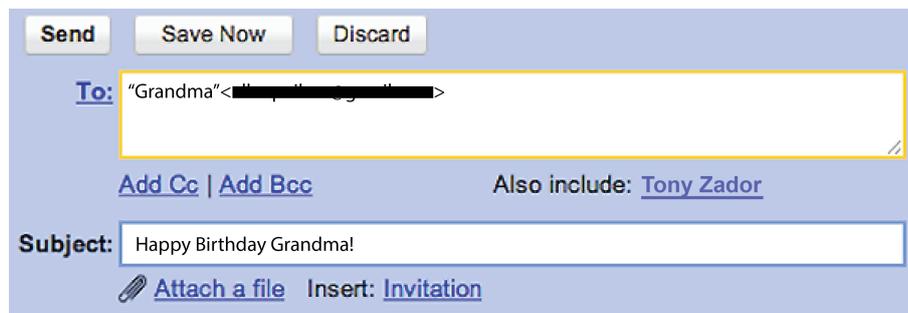
Specifically, I'd like to thank Jen Troge for all of her help with high-throughput sequencing, and Stephanie Muller for her help with PacBio libraries.

I also had the benefit of interacting with several fantastic external collaborators who really made the whole process a lot more fun. I'd specifically like to thank Steve Bates, Fei Chen, and Ed Boyden for stimulating conversations about applying molecular/synthetic biology to problems in neurobiology. Huge thanks to Ivan Correa at NEB for all of his help with chemical synthesis and all things CLIP and SNAP. Scott Silverman was extremely helpful with all of the RNA ligation methods. For several years I threw tons of questions his way and he never failed to give a thorough and thoughtful answer.

I'd also like to thank my thesis committee – Josh Dubnau, Bruce Stillman, and Mickey Atwal – for their time, effort, and countless helpful suggestions.

The Marks building was a fantastic place to work and I thank all of the Marxists for their friendship, advice, and experimental help. A special thanks to the molecular biologists who taught me when I didn't even know how to make agarose: Hassana Oyibo, Huiqing Zhan, Gang Cao, and Peter Znamenskiy. Thanks to Justus Keschull, my partner in crime through the hardest part of this project. And my biggest thanks to Diana Gizatullina who really went above and beyond her job description. She played a major role in everything in this thesis.

Finally, I want to thank Tony Zador for being the best advisor I could have imagined, and for being a great friend. I'm grateful that Tony had enough faith to put me on this project despite my completely orthogonal skill-set when I joined the lab. Tony has been an amazing leader in the brainstorming, troubleshooting, and design that this thesis entailed. He was always available for open and meaningful discussion of any and all ideas (including those outside the scope of this project), with zero judgement. We did this so much that Gmail still suggests I add Tony to every email, regardless of who the primary recipient is. It has been fun learning from Tony, and learning molecular biology together.



The image shows a screenshot of a Gmail compose email interface. At the top, there are three buttons: "Send", "Save Now", and "Discard". Below these is the "To:" field, which contains the text "Grandma" followed by a redacted email address. Below the "To:" field are links for "Add Cc" and "Add Bcc", and a note that says "Also include: Tony Zador". The "Subject:" field contains the text "Happy Birthday Grandma!". At the bottom, there are links for "Attach a file" and "Insert: Invitation".

Contents

| | |
|--|------------|
| List of Figures | xi |
| List of Tables | xiv |
| Glossary | xv |
| 1 Introduction | 1 |
| 1.1 The importance of single-neuron connectivity | 2 |
| 1.2 Current approaches to the connectome | 4 |
| 1.3 Sequencing the Connectome | 5 |
| 1.3.1 Costs | 8 |
| 1.3.2 Advantages and limitations of the sequencing approach | 9 |
| 1.3.3 Extensions of the method | 10 |
| 1.3.4 Conclusions and perspectives | 12 |
| 1.4 Acknowledgements | 12 |
| 2 In vivo generation of DNA sequence diversity for cellular barcoding | 13 |
| 2.1 Abstract | 13 |
| 2.2 Introduction | 14 |
| 2.3 Materials & Methods | 15 |
| 2.3.1 <i>In silico</i> simulations | 15 |
| 2.3.2 Synthesis of barcode cassettes | 16 |
| 2.3.3 Plasmid construction | 16 |
| 2.3.4 Bacterial culture & shuffling | 17 |
| 2.3.5 Barcode reconstruction | 18 |
| 2.3.6 Sequences | 18 |

| | | |
|----------|---|-----------|
| 2.4 | Results | 19 |
| 2.4.1 | Design of Cre-based barcoding | 19 |
| 2.4.2 | Employing DNA invertases for cassette shuffling | 22 |
| 2.4.3 | Design and synthesis of a 5BC cassette | 23 |
| 2.4.4 | Testing of the 5BC cassette <i>in vivo</i> | 25 |
| 2.4.5 | High-throughput sequencing of shuffled 5BC | 27 |
| 2.4.6 | High-throughput sequencing of shuffled 11BC | 30 |
| 2.5 | Discussion | 30 |
| 2.6 | Supplementary Material | 36 |
| 2.6.1 | Supplementary Notes | 36 |
| 2.6.1.1 | Supplementary Note 1 | 36 |
| 2.6.1.2 | Supplementary Note 2 | 36 |
| 2.6.2 | Supplementary Figures | 37 |
| 2.6.3 | Sequences of relevant genetic elements | 42 |
| 2.6.3.1 | 5BC_Cassette (synthesized by IDT) | 42 |
| 2.6.3.2 | BCextension (synthesized by GeneWiz) | 42 |
| 2.6.3.3 | Rci sequence (from NCBI Reference Sequence: NC_013120.1 REGION: complement 7291274066) | 43 |
| 2.6.4 | Plasmid maps | 44 |
| 2.6.5 | Description of Supplementary Files | 44 |
| 2.6.6 | Sequencing data files | 45 |
| 2.7 | Acknowledgements | 45 |
| 3 | Encoding neural connectivity into DNA | 46 |
| 3.1 | Abstract | 46 |
| 3.2 | Introduction | 46 |
| 3.3 | Results | 49 |
| 3.4 | Materials and Methods | 66 |
| 3.4.1 | HEK293 cultures | 66 |
| 3.4.2 | Neuronal cultures | 66 |
| 3.4.2.1 | Xona microfluidic cultures | 66 |
| 3.4.3 | Plasmid construction | 66 |
| 3.4.4 | Sindbis Production | 67 |

| | | |
|---------|--|----|
| 3.4.4.1 | Sindbis titering | 67 |
| 3.4.5 | Barcode Library Production | 68 |
| 3.5 | Tagging with BG/BC derivatives | 69 |
| 3.5.1 | Immunoprecipitation | 69 |
| 3.5.1.1 | Protein immunoprecipitation | 69 |
| 3.5.1.2 | Elution | 69 |
| 3.5.1.3 | Western Blotting | 69 |
| 3.5.1.4 | RNA immunoprecipitation | 70 |
| 3.5.2 | Proximity Ligation Assay | 70 |
| 3.5.2.1 | Electron Microscopy of PLA signals | 71 |
| 3.5.3 | Emulsion Overlap PCR | 72 |
| 3.5.3.1 | Design and construction of microfluidic device | 72 |
| 3.5.3.2 | Generation of emulsion droplets | 72 |
| 3.5.3.3 | One-step overlap reverse-transcription PCR | 72 |
| 3.5.4 | Nested PCR & DNA sequencing | 73 |
| 3.6 | Supplementary Material | 74 |
| 3.6.1 | RNA design | 74 |
| 3.6.2 | Sequences of RNAs | 74 |
| 3.6.2.1 | preRNA coding sequence | 74 |
| 3.6.2.2 | postRNA coding sequence | 75 |
| 3.6.3 | Protein design | 76 |
| 3.6.3.1 | Neurexin1B Amino Acid sequence | 76 |
| 3.6.3.2 | Neurologin1AB Amino Acid sequence | 76 |
| 3.6.3.3 | Protein-protein interaction | 77 |
| 3.6.3.4 | RNA-binding | 77 |
| 3.6.3.5 | Coding Sequence: Myc-CLIP-Nrx1B- λ N(i) | 78 |
| 3.6.3.6 | AA Sequence: Myc-CLIP-Nrx1B- λ N(i) | 79 |
| 3.6.3.7 | Coding Sequence: HA-SNAP-Neurologin1AB- λ N(i) | 80 |
| 3.6.3.8 | AA Sequence: HA-SNAP-Neurologin1AB- λ N(i) | 82 |
| 3.6.4 | Protein-RNA interaction | 83 |
| 3.7 | Acknowledgements | 83 |

| | |
|---|------------|
| 4 Discussion | 85 |
| 4.1 Scaling to whole brains | 85 |
| 4.1.1 Transgenics | 86 |
| 4.1.2 Enabling bi-directional tracing | 86 |
| 4.1.3 Preventing endogenous interactions | 86 |
| 4.1.4 Efficient RNA joining | 87 |
| 4.2 Porting to other organisms | 87 |
| 4.3 Missing information | 88 |
| 4.3.1 Spatial information | 88 |
| 4.3.2 Cell types | 89 |
| 4.3.3 Cell physiology | 89 |
| 4.3.4 Neuromodulation | 90 |
| 4.3.5 Gap Junctions | 90 |
| 4.4 Future directions | 91 |
| 4.4.1 <i>In situ</i> sequencing | 91 |
| 4.5 Conclusions | 91 |
| References | 92 |
| 5 Declaration | 97 |
| Appendices | 99 |
| A Rci in mammalian cells | 100 |
| A.1 Introduction | 100 |
| A.2 Expression of Rci in mammalian cells | 100 |
| A.2.1 Cellular localization of Rci | 100 |
| A.3 Functional testing of Rci in mammalian cells | 101 |
| A.3.1 Rci-mediated inversion | 101 |
| A.3.1.1 Rci-mediated inversion on a plasmid substrate | 102 |
| A.3.1.2 Rci-mediated inversion on a genomic substrate | 102 |
| A.3.2 Rci-mediated deletions | 103 |
| A.4 Discussion | 106 |
| A.5 Acknowledgements | 107 |

| | |
|---|------------|
| B Mapping long-range projections with high-throughput sequencing | 108 |
| B.1 Introduction | 108 |
| B.2 Results | 110 |
| B.2.1 Tracing of neurons originating in auditory cortex | 110 |
| B.3 Discussion | 112 |
| B.4 Methods | 113 |
| B.4.1 Injections | 113 |
| B.4.2 RNA isolation | 114 |
| B.4.3 RT-PCR | 114 |
| B.5 Acknowledgements | 114 |
| C RNA ligation methods | 115 |
| C.1 Introduction | 115 |
| C.2 General considerations | 115 |
| C.2.1 Maintaining the Protein-RNA complexes | 115 |
| C.2.2 Stoichiometry of ligation reactions | 116 |
| C.2.3 Pol-II transcripts | 116 |
| C.3 ssRNA ligation | 116 |
| C.4 Splint ligation | 117 |
| C.4.1 Splinted ligation of short oligonucleotides | 117 |
| C.4.2 Adapting splinted ligation for Pol-II transcripts | 118 |
| C.4.2.1 RNaseH cleavage | 120 |
| C.4.2.2 Ligation | 122 |
| C.4.2.3 Strand replacement | 125 |
| C.4.2.4 Strand replacement and ligation | 125 |
| C.5 Catalytic nucleic acids | 127 |
| C.5.1 Ribozymes | 127 |
| C.5.2 Deoxyribozymes | 129 |
| C.6 Joining by reverse transcription | 129 |
| C.6.1 Overlap extension reverse transcription | 129 |
| C.6.2 dsDNA primed reverse transcription | 130 |
| C.7 Discussion | 131 |
| C.8 Methods | 132 |

| | | |
|----------|---|------------|
| C.8.1 | Splint ligation | 132 |
| C.8.2 | RNAseH digestion | 132 |
| C.8.3 | Reverse transcription | 132 |
| C.9 | Acknowledgements | 132 |
| D | Non-transgenic, cell-type specific expression of Cre recombinase | 133 |
| D.1 | Introduction | 133 |
| D.1.1 | Current approaches | 134 |
| D.1.2 | Ribozyme mediated trans-splicing | 135 |
| D.2 | Results | 139 |
| D.2.1 | Cre recombinase | 139 |
| D.2.2 | Construction and testing of a SOM-targeting Ribozyme | 140 |
| D.2.3 | Identification of efficient trans-splicing sites on target mRNA | 142 |
| D.2.4 | Addition of an extended guide sequence | 142 |
| D.2.5 | Trans-splicing in stable cell-lines | 143 |
| D.3 | Discussion | 143 |
| D.4 | Materials and Methods | 145 |
| D.4.1 | Trans-splicing | 145 |
| D.4.2 | Flow Cytometry | 145 |
| D.5 | Acknowledgements | 145 |
| E | Sparse transcriptome profiling via nonspecific trans-splicing | 146 |
| E.1 | Introduction | 146 |
| E.2 | Results | 147 |
| E.3 | Discussion | 148 |
| E.4 | Materials and Methods | 151 |
| E.4.1 | Design of splicing cassette | 151 |
| E.4.2 | Mapping and analysis | 151 |
| E.5 | Acknowledgements | 151 |
| F | Color by number Brainbow | 152 |
| F.1 | Abstract | 152 |
| F.2 | Introduction | 153 |
| F.3 | Results | 154 |

| | | |
|---------|--|-----|
| F.4 | Discussion | 159 |
| F.5 | Supplementary Materials | 161 |
| F.5.1 | Injections | 161 |
| F.5.2 | Cryoslicing | 161 |
| F.5.3 | FISSEQ | 161 |
| F.5.3.1 | FISSEQ of endogenous transcripts | 161 |
| F.5.3.2 | FISSEQ of barcode transcripts | 162 |
| F.5.3.3 | FISSEQ Barcode RNA sequence | 162 |
| F.6 | Acknowledgements | 163 |

List of Figures

| | | |
|------|--|----|
| 1.1 | The wiring of neural circuits is highly structured | 3 |
| 1.2 | Barcoding of Individual Neuronal Connections | 7 |
| 1.3 | Cell-type reveals additional structure in the connectivity matrix | 11 |
| 2.1 | Design of Cre barcode cassettes | 21 |
| 2.2 | Design of Rci barcode cassettes | 24 |
| 2.3 | <i>In vivo</i> testing of Rci recombination on barcode cassettes | 26 |
| 2.4 | High-throughput sequencing of 5BC Rci cassette after <i>in vivo</i> shuffling . | 29 |
| 2.5 | High-throughput sequencing of 11BC Rci cassette after <i>in vivo</i> shuffling | 31 |
| 2.6 | Overview of an example <i>in vivo</i> barcoding experiment | 35 |
| 2.7 | Biases of the Cre barcode architecture | 37 |
| 2.8 | 5BC cassette stability during growth | 38 |
| 2.9 | T7 induced Rci expression | 39 |
| 2.10 | 11BC cassette stability during growth | 40 |
| 2.11 | Cassettes approach complete randomness as the number of recombina- tion events increase | 41 |
| 3.1 | Overview of NO-C | 48 |
| 3.2 | BG-PEG-(S-S)-Biotin-PEG-BC cross-linker | 49 |
| 3.3 | Protein design and testing in HEK cells | 51 |
| 3.4 | Membrane protein tagging in neurons | 52 |
| 3.5 | Proximity ligation assay for detecting transneuronal interactions | 53 |
| 3.6 | Employing PLA to find interacting protein pairs | 54 |
| 3.7 | Employing PLA to test for compatible RNA binding domains | 55 |
| 3.8 | RNA-IP with 1xλN fusion proteins | 57 |

LIST OF FIGURES

| | | |
|------|---|-----|
| 3.9 | Optimization of biochemistry in HEK cells | 58 |
| 3.10 | Characterization of system components in neurons | 60 |
| 3.11 | Schematic of overlap RT-PCR | 61 |
| 3.12 | Emulsion PCR for barcode joining | 62 |
| 3.13 | Probing emulsion occupancy via qPCR | 63 |
| | | |
| A.1 | Rci expression in mammalian cells | 101 |
| A.2 | Rci-mediated inversion in mammalian cells | 104 |
| A.3 | Rci-mediated excision in mammalian cells | 105 |
| | | |
| B.1 | Employing DNA sequencing for projection mapping | 111 |
| B.2 | Reconstruction of projectome from left auditory cortex | 112 |
| | | |
| C.1 | Splint Ligation of RNA | 118 |
| C.2 | Splinted ligation of short oligonucleotides | 119 |
| C.3 | Site-specific RNaseH cleavage directed by 2'Ome/DNA hybrids | 120 |
| C.4 | RNaseH-mediated cleavage of <i>in vitro</i> transcribed RNA | 121 |
| C.5 | Competition of targeting oligo and splint | 123 |
| C.6 | Dual cleavage and splint oligo | 124 |
| C.7 | Monitoring strand replacement with fluorescent oligos | 126 |
| C.8 | Strand replacement and ligation | 127 |
| C.9 | Bifunctional ribozymes for ligation | 128 |
| C.10 | Overlap reverse transcription | 130 |
| | | |
| D.1 | Cell-type specific expression via trans-splicing | 136 |
| D.2 | Function of Cre truncations | 141 |
| D.3 | Trans-splicing <i>in vivo</i> | 142 |
| | | |
| E.1 | Tagging cellular transcripts via promiscuous trans-splicing | 147 |
| E.2 | Barcode trans-splicing to defined transcripts | 149 |
| E.3 | Barcode trans-splicing genome wide | 150 |
| | | |
| F.1 | Overview of the FISSEQ neuronal tracing | 155 |
| F.2 | Barcode FISSEQ in a coronal brain slice | 157 |
| F.3 | Negative control - unbarcoded slice after FISSEQ | 159 |

LIST OF FIGURES

F.4 FISSEQ of endogenous transcripts in a coronal slice 160

List of Tables

| | | |
|-----|---|-----|
| 2.1 | Primers used in the Rci study | 18 |
| 2.2 | <i>sfx</i> sites used in Rci study | 19 |
| 2.3 | Barcode sequences used in the Rci study | 19 |
| 2.4 | Plasmids used in the Rci study | 19 |
| 3.1 | Summary of presynaptic proteins tested by PLA | 56 |
| 3.2 | Summary of postsynaptic proteins tested by PLA | 56 |
| 3.3 | Overlap reverse-transcription primers | 73 |
| A.1 | Constructs used for inversion and excision testing of Rci | 103 |
| F.1 | FISSEQ gene hits from coronal mouse brain sample | 158 |

Glossary

| | | | |
|------------------------------|--|---------------|--|
| λN | RNA binding domain N from λ phage | EtOH | ethanol |
| barcode | Unique, cell-identifying sequence of nucleic acids | FISSEQ | Fluorescence <i>in situ</i> sequencing |
| BC | Benzylcytosine - reactive moiety for CLIP-tag | GRASP | GFP Reconstitution Across Synaptic Partners |
| BG | Benzylguanine - reactive moiety for SNAP-tag | HEK | Human embryonic kidney cell line |
| boxB | 15bp RNA hairpin that binds with picomolar affinity to λ N protein | IP | Immunoprecipitation |
| bp | base pair | Kb | Kilobases |
| BSA | Bovine serum albumin | mRNA | messenger RNA |
| Bx | BG-PEG-Biotin-PEG-BC cross-linker | NLS | Nuclear localization signal |
| cDNA | complementary DNA | nt | nucleotide |
| CLIP | CLIP-tag self labeling protein | ORF | Open reading frame |
| connectome | The full wiring diagram of a brain | PBS | Phosphate buffered saline |
| DAPI | 4,6-diamidino-2-phenylindole | PCR | Polymerase chain reaction |
| DTT | Dithiothreitol | PEG | Polyethylene glycol |
| EM | Electron microscopy | PFA | Paraformaldehyde |
| | | PLA | Proximity Ligation Assay |
| | | qPCR | Quantitative real-time PCR |
| | | RCA | Rolling circle amplification |
| | | RNAseq | Sequencing of the RNA |
| | | rolony | Rolling circle colony |
| | | rRNA | ribosomal RNA |
| | | RT | Reverse transcription or Reverse transcriptase |
| | | SNAP | SNAP-tag self labeling protein |

1

Introduction

Heterogeneity is a defining feature of biological systems – from populations of bacterial cells to complex tissues such as the brain. A complete understanding of complex biological systems will require simultaneous high-throughput measurements across multiple scales. For example, the brain is composed of several dozen cell types connected by orders of magnitude more synapses. Deconstruction of a neural circuit, therefore, would minimally require measurements of connectivity between all of the cells in the circuit as well as their individual transcription profiles. The advent of high-throughput sequencing has provided researchers with access to data streams of unprecedented scale. By re-encoding information about individual cells and their interactions within a network into deoxyribonucleic acids (DNA), high-throughput sequencing can be harnessed to probe the dynamics of complex biological networks.

A surprising amount of evidence exists to suggest that individual cells, even when derived from apparently homogenous sources, exhibit a wide range of characteristics. This heterogeneity amongst cells has a profound implication on the overall function of the system. For example, bacterial populations exploit pre-established heterogeneity to survive during periods of atypical stress due to the presence of antibiotics and/or lack of available nutrients [Ryall et al. 2012]. Cells which make up tissues of higher organisms, despite their concordant function, have also been shown to consist of a heterogenous population of cells. This is perhaps most striking in the brain, where several dozen cell types interact to endow an animal with abstract functions such as perception, thought, and behavior. Neuroscientists seek cellular and circuit explanations of these complex

1.1 The importance of single-neuron connectivity

traits. For example, what are the roles of individual cell-types, and their pattern of connectivity, in computing the confidence of a decision?

Circuit level explanations of computation and behavior represent the gold standard. However, the circuit architectures that underlie these complex computations are mostly unknown. This is largely because of a lack of high-throughput methods for mapping neural connectivity at single neuron resolution. We have established a method and a suite of molecular components for achieving this goal that relies on DNA Barcoding of Individual Neuronal Connections (BOINC), which can subsequently be read out by standard high-throughput sequencing techniques [Zador et al. 2012].

1.1 The importance of single-neuron connectivity

Connectivity can be studied at different spatial scales. Conventional neuroanatomical methods probe the connectivity between brain regions. Such analysis reveals, for example, that the retina is connected to the visual thalamus, which in turn is connected to the visual cortex. The importance of mesoscopic connectivity in the mammalian brain is uncontroversial – different brain areas represent different kinds of information and have clearly distinct functions, so it is easy to see how knowing the connections among areas at the mesoscopic level will be useful. There are currently several major efforts to describe systematically the mesoscopic-scale connectivity of the mouse, macaque, and human brain [Bohland et al. 2009, Oh et al. 2014].

Mesoscopic connectivity represents the natural anatomical complement to conventional physiological approaches, such as extracellular recording, for studying how populations of neurons encode information and control behavior. However, such physiological approaches tend to obscure the identity and heterogeneity of the neurons under study. From the point of view of conventional extracellular recording, neurons within a brain area (e.g., visual area MT) differ only by their responses to sensory inputs and other external variables. Indeed, in physiological studies neurons are often referred to as interchangeable units; differences among nearby neurons are often attributed to random variation. Such assumptions are often incorporated into theoretical models, in which it is often assumed that cortical wiring is random, and therefore, only the statistical properties of neural connections, such as the average number of inputs per

1.1 The importance of single-neuron connectivity

neuron, need be specified [Hill et al. 2012, Koulakov et al. 2009, Shadlen and Newsome 1998].

In the absence of data about the relationship between the function of a neuron, its cell type, and its position within the local circuit, a description of connectivity at the mesoscopic level may seem sufficient. However, the circuits in Figure 1.1 illustrate how connectivity beyond the mesoscopic – at the level of synaptic contacts between pairs of individual neurons – can also be useful. In the motion detection circuit in Figure 1.1A, sequential activation of input neurons from left to right (1, 4, 7, 10) will generate less activity in the output neuron (0) than activation from right to left (10, 7, 4, 1). The lateral inhibition circuit in Figure 1.1B is wired similarly, but the addition of a few extra inhibitory connections renders it insensitive to directional motion.

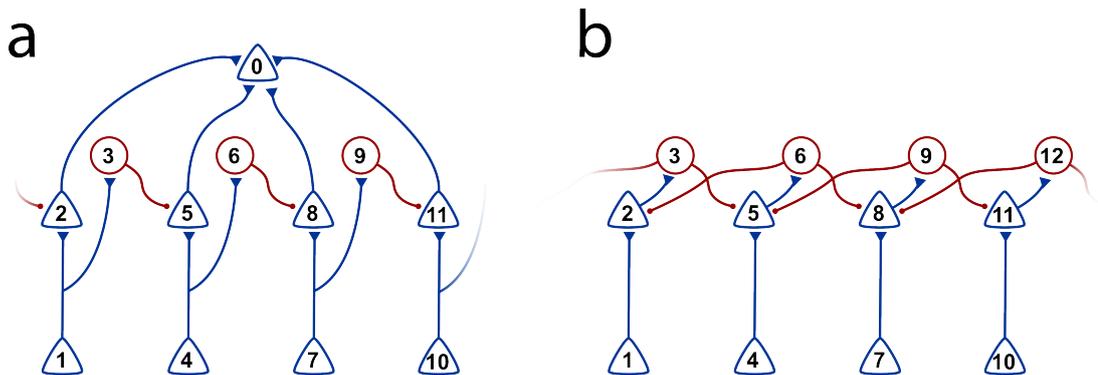


Figure 1.1: The wiring of neural circuits is highly structured - Two similar circuits in which the computation is readily deduced from the wiring. Circuit (a) is directionally selective, whereas (b) performs a center-surround computation. Red cells are inhibitory and blue cells are excitatory.

These simple examples reveal how detailed information about cell-types and connectivity can provide immediate insight into the computations a circuit performs and can generate hypotheses that can be tested physiologically. In practice, most computations are not understood at this level of precision. In part the reason is simply that detailed circuit information is largely unavailable. Indeed, the complete wiring diagram, or “connectome”, is known for only a single nervous system, that of the tiny worm *C. elegans*, with 302 neurons connected by about 7,000 synapses [Varshney et al. 2011, White et al. 1986]. Interestingly, the utility of the connectome in *C. elegans* is somewhat limited because function is highly multiplexed, with different neurons performing

1.2 Current approaches to the connectome

different roles depending on the state of neuromodulation [Bargmann 2012], possibly as a mechanism for compensating for the small number of neurons. Mammalian circuits contain orders of magnitude more neurons than *C. elegans*. Although neuromodulation is important in mammalian circuits – and should be measured in addition to wiring (see chapter 4), the need to multiplex function may not be as severe as in *C. elegans*, which may render the relationship between circuitry and function more transparent.

In mammals there is ample evidence that the connectivity of a neuron correlates with its function. For example, whether a neuron in primary visual cortex is simple or complex is correlated with cell layer; cell layer is in turn a surrogate for connectivity. Even more striking is the finding that neurons in primary visual cortex that project to the motion-sensitive area MT represent a homogenous population whose motion sensitivity is more similar to that of neurons in MT than to other V1 neurons [Movshon and Newsome 1996]. Observations such as these reinforce the notion that connectivity predicts function.

1.2 Current approaches to the connectome

There are currently two main approaches to determining single cell connectivity. The first is based on physiology. This approach can be quite powerful and has yielded tantalizing evidence of the precise nature of the connectivity of cortical circuits. In one series of experiments, Callaway and colleagues used laser scanning photostimulation to probe connectivity in visual cortex [Yoshimura and Callaway 2005]. They found that if two nearby neurons in layer 2/3 are connected, then they share input from single neurons in layer 4, but if they are not connected they do not share input. Thus, the input from layer 4 to layer 2/3 appears to consist of at least two independent subnetworks, which happen to overlap in space. In a different set of experiments, Chklovskii and colleagues [Song et al. 2005] used whole cell methods to assess connectivity among triplets of neurons. By enumerating all 16 possible ways that three neurons can be connected, they discovered that several connectivity motifs were overrepresented above the chance levels predicted by the pairwise connection probabilities. Thus connectivity among triplets of cortical neurons deviates markedly from the null hypothesis of random connectivity. Unfortunately, physiological approaches do not readily scale up to an entire brain. Nevertheless, findings such as these hint at the rich structure yet to

be uncovered in cortical circuits and motivate the development of higher throughput technologies.

The second approach is based on electron microscopy (EM). EM is required because light microscopy (LM) does not have sufficient resolution to establish whether two nearby neuronal processes are merely close or whether they have actually formed a synapse. Reconstruction of serial electron micrographs has yielded what to date is the only complete connectome, that of *C. elegans* [Varshney et al. 2011, White et al. 1986]. However, even for this simple nervous system, the reconstruction required a heroic effort – over 50 person-years of labor to collect and analyze the images. The difficulty of EM-based reconstruction arises from the fact that stacks of many individual images need to be aligned to track each axonal or dendritic process back to the soma; misalignment of even a single pair of images can result in an error in the wiring diagram, rendering the reconstruction of long-range connections particularly challenging. It is a testament to the importance of the connectivity problem that several research groups have made remarkable progress in automated EM reconstruction [Bock et al. 2011, Chklovskii et al. 2010, Lichtman and Denk 2011].

Several recent technical advances raise the possibility that a third class of approaches, based on light microscopy, may succeed in mapping circuit connectivity. GRASP (GFP Reconstitution Across Synaptic Partners) [Feinberg et al. 2008, Kim et al. 2012, Yamagata and Sanes 2012] and ID-PRIME [Liu et al. 2013] allow synaptic contacts to be resolved at the level of light microscopy. Brainbow [Livet et al. 2007] can be used to trace axons and dendrites over considerable distance. Spatial resolution, difficulties in tracing from cell body to synapse, and the limited color depth of fluorophores limit these approaches in practice to sparse connectivity mapping.

1.3 Sequencing the Connectome

Here we propose to exploit high-throughput DNA sequencing to probe the connectivity of neural circuits at single-neuron resolution. Sequencing technology has not previously been applied in the context of neural connectivity, but the sequencing approach has tremendous potential. The advantage of sequencing is that it is already fast – sequencing billions of nucleotides per day is now routine – and, like microprocessor technology,

getting faster exponentially. Moreover, the cost of sequencing is plummeting. It currently costs less than \$5,000 to sequence an entire human genome, and the race is on to reach the \$1,000 genome. Thus, by converting brain connectivity from a problem of microscopy to a problem of sequencing, it becomes tractable using current technology.

BOINC, the method we propose for converting connectivity into a sequencing problem, can be broken down conceptually into three components (Figure 1.2). First, each neuron must be labeled with a unique sequence of nucleotides – a DNA barcode (Figure 1.2A). The requisite barcoding is conceptually similar – though different in detail – to the generation of antibody diversity by B cells in the immune system through somatic recombination. The idea of barcoding individual neurons (see chapter 2 for details) is inspired by Brainbow, except that here DNA sequences substitute for fluorophores (XFPs). The advantage of using sequences is diversity: whereas Brainbow allows for at most hundreds of color combinations, a barcode consisting of even 20 random nucleotides can uniquely label $4^{20} = 10^{12}$ neurons, far more than the number of neurons (10^8) in a mouse brain.

Second, barcodes from synaptically connected neurons must be associated and amplified (Figure 1.2B). We set out to create such a system with several design criteria: (1) scalable to a full brain, (2) easily ported to a variety of organisms (without major re-design), (3) minimally toxic to the host organism, and (4) unperturbed synapse formation and maintenance. One potential way to associate a pre- and postsynaptic barcode is by means of a transsynaptic virus such as rabies virus (RV) [Wickersham et al. 2007] or pseudo-rabies virus (PRV) [Ekstrand et al. 2008]. These viruses have evolved exquisite mechanisms for moving genetic material across synapses and have been used extensively for tracing neural circuits in rodents. However, scaling viral transsynaptic spread to a full brain would require complex engineering in order to achieve exclusively monosynaptic spread from every neuron. Moreover, transsynaptic viruses such as RV and PRV are highly toxic and exhibit limited host range. Therefore, we instead decided to associate barcodes by way of a highly engineered synapse-specific protein-protein interaction (for details see chapter 3). Once associated, barcode pairs are joined and amplified.

Finally, barcode pairs are sequenced and the data is processed. If upon sequencing we observe pre-synaptic barcode 2 with post-synaptic barcodes 21 and 24, we can infer that neuron 2 is connected to neurons 21 and 24. Since most neurons are only sparsely

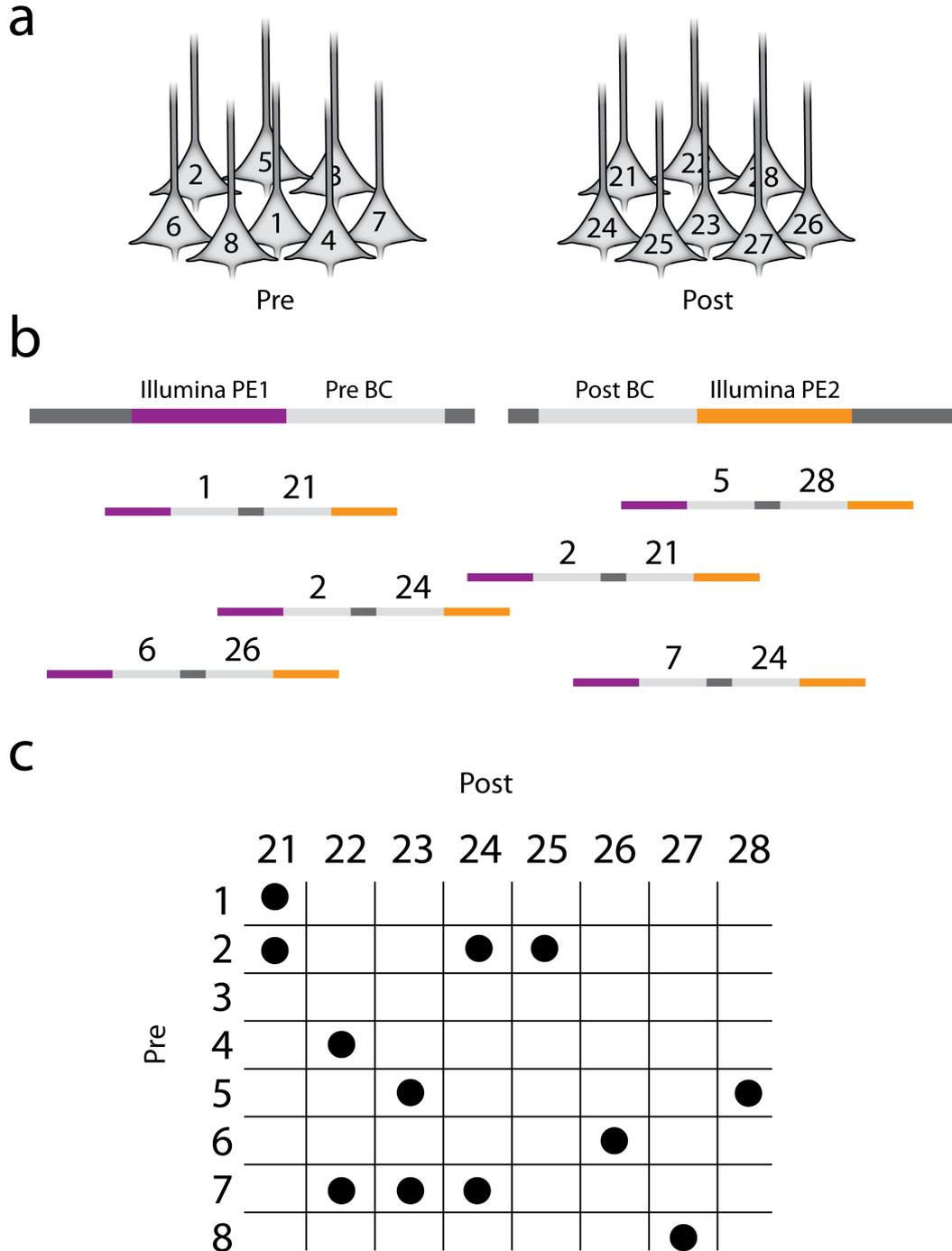


Figure 1.2: Barcoding of Individual Neuronal Connections - There are many possible implementations for each step of BOINC. (a) First, each neuron must be labeled with a unique sequence of nucleotides – a DNA barcode. For simplicity, here we consider a unidirectional connection between two populations of neurons, pre (1-8) and post (21-28). (b) Second, barcodes from synaptically connected neurons must be associated, joined, and amplified. (c) Finally, barcode pairs are subjected to high-throughput sequencing to reconstruct a connectivity matrix.

connected to other neurons in the brain – for example, in the mouse cortex a typical neuron is connected with perhaps 10^3 of its 10^8 potential partners – only a small subset of the potential pre-post barcode pairs will actually be observed. Thus upon high-throughput sequencing, we can fill in the non-zero elements of the sparse connectivity matrix (Figure 1.2C).

In its simplest form the sequencing approach yields only a connectivity matrix. Missing from this matrix are at least two kinds of useful information typically obtained with conventional methods based on microscopy: information about the brain region (e.g., primary auditory cortex, striatum, etc.) from which each barcode originates, and information about the cell type (e.g., dopaminergic, fast-spiking GABAergic, etc.) of each barcoded neuron. However, several strategies can be used to augment the connectivity matrix with both kinds of information (see subsection 1.3.3 and chapter 4). Thus, as sequencing-based connectivity analysis matures, it may generate a view of connectivity similar to that provided by traditional methods.

1.3.1 Costs

In the 3.5 years between the introduction of next generation DNA sequencing technologies in January 2008 to the most recent data in July 2011, the cost of sequencing fell by a factor of 1,000 [Wetterstrand 2013]. This 15-fold yearly rate of improvement far exceeds even Moores law, according to which computer costs drop 2-fold every 2 years. Just as Moore’s law drove and was driven by the computer revolution, so the drop in sequencing costs is driven by the prospect of a genomics revolution in medicine. Although such a precipitous rate of improvement of sequencing cannot be sustained indefinitely, it would not be surprising if commercial pressures were to drive costs down by another factor of 100 or more over the next few years.

How much would it cost to sequence the cortex of a mouse? We can put a lower bound on the current sequencing cost as follows. The mouse cortex consists of about $4.6 * 10^6$ neurons [Roth and Dicke 2005]. Suppose that each cortical neuron connects to about 10^3 other cortical neurons, so that there are $4.6 * 10^6 * 10^3 = 4.6 * 10^9$ connections. If we assume that each barcode is 20 nucleotides, then we have $4.6 * 10^9$ connections * 20 nucleotides/barcode * 2 barcodes/connection = $1.66 * 10^{11}$ nucleotides. Assuming that the fraction of unsampled connections is $exp^{-\frac{2k}{N}}$, where k is the number of reads and N is the number of barcodes, then with 3-fold oversampling ($4.6 * 10^{11}$ nucleotides)

we would expect to sample 95% of connections. At $\$0.1/10^6$ nucleotides (July 2013, for up to date information see: sequencing cost per megabase), this would cost \$48,000 and could easily drop several orders of magnitude in a few years. A similar calculation for *Drosophila melanogaster*, with 10^5 neurons and 10^7 connections, yields \$1/brain; and for *C. elegans*, with 302 neurons and 7,000 connections, sequencing costs would be essentially negligible. Although these are best case estimates and do not include costs other than sequencing, the possibility of a \$1 *Drosophila* connectome, or a \$1,000 mouse cortical connectome, emphasizes the promise of recasting neural connectivity as a sequencing problem. For an in-depth review of the costs of current and future connectomics methods see: [Marblestone et al. 2013].

1.3.2 Advantages and limitations of the sequencing approach

Like any method, the sequencing approach is subject to false positives (i.e., inferred connections that do not exist) and false negatives (actual connections that are missed). Although the prevalence of each type of error will depend on the details of the implementation, with the sequencing approach most errors will likely be false negatives. Possible sources of false negatives include failure of barcode association and undersampling of the amplified barcode pairs. Most sequencing errors will also result in false negatives, but these can be minimized by judicious design of the barcodes. Possible sources of false positives include loss of synapse specificity in the association of barcodes and insufficient diversity in the pool of possible barcodes. False positives are likely to be an important source of error in microscopy-based approaches in which inaccurate tracing of a neuronal process across tissue sections can lead to misattribution of a synaptic connection to the wrong parent.

The sequencing approach provides different information from conventional microscopy-based approaches. Electron microscopy provides a wealth of data not available in the sequencing approach, including information about neuronal morphology, as well as about the subcellular placement, number, and size of synapses. It is possible that some of this information could be preserved with extensions to the sequencing approach (see subsection 1.3.3), but in the first implementations this information will be missing. On the other hand, the sequencing approach has the potential to provide direct access to the molecular expression profile of individual neurons – whether it is dopaminergic or expresses a marker, such as parvalbumin, that tags the neuron as belonging to a

particular subtype of interneuron. Moreover, with the sequencing approach, local and long-range connections are equally accessible; by contrast, with microscopy the probability of inaccurately tracing a synaptic connection increases with distance, rendering the reconstruction of inter-areal connections a particular challenge.

1.3.3 Extensions of the method

In theory any cellular information that can be encoded into DNA can be incorporated into the sequencing approach. By tagging DNA-encoded information with the same cellular barcodes used for connectivity mapping, our approach can be extended to include a rich array of information. Information that is endogenously encoded in DNA, such as the epigenetic state of the cell or the cellular transcriptome profile, can very easily be encoded by joining the barcode with the nucleic acid information. Indeed, the structure of the connectivity matrix may largely depend upon information such as cell-type, as shown in Figure 1.3 [Pfeffer et al. 2013]. One implementation of this idea is discussed in Appendix E.

A new method, which relies on fluorescence *in situ* sequencing (FISSEQ) [Lee et al. 2014], has the promise of combining both the advantages of sequencing and imaging for circuit mapping. The technique was originally conceived for the high throughput sequencing of mRNA transcripts *in situ*, but can be repurposed for the sequencing of cellular barcodes. In theory, this technique could allow for the combination of functional imaging of cells during behavior (via 2-photon microscopy), mapping of neural circuit connectivity (via FISSEQ of cellular barcodes), and identification of individual cell types (via mRNA FISSEQ) in a single specimen [Marblestone et al. 2013; 2014].

Recently it has been proposed that other signals, such as the activity of a neuron, could be encoded in DNA [Alivisatos et al. 2012, Kording 2011]. There are countless other examples of additional information of interest including: the spatial coordinates of a cell within the brain, whether the synapse has been recently potentiated, the presence of neuromodulators, etc. that can theoretically be encoded into DNA. Because of its high information storage capacity, DNA is an attractive medium for encoding of different signals. Thus, the DNA-based connectome may one day serve as a foundation upon which additional layers of complex information about neural circuits can be added to uncover the fundamentals of neural circuit computation.

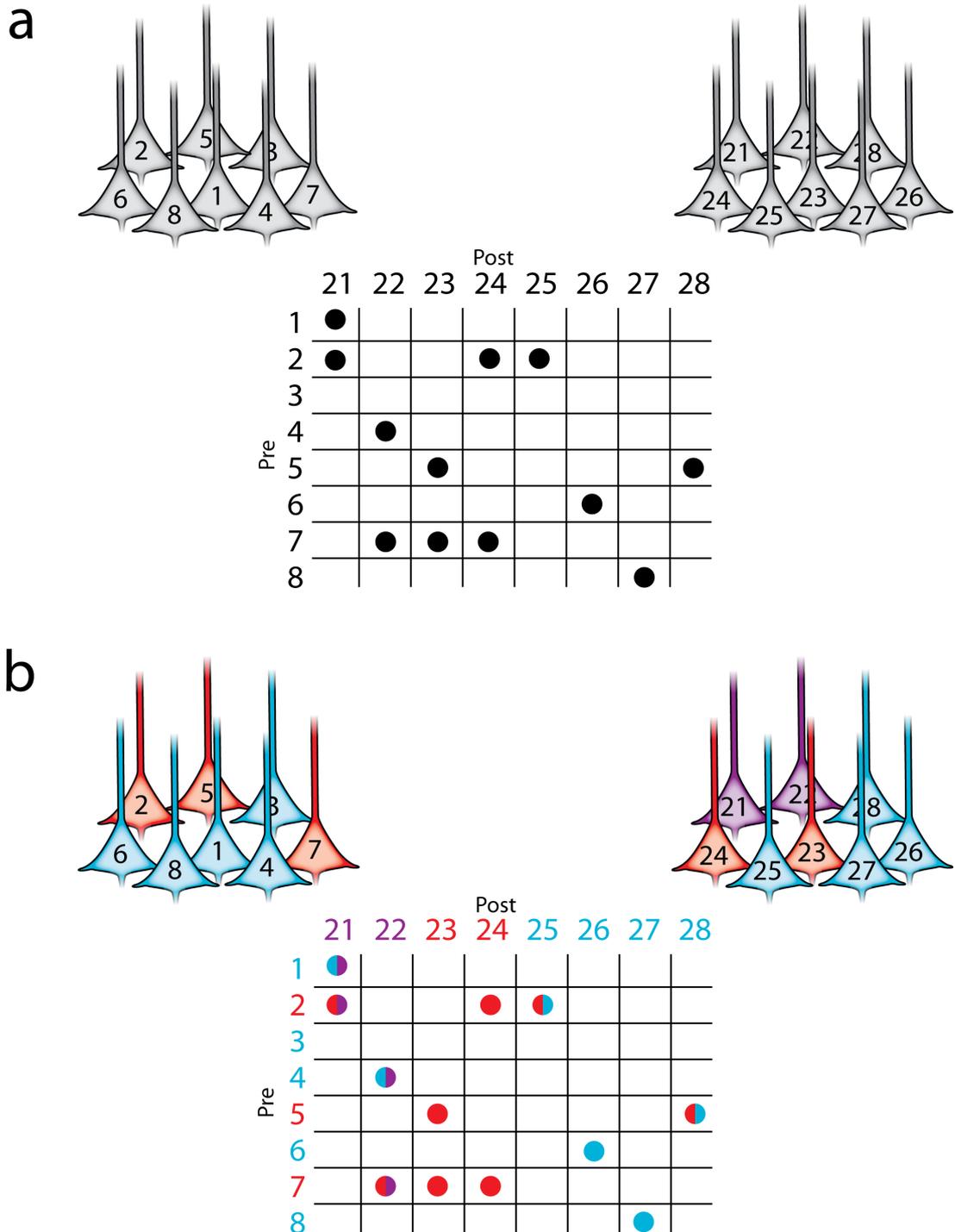


Figure 1.3: Cell-type reveals additional structure in the connectivity matrix -
 (a) Without knowledge of cell-type, a simple connectivity matrix can be reconstructed. (b) Knowledge of cell-type reveals additional high-order structure in the connectivity matrix (i.e. red cells receive input only from other red cells, whereas blue cells receive input from both red and blue cells.)

1.3.4 Conclusions and perspectives

The appeal of the sequencing approach rests in its promise of high throughput, as defined by cost and mapping time. Low-cost sequencing of brain circuits could be used as a screening test to generate hypotheses about how circuits change with development, learning, genetic manipulations, or any other experimental factor. For example, autism has been hypothesized to arise from genetic lesions that disrupt local and long-range connectivity, but different autism candidate genes may disrupt circuits differently [Geschwind and Levitt 2007]. High-throughput circuit screening would enable a systematic comparison of the similarities and differences among brain circuits in animal models of autism. A high-throughput circuit screen has the potential to transform how experiments are designed.

What will we learn from sequencing the connectome? Perhaps it is instructive to turn to the lessons learned from sequencing the human genome. Knowledge of the complete genome provides the starting point for much of modern biological research, transforming how research is conducted in the post-genomic era. A cheap and rapid method for deciphering the wiring diagram of an entire brain may have a comparably profound impact on neuroscience research.

1.4 Acknowledgements

The idea and original concepts for this project were devised by Anthony Zador. A version of this chapter was published in 2012 [Zador et al. 2012]. I contributed to the idea of barcoding via an invertase (see chapter 2) and extensions of the technique to identify cell-types, as well as to other ideas outlined in the text.

2

In vivo generation of DNA sequence diversity for cellular barcoding

2.1 Abstract

We have developed a novel method of uniquely tagging individual cells *in vivo* with a genetic “barcode” that can be recovered by DNA sequencing. Our method is a two-component system comprised of a genetic barcode cassette whose fragments are shuffled by Rci (recombinase for clustered inversion), a site-specific DNA invertase. The system is highly scalable, with the potential to generate theoretical diversities in the billions. We demonstrate the feasibility of this technique in *Escherichia coli*. Currently, this method could be employed to track the dynamics of populations of microbes through various bottlenecks. Advances of this method should prove useful in tracking interactions of cells within a network, and/or heterogeneity within complex biological samples.

2.2 Introduction

Reverse engineering any complex system requires the simultaneous monitoring of individual components. Recent advances in high-throughput DNA sequencing have given biologists unprecedented access to massively parallel data streams. Genetic barcoding – the labeling of individual cells with a unique DNA sequence – when combined with these technologies, will enable monitoring of millions or billions of cells within complex populations. This approach has proved useful in tagging neurons [Golden et al. 1995] and hematopoietic stem cells [Gerlach et al. 2013, Lu et al. 2011, Naik et al. 2013] for lineage analysis and could be applied to the normal and/or abnormal development of other cell populations or tissues, including tumours. Indeed, *in vivo* barcoding of individual neurons is the requisite first step towards converting neuronal connectivity into a form readable by high-throughput DNA sequencing [Zador et al. 2012].

Most current approaches for tagging individual cells with a genetic barcode rely on the creation of diverse libraries *in vitro* and subsequent delivery of genetic material into a host cell at low-copy number. Such *in vitro* approaches are limited by cloning bottlenecks that cause reduced library diversities and sequence biases, by incomplete labeling of all cells within a population, by the possibility of introducing multiple barcodes per cell, and by the challenges of working across organisms (e.g. retroviral barcoding cannot be applied in some organisms like *Caenorhabditis elegans*).

In vivo barcoding has the potential to overcome all of these limitations. Mechanisms for generating diversity *in vivo* exist, endogenously, in many organisms – most notably the mammalian immune system. However, efforts to repurpose the immune system’s V(D)J recombination for *in vivo* cellular barcoding [Gerlach et al. 2013] yielded limited barcode diversity – on the order of a dozen unique sequences – in cells other than lymphocytes [Schumacher 2014]. Exogenous recombinases have been successfully applied to generate diverse combinations of colors for cellular tagging purposes. This technique, better known as Brainbow [Livet et al. 2007], relies on Cre recombinase to rearrange a cassette resulting in the stochastic expression of a subset of different colored fluorescent proteins (XFPs) in neurons. The theoretical diversity of Brainbow is in the hundreds, but cannot be easily assayed with DNA sequencing because it relies heavily on gene copy number variation as well as recombination. We reasoned that by replacing XFPs with unique sequences, we could design a barcoding system with the

potential to achieve diversities that matched the scale of high-throughput sequencing technologies.

We have developed a novel method of generating sequence diversity *in vivo* for the purpose of cellular barcoding. Our method, which relies on a DNA invertase – Rci, recombinase for clustered inversion [Kubo et al. 1988] – to shuffle fragments of DNA, has the potential to easily achieve diversities over 10^9 unique sequences. Here, we show that this method can be applied for the *in vivo* generation of diversity in *E. coli*.

2.3 Materials & Methods

2.3.1 *In silico* simulations

We performed *in silico* simulations to determine the behavior of different cassette architectures. For Cre-based cassettes, *ncell* cassettes ($ncell = 10000$) of *nfrag* fragments ($nfrag = 100$) were operated on independently. Each fragment was flanked on its 5' end with a *loxP* site in sense orientation (5'-GCATACAT-3') and on its 3' end with a *loxP* site in the antisense orientation. Concatenation of fragments resulted in cassettes in which adjacent fragments (excluding ends) were separated by two *loxP* sites in opposing orientation. We defined Cre recombination as two independent binding events to *loxP* sites. Binding of Cre to a pair of *loxP* sites always resulted in recombination, where the result (inversion or excision) was defined by the relative orientation of the sites defined in a look up table (updated after each event). Completion was defined to be the point at which Cre can no longer mediate an excision event. The number of recombination events required to reach completion was tracked for each cassette. For Rci-based cassettes, *ncell* cassettes ($ncell = 10000$) of *nfrag* fragments ($nfrag = 10$) were operated on independently. Here, we considered two architectures. For the first architecture, each fragment of the cassette was flanked on its 5' end with an *sfx* site in sense orientation and on its 3' end with an *sfx* site in the antisense orientation. Concatenation of fragments resulted in cassettes in which adjacent fragments (excluding ends) were separated by two *sfx* sites in opposing orientation. For the purpose of simulations, we consider each pair of *sfx* sites between fragments to be equivalent to a single bidirectional *sfx* site. We defined Rci recombination as two independent binding events to *sfx* sites. In this case, binding of Rci to a pair of *sfx* sites always resulted in recombination (inversion). Simulations were allowed to proceed for m recombination

events per cassette. For the second Rci architecture, the 5' end of the cassette begins with a single *sfx* site in sense orientation, followed by a single sequence fragment. The cassette is extended by addition of an *sfx* site and a sequence fragment, with the orientation of *sfx* sites alternating throughout the cassette. The cassette is terminated at its 3' end by an *sfx* site in antisense orientation. We defined Rci recombination as two independent binding events to *sfx* sites. Binding of Rci to a pair of *sfx* sites only resulted in recombination if the *sfx* sites were in opposite orientations (inversion only). Simulations were allowed to proceed for m recombination events per cassette. The code for running all simulations is provided in supplementary materials (section 2.6).

2.3.2 Synthesis of barcode cassettes

A 5-fragment barcode cassette was synthesized as a minigene by IDT and inserted into a standard plasmid (IDP190). Several different *sfx* sites were used to avoid perfect inverted repeats to simplify DNA synthesis. The cassette was constructed as ANCHOR105-*sfx*101R-BC1-*sfx*102L-BC2-*sfx*106R-BC3-*sfx*112L-BC4-*sfx*109R-BC5-*sfx*101L-ANCHOR56 where R indicates the *sfx* sequence in 5'-3' orientation and L indicates the reverse orientation (where orientation is determined by the core sequence: 5'-GTGCCAA-3'). The barcode cassette was synthesized with flanking sequences for use in PCR amplification and sequencing (#3:ANCHOR105 and #4:ANCHOR56). In addition, the cassette was flanked by restriction sites (PciI) and flanking primer sequences (#1:BC.F and #2:BC.R) for subsequent polymerase chain reaction (PCR) and cloning.

2.3.3 Plasmid construction

The 5-fragment barcode cassette (from plasmid #IDP190) was amplified using primers #1:BC.F and #2:BC.R, digested with PciI, and cloned into pet22b. A strain of *E. coli* harbouring the pEK204 plasmid, which encodes Rci recombinase (NCBI Reference Sequence: NC_013120.1), was ordered from NCTC (NCTC 13452: J53-derived *E. coli*. GenBank accession number EU935740). The open reading frame of Rci was obtained by PCR of the pEK204 plasmid with primers #7:NdeI-Rci.F and #8:NotI-Rci.R, which add restriction sites NdeI and NotI, respectively. Rci was cloned into plasmid pet22b using restriction sites NdeI and NotI (thus removing the periplasmic localization signal of pet22b) to create Plasmid #IDP205:(T7→Rci; 5BC). The constitutively

active promoter, pKat (Registry of Standard Biological Parts: BBa_I14034, http://parts.igem.org/Part:BBa_I14034), and a ribosomal binding site (RBS): AGGAGG, flanked by restriction sites BglIII and NdeI were synthesized as complementary oligos #9:s-pKat_promoter and #10:as-pKat_promoter, annealed, digested, and subsequently cloned into #IDP205 to make plasmid #DIG35:(pKat→Rci, 5BC). To extend the barcode cassette we amplified #IDP205 using primers #11:BamHI-BC5_F and #12: NheI-BC5_R, and cut with BamHI and NheI. The insert, BC6-BC11, was digested from BCextension (synthesized by GeneWiz) with BamHI and NheI. The backbone and insert were ligated to make plasmid #DIG70:(T7→Rci, 5BC). The constitutively active promoter, pKat, and a ribosomal binding site (RBS) flanked by restriction sites BglIII and NdeI were synthesized as complementary oligos #9:s-pKat_promoter and #10:as-pKat_promoter, annealed, digested, and subsequently cloned into #DIG70 to make plasmid #DIG71:(pKat→Rci, 11BC). All cloning was performed using Top10 chemically competent cells (Invitrogen) with growth at 37°C.

2.3.4 Bacterial culture & shuffling

For initial tests with T7 induced expression of Rci, plasmids IDP205 or DIG35 were transformed into *E. coli* strain BL21(DE3) (NEB) and grown in 5ml of normal or OvernightExpress (Millipore) supplemented media overnight. Plasmid DNA was isolated and the Rci coding sequence was removed (to prevent further shuffling) by double digestion (NdeI-NotI), blunting (Mung Bean), and re-ligation (Roche Rapid Ligation kit). The transformed ligation was plated for clonal analysis. Clonal analysis involved the selection of single colonies, growth in LB for 16 hours, plasmid isolation, and Sanger sequencing with ANCHOR105 used as a primer. For tests of the pKat driven expression, plasmids IDP205 or DIG35 were transformed into *E. coli* strain Top10 cells (Invitrogen) and grown in 5ml of LB overnight. Plasmid DNA was isolated and the Rci coding sequence was removed (to prevent further shuffling) by double digestion (NdeI-NotI), blunting (Mung Bean), and re-ligation (Roche Rapid Ligation kit). The transformed ligation was plated for clonal analysis. Clonal analysis involved the selection of single colonies, growth in LB for 16 hours, plasmid isolation, and Sanger sequencing with ANCHOR105 used as a primer. For high-throughput sequencing by PacBio, plasmids IDP205, DIG35, and DIG71 were transformed into Top10 cells (Invitrogen) and grown overnight in 50mL of LB. Plasmid DNA was isolated and digested with PciI to release

the barcode cassette. The barcode cassette was prepared for PacBio sequencing using the PacBio SMRTbell Template Prep Kit according to the manufacturers instructions and cassettes from each original plasmid were sequenced on a single SMRT cell. PacBio sequences were collapsed into circular consensus reads using the PacBio command line tools.

2.3.5 Barcode reconstruction

Our sequence alignment algorithm is written in Matlab and uses the Matlab Bioinformatics Toolbox. Sequencing reads are processed independently. Each known barcode fragment is aligned to the sequence read (in both orientations) with a thresholded Smith-Waterman alignment [Smith and Waterman 1981] and the position of the segment along the read is stored. The threshold is set by a bootstrap method. Briefly, one hundred randomly generated 100-mers are aligned to all of the sequence traces and a score is associated with each alignment. The mean score plus two standard deviations is considered the threshold for all subsequent alignments. The complete barcode can be reconstructed based on the positions of each segment. Because the algorithm relies only on local alignment, this method is extremely robust to sequencing errors.

2.3.6 Sequences

Table 2.1: Primers used in the Rci study

| | | |
|----|------------------|---|
| 1 | BC_F (5BC) | gctttacctgcactgcccagagtg |
| 2 | BC_R (5BC) | agactcatgcatcgtgccactgtgttcg |
| 3 | ANCHOR56 | ttgcgaacctcatcactcgttgc |
| 4 | ANCHOR105 | tgaggcaaggaagatgctgtcc |
| 7 | NdeI-Rci_F | tctacatatgccgtctccacgcatccgt |
| 8 | NotI-Rci_R | actggcgccgcttacagcggtgtgctgc |
| 9 | s-pKat_promoter | acctagatctcattattgcaattaataaacaactaacggacaattctacctaacaaggaggtaccatgatgacct |
| 10 | as-pKat_promoter | aggtcatatggtacctccttgtagtagtagaattgtccgtagttgtttattaattgcaataatgagatctaggt |
| 11 | BamHI-BC5_F | tctagatccggcaatacttctgtccaatccgg |
| 12 | NheI-BC5_R | actggctagcgaacgagtgatgaggttcgcaa |

Table 2.2: *sfx* sites used in Rci study

| | |
|---------------|----------------------------------|
| <i>sfx101</i> | ggcaataactttcgtgccaatccggtacgtgg |
| <i>sfx102</i> | ctgctggcctacgtgccaatccggtacgtgg |
| <i>sfx106</i> | gcacgtaggccagtgccaatccggtacctgg |
| <i>sfx109</i> | ttctctgcaagcgtgccaatccggtacgtgg |
| <i>sfx112</i> | gccaagcttgttgtgccaatccggtacgtgg |

Table 2.3: Barcode sequences used in the Rci study

| | |
|------|--|
| BC1 | ccaattggtagtttgcagaactcagattttaacagcagaggacgcatgctctaccttcatgatccactgacgtccctgaggctgcaatacatgcaacg |
| BC2 | aggcagctccggcgttaagtcttagtgcaatggcgctttttacctcgtcctcgagaagaggggacgccagtgagatctttaatgfggtaattggg |
| BC3 | aggactcttggccctccgcttaggcagtgacactctccataaacgggctgttagttatggcgtccgaggattcaaaaaggtagcgaactcggccg |
| BC4 | atccggagagacgggcttcaagctgcctgacgacggttgcgcgtccgtatcaaaatcctcctaataagccccctcactgttgggtgaagagcccagga |
| BC5 | cgggttggccagatgtgcgattatatacgttaaatggctcttggcccggtgcttacctgagggaattgagccctccgttaattcccttgcataca |
| BC6 | ggatccgtcaaatgtgattgatccctcgatccccgtggagatgagatgcctgctggtcggggtgcaaaccgatcaataacaatcgtcacttgcaggt |
| BC7 | ccccgactgatgcctaaacctgcaggtctttaggatattgacgcttgacgatgtcccagattaaaccggtgtgcaaccttggatgctgattaatcgc |
| BC8 | gcgagatgaatggacgggtctggtcgaacgatgtattaatgatgcaggagccccgcacacctaataatcgatccgggtatgttaattggtgcgatgc |
| BC9 | ggctccctctcactactacgggtatcggctcccagccccagctgcgctctgttctcaaccaactaggatctgatgcacgagattaaccttgacgttgt |
| BC10 | tacaccggcccgagcgtcgtcctttctatagatggcttccagcagctaccaggaggttctggttgggaactacttcaaacgctatgaccttccctat |
| BC11 | tgctgaacagtttaataactgggacttattctgggactgatagggtatgacgcttcttatgttctccggcgtacagtgaaaaaaatcaatcctatac |

Table 2.4: Plasmids used in the Rci study

| | |
|--------|------------------------|
| IDP190 | 5BC Cassette |
| IDP205 | T7→Rci; 5BC Cassette |
| DIG35 | pKAT→Rci; 5BC Cassette |
| BCext | BC6-BC11 extension |
| DIG70 | T7→Rci; 11BC Cassette |
| DIG71 | T7→Rci; 11BC Cassette |

All other relevant sequences and plasmid maps are provided in Supplementary Materials (section 2.6).

2.4 Results

2.4.1 Design of Cre-based barcoding

Our design goals were to create a modular genetically encoded barcode system that is easily scalable, cross-platform (applicable across model organisms), compatible with high-throughput sequencing technologies, and robust to sequencing errors. Our ulti-

mate goal is to generate unique barcodes to label all of the cells of the mouse cortex – approximately 1×10^7 neurons [Roth and Dicke 2005]. In general, if the repertoire of possible barcodes is substantially greater than the number of cells in the population of interest, then even randomly generated barcodes will label most cells uniquely. Specifically, if n is the number of cells and k is the number of possible barcodes, then under simple assumptions the fraction of uniquely labeled cells will be $e^{-n/k}$. Thus assuming one barcode per cell, a barcode repertoire exceeding the number of cells by a factor of 100 will yield 99% uniquely labeled cells. Therefore, we sought a barcode system with the potential to scale to at least $100 \times 10^7 = 10^9$ unique barcodes. Our initial design relied on Cre recombinase to shuffle and pare down a cassette of n barcode fragments (each fragment flanked by *lox* sites in alternating orientation) by stochastic inversion and excision events to leave a single fragment (Figure 2.1A).

Cre acts by binding to and mediating recombination between two of its cognate DNA sequences, called *lox* sites. Cre-mediated recombination between any two compatible *lox* sites in the same orientation causes the excision of the intervening sequence elements. In contrast, recombination between *lox* sites in opposing orientation causes the inversion of any intervening sequence elements.

This architecture (shown in Figure 2.1A) has a theoretical diversity of $2n$ after completion, since any fragment, j , of the n fragments can end in either the forward or inverted orientation (Figure 2.1A). Several variant *lox* sites have been discovered with a wide variety of characteristics [Lee and Saito 1998, Missirlis et al. 2006, Siegel et al. 2001]. The Brainbow [Livet et al. 2007] system, for example, employed three *lox* sites that were shown to be mutually incompatible including *loxP* (the wild-type *lox* site), *loxN* [Livet et al. 2007], and *lox2272* [Lee and Saito 1998] (Figure 2.1B). We reasoned that a Cre-based barcoding approach could be extended to achieve higher diversities by concatenating k cassettes of n fragments, where each cassette employs one of a subset of incompatible *lox* sites (Figure 2.1C). Here, the theoretical diversity is $(2n)^k$ because each cassette operates independently. Thus with $n = 100$, $k = 4$, theoretical diversities reach our goal of 10^9 unique sequences.

Unfortunately, many of the reported *lox* variants have not been validated for complete incompatibility or pairwise efficiency. Moreover, because of the repetitive nature of *lox* sites it would be difficult to synthesize cassettes with the hundreds of fragments required to achieve our target diversity. For example, an architecture with $n = 100$,

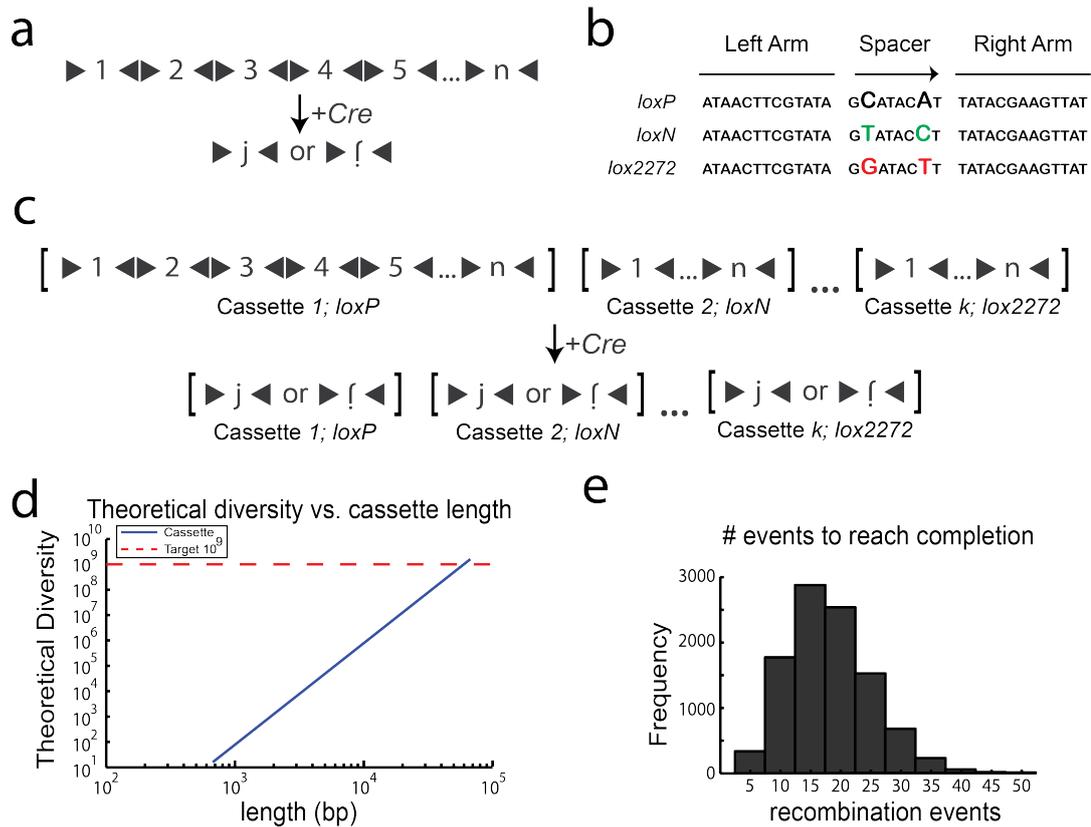


Figure 2.1: Design of Cre barcode cassettes - (a) n sequence fragments, each flanked by *loxP* sites (black triangles) in opposing orientations are concatenated to form a cassette. Upon Cre recombination, the cassette is shuffled and pared down until a single fragment, j , remains in either orientation. (b) Several variant mutually incompatible *lox* sites have been described, each of which has a pair of mutations in the spacer sequence. (c) The diversity attainable by Cre recombination can be increased by concatenating k cassettes, where each cassette utilizes an incompatible *lox* site. Cre recombination results in the independent shuffling of each of the k cassettes. (d) The diversity as a function of length for cassettes containing $n = 100$ fragments (100bp each) concatenated together. Lengths $> 60kb$ are needed to reach the requisite diversity of 10^9 . (e) The distribution of the number of recombination events needed for each cassette to reach completion.

$k = 4$, where each fragment is above the minimum length ($\sim 100\text{bp}$) for efficient recombination [Hoess et al. 1985], requires a large genomic insertion with a length greater than 60kb (Figure 2.1D). Finally, simulations suggested that Cre-based architectures are subject to considerable biases that limit the diversities that can be achieved in practice (subsubsection 2.6.1.1, Figure 2.7).

2.4.2 Employing DNA invertases for cassette shuffling

The key limitation of Cre-based designs is that Cre mediates both inversion and excision/insertion. Because insertion is a bimolecular reaction whereas excision is a unimolecular reaction, in general equilibrium will favour excision over insertion. Thus the equilibrium diversity of a Cre-based cassette scales linearly with the number of fragments n . In simulations (see subsection 2.3.1), 10 to 40 recombination events were performed on each cassette before reaching completion (Figure 2.1E). The diversity of the Cre-based cassettes could, in principle, be increased by preventing the reaction from proceeding to completion. In the limit, if only inversions were permitted, then the fragments of the cassette would be shuffled rather than pared down. Intuitively, the advantage of eliminating excision events can be understood by analogy to a deck of n playing cards, in which each card can occur in either orientation (face up or face down). If excisions dominate, then the diversity is given by eliminating all but one card. If inversions dominate, then the diversity is given by all the possible sequences of n shuffled cards ($n!$) multiplied by all of the possible orientations ($2n$). Thus eliminating excisions allows the diversity (given by Equation 2.1) to increase supra-exponentially, rather than linearly, with the number of elements n .

$$D(n) = 2^n n! \tag{2.1}$$

With only 10 fragments, the diversity reaches $> 3 \times 10^9$ unique sequences. Importantly, the equilibrium state of this architecture, as the number of recombination events m approaches infinity, is an equal distribution of all potential unique sequences [Wei and Koulakov 2012]. In practice, m will reach some finite number that may result in cassette biases. However, simply extending the cassette by one fragment can compensate for any modest biases of this architecture. Moreover, because the barcodes are made of a small number of known sequence fragments, reconstruction of barcode

sequences even from highly error prone sources becomes possible. The order and orientation of each fragment within the cassette after recombination can be determined simply by performing $2n$ pairwise alignments (each fragment in both orientations is aligned to the recovered sequence). This results in barcodes that are robust to many classes of sequencing errors.

We thus adopted a strategy based on DNA invertases - recombinases that can mediate only inversions [Johnson 2002]. Rci (recombinase for clustered inversion) is a site-specific recombinase (SSR) of the integrase (Int) family, of which Cre is also a member [Kubo et al. 1988]. Rci recognizes 31bp *sfx* sites and mediates recombination events only between sites in inverted orientation (inversions) – it cannot mediate excision events between sites in the same orientation. Unlike other inversion systems, such as Hin and Gin [Johnson et al. 1988], Rci does not appear to require any co-factors or enhancer sequences [Gyohda and Komano 2000]. Because of this, we selected Rci as our recombinase and designed a new barcode cassette in which segments of DNA are shuffled by inversion-only recombination events (Figure 2.2A).

2.4.3 Design and synthesis of a 5BC cassette

Initially, we planned to synthesize DNA in which $n = 5$ fragments, each flanked by *sfx* sites in opposite orientation, were concatenated to form a barcode cassette. After concatenation, each fragment is separated from its immediate neighbor by two *sfx* sites in opposite orientation (similar to the architecture proposed for Cre in Figure 2.1A). However, DNA synthesis constraints and plasmid stability forced us to redesign the cassette to minimize the effects of repetitive elements (the *sfx* sites). Ultimately, we employed an architecture in which individual fragments are separated by a single *sfx* site the orientation of which alternates between successive fragments (Figure 2.2A), relying on several compatible *sfx* site variants to further reduce complexity [Gyohda et al. 2002] (Figure 2.2B). This architecture achieves a somewhat lower diversity. Fragments originating in odd positions within the cassette (i.e. position 1, 3, 5, etc.) can only occupy odd positions after recombination. Likewise, fragments originating in even positions can only occupy even positions after recombination. This leads to a reduced diversity, given by Equation 2.2.

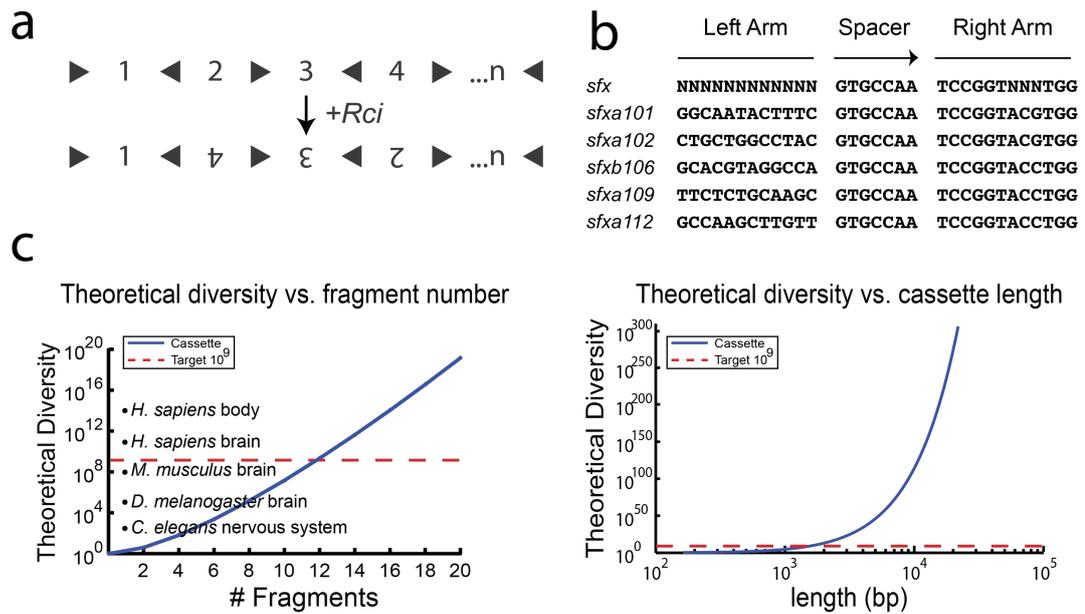


Figure 2.2: Design of Rci barcode cassettes - (a) The Rci cassette is formed by concatenation of n sequence fragments, in which individual fragments are separated by a single *sfx* site the orientation of which alternates between successive fragments. Upon Rci recombination, the cassette is shuffled. Because Rci cannot mediate excision events, the cassette remains the same length, but the relative position of each fragment within the cassette can change. (b) *sfx* sites share a central spacer sequence and right arm sequence, but have little sequence conservation in the left arm. 5 different *sfx* sequences were used in this study. (c) The diversity of the Rci cassette as a function of the number of fragments, n . Cassettes with > 20 fragments can generate sufficient diversity to label all cells in complex organs/organisms. (d) The diversity as a function of length for Rci cassettes (assuming 100bp fragments). Lengths ~ 1 -2kb are needed to reach the requisite diversity of 10^9 .

$$D(n) = 2^n \left(\frac{n}{2}!\right)^2; \quad n \text{ is even} \quad (2.2a)$$

$$D(n) = 2^n \left(\frac{n+1}{2}\right) \left(\frac{n-1}{2}!\right)^2; \quad n \text{ is odd} \quad (2.2b)$$

where D is the total diversity and n is the number of fragments in the cassette. Despite the reduction of diversity due to the modified architecture, only 12 fragments are required ($n = 12$) to achieve our target diversity of $> 10^9$. Additional segments greatly increase the diversity making this a scalable approach (Figure 2.2C). Moreover, unlike the Cre-based architecture, the Rci-based cassettes reach our requisite diversity at reasonable cassette lengths of ~ 1 -2kb (Figure 2.2D). Ultimately, a 5-fragment (100bp fragments) barcode cassette was synthesized utilizing 5 different *sfx* site sequences to decrease the repetitive nature of the cassette to aid in synthesis and replication. In addition, known anchor sequences (ANCHOR105 and ANCHOR56) positioned at either end of the cassette, were added outside of the recombination region to aid in sequence reconstruction. The final 5-fragment cassette, 5BC (Figure 2.3A), was cloned into a low-copy plasmid containing the Rci gene (resulting in plasmid IDP205). This ensures that all barcode cassettes that are transformed into bacterial cells will be exposed to the Rci coding sequence. Plasmid IDP205 (T7 \rightarrow Rci; 5BC) contains the Rci gene driven by the inducible T7 promoter. This plasmid is remarkably stable, showing no signs of recombination in the absence of induction across many generations (Figure 2.8).

2.4.4 Testing of the 5BC cassette *in vivo*

We transformed two populations of *E. Coli* BL21-DE3 (NEB) cells with plasmid IDP205 (T7 \rightarrow Rci; 5BC) and grew 10mL cultures overnight. One culture was grown under conditions that induce the expression of Rci from the T7 promoter (see methods). After growth, cells were plated for clonal analysis on plates that did not support Rci expression. Twenty colonies were chosen for each condition (+ and - Rci induction) and analyzed by Sanger sequencing. Without induction of Rci expression, no recombination was observed (0 of 20 colonies sequenced, data not shown). Moreover, the induction of Rci led only to modest recombination - shuffling the cassette in 8 of the 20 reconstructed barcode sequences (Figure 2.3B). Interestingly, each of the final products could be explained by a single recombination event (Figure 2.9).

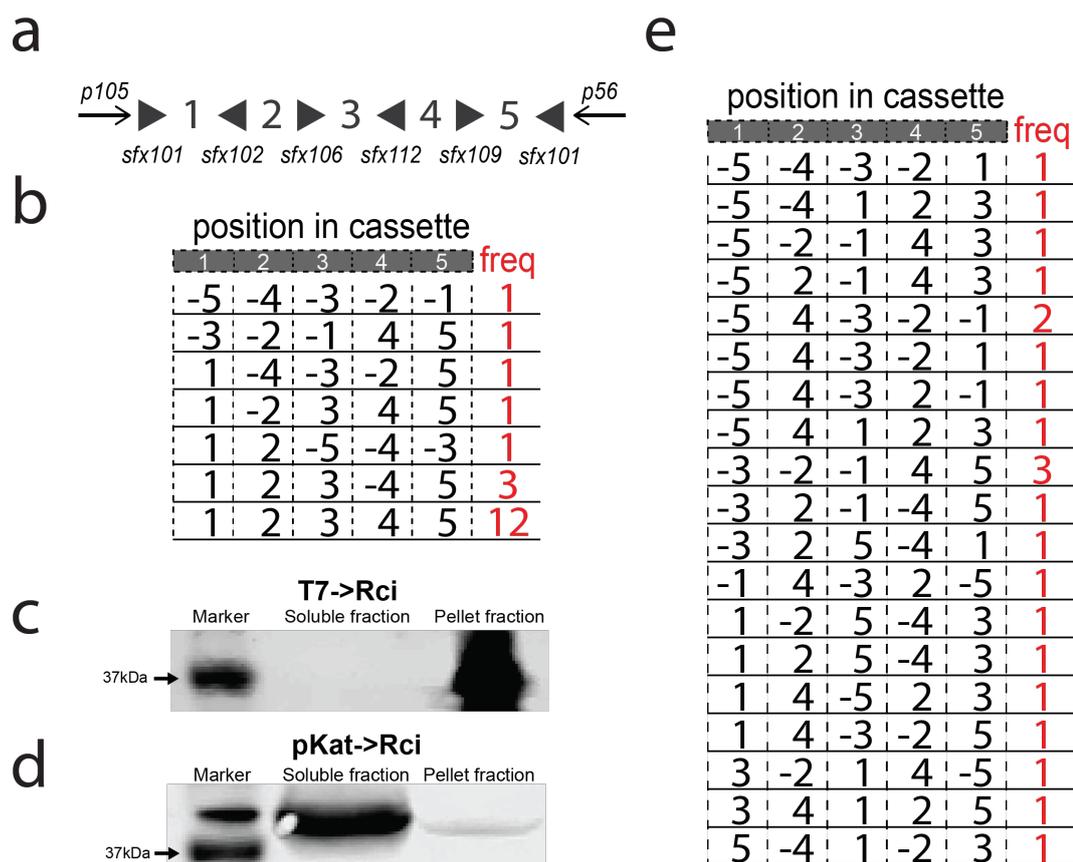


Figure 2.3: *In vivo* testing of Rci recombination on barcode cassettes - (a) A barcode cassette containing 5 fragments, which were separated by various *sfx* sites in alternating orientations, was synthesized. Anchor sequences flanked the cassette for use in PCR and sequence reconstruction. (b) Induction of Rci expression from the T7 inducible promoter results in shuffling in some, but not all cassettes. (c) Induction of Rci expression from the T7 inducible promoter results in Rci protein that is largely insoluble. (d) Expression of Rci from a medium-strength, constitutively active promoter results in expression of soluble Rci protein and (e) robust shuffling of the cassette.

We reasoned that the inefficient shuffling might be due to protein aggregation and insolubility due to high overexpression, as is often the case with T7 overexpression [Moore et al. 1993]. To test this we fused an HA-tag at the N-terminus of Rci and tested the expression level and solubility via western blot. Indeed, we found that HA-Rci was found only in the insoluble fraction (Figure 2.3C), perhaps explaining the inefficient shuffling observed. Thus, we tested expression of HA-Rci from a different promoter, a medium strength constitutively active promoter, pKat (Registry of Standard Biological Parts: BBa_I14034), and found that the protein was soluble when expressed from this promoter (Figure 2.3D). Therefore, we cloned the pKat promoter in place of T7 to make plasmid DIG35 (pKat→Rci; 5BC) and tested for shuffling efficiency. Briefly, we transformed DIG35 (pKat→Rci; 5BC) into Top10 competent cells and grew cultures overnight. To stop shuffling, plasmid DNA was isolated and the sequence for the Rci gene was removed via restriction enzyme digestion. Plasmids were re-transformed and colonies were selected for Sanger sequencing. DNA was isolated from each colony and subjected to Sanger sequencing. Sequence reads were analyzed with our alignment algorithm in order to reconstruct full barcodes. Reads that could not be fully reconstructed from sequencing data were discarded from further analysis.

Expression of Rci from this promoter, pKat, resulted in robust shuffling (Figure 2.3E). Of the 22 reconstructed (3 sequences failed reconstruction) cassette sequences, each had undergone shuffling to yield 19 unique barcode sequences. Moreover, all of the final barcode sequences could only be explained by multiple (>1) recombination events. Based on these positive preliminary results, we next subjected the 5BC cassette to high-throughput DNA sequencing.

2.4.5 High-throughput sequencing of shuffled 5BC

Advances in high-throughput sequencing have allowed for unprecedented access to massive quantities of DNA sequence data. We took advantage of high-throughput sequencing to sequence the shuffled 5BC cassettes at depths that allowed for a more thorough analysis of the actual *in vivo* behavior of barcode generation by Rci. Because of the length of our potential cassettes (~ 1 -2kb to reach diversities of 10^9), we chose the PacBio sequencing platform.

Briefly, we transformed *E. coli* Top10 (Invitrogen) cells with either plasmid IDP205 (T7→Rci; 5BC – negative control) or DIG35 (pKat→Rci; 5BC) and allowed cultures

to grow overnight. DNA was isolated and digested to release the barcode cassette. The barcode cassettes were then prepared for high-throughput sequencing on the PacBio RS II and sequenced.

Using our algorithm, we reconstructed 5887 barcodes from 7203 circular consensus reads (see subsection 2.3.5) obtained from cells in which Rci was not expressed (plasmid IDP205; T7→Rci, 5BC). Of the reconstructed barcodes, 5886/5887 gave the original sequence (Figure 2.4A).

These data indicated that the PacBio sequencing platform could handle the highly repetitive nature of the barcode cassettes and would allow for high-throughput sequencing of cassettes without the introduction of recombination during sequencing from template switching or other sources. When Rci was expressed off of the pKat promoter (DIG35) the cassette was shuffled robustly (Figure 2.4B). Here, we reconstructed 5243 barcodes from 6105 circular consensus reads, of which there were 203 unique sequences. After shuffling, each position along the cassette is populated with a relatively even distribution of all of the possible fragments, with only slight biases at the ends of the cassette (Figure 2.4B). In other words, the occupancy at each position of the cassette, while still preferential for the original fragment sequence, approaches randomness (Figure 2.4C). This important observation indicates that there are no biological constraints on our design that prohibit the full exploration of the barcode space. Simulations of shuffling of a 5BC cassette under simple assumptions (see section 2.3 for details) show that the occupancy at any given position reaches equilibrium after ~ 6 or more recombination events per cassette (Figure 2.4D). Comparison of the simulated data and the data collected from *in vivo* shuffling suggests that our cassettes likely experienced 2-3 recombination events on average *in vivo* (Figure 2.4D).

Even this limited amount of recombination resulted in an observed 203 unique sequences (out of a theoretical 384; $n = 5$, diversity = $2^5 x 3x(2!)^2$). Intuitively, after a small number of recombination events, the cassette remains biased at each position to its initial state (Figure 2.4D). As the number of recombination events increases, however, the cassette goes to equilibrium, and there is a nearly uniform distribution of each fragment at each position in the cassette (Figure 2.4D). In the limit, the cassette will reach an equilibrium state in which every possible barcode is equally probable (16).

This proof of principle experiment shows that *in vivo* recombination by Rci on a cassette is efficient at shuffling the original sequence into a unique barcode. However,

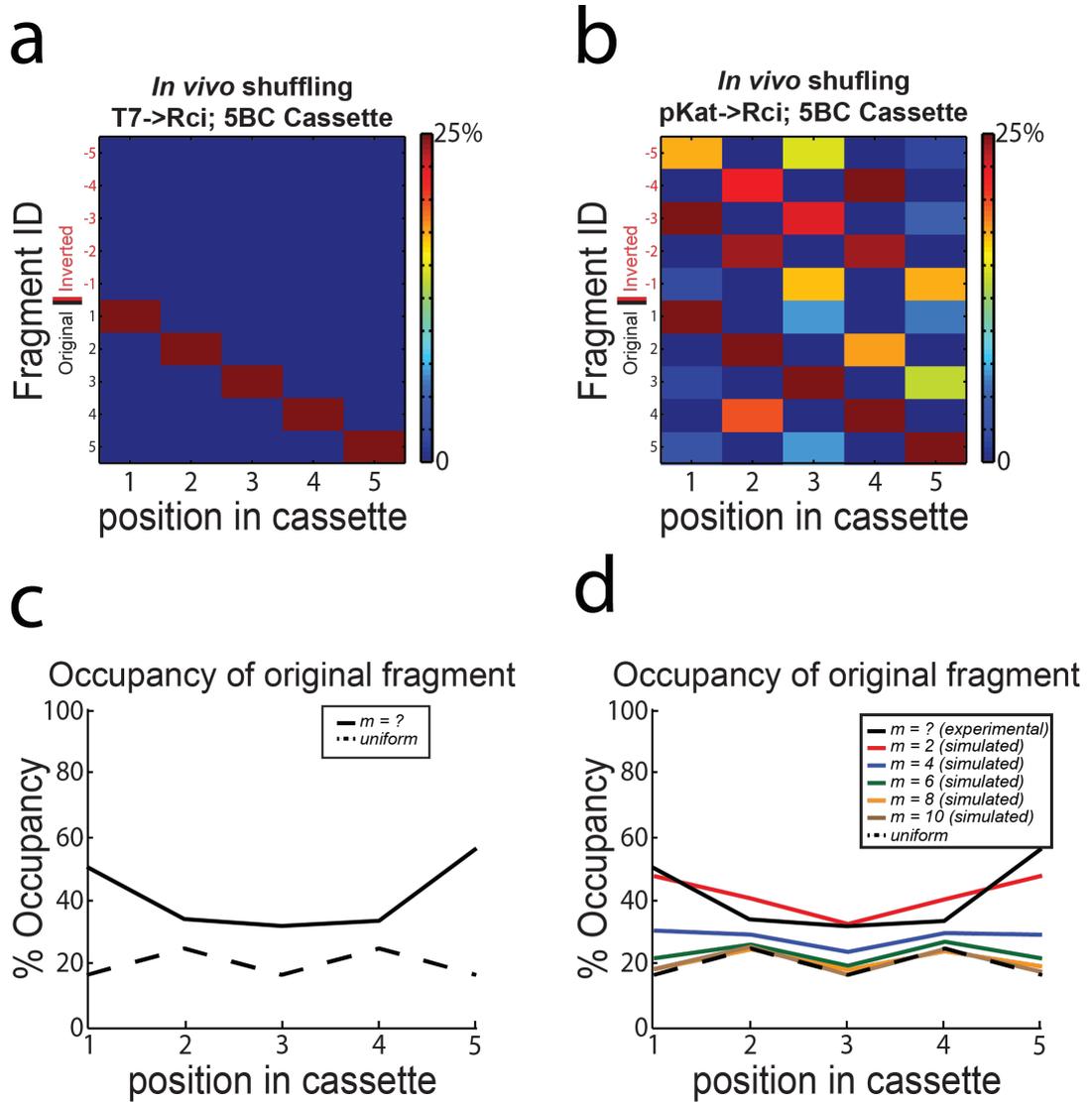


Figure 2.4: High-throughput sequencing of 5BC Rci cassette after *in vivo* shuffling - (a) Cassettes are stable in the absence of Rci expression and can be reconstructed from PacBio sequencing data. (b) Expression of Rci from the constitutively active pKat promoter results in robust shuffling of the 5BC cassette. The colormaps show the distribution of fragment occupancy at each position in the cassette. Colors are scaled from 0 to 25% for visualization. (c) The bias at each position was calculated as the number of times the original fragment appeared in its original position divided by the number of cassettes. The dotted black line indicates the expected occupancy of the original fragment at each position in a cassette with completely random occupancy. (d) The bias at each position within a cassette subjected to $m = 2, 4, 6, 8,$ or 10 recombination events was simulated. The dotted black line indicates the expected occupancy of the original fragment at each position in a cassette with completely random occupancy. The solid black line indicates the data observed *in vivo*.

the theoretical diversities of the 5BC cassette are well below our initial goals. Therefore, we sought to expand the cassette to achieve higher diversities.

2.4.6 High-throughput sequencing of shuffled 11BC

To explore the feasibility of achieving diversities that are capable of labeling large populations of cells, we expanded the cassette to 11 fragments. We synthesized a 6-fragment extension to our original cassette and concatenated this with our original 5BC cassette to create an 11BC cassette (Figure 2.5A). Importantly, plasmids harbouring this cassette were stable across many generations and showed no evidence of recombination in the absence of Rci expression (Figure 2.10).

The theoretical diversity of this cassette is $2^{11} \times 6 \times (5!)^2 = 176,947,200$. Unfortunately, there is currently no sequencing technology that has both the requisite depth and read-length to appropriately cover the potential diversity of the 11BC cassette. Nevertheless, we used high-throughput sequencing on the PacBio platform to sample the barcodes produced by the recombination of the 11BC cassette. Briefly, we transformed Top10 bacterial cells with plasmid DIG71 (pKat→Rci; 11BC cassette) and cultured the cells overnight. The barcode cassette was released by restriction digestion and subjected to high-throughput sequencing on the PacBio RS II.

Shuffling of the 11BC cassette *in vivo* was efficient (Figure 2.5B). We reconstructed 1786 barcodes, of which 1723 were unique. Again we observe that, after shuffling *in vivo*, each position along the cassette is populated with a relatively even distribution of all of the possible fragments (Figure 2.5B), approaching a completely random cassette (Figure 2.5C). Based on simulations, we estimate that the 11BC cassettes have experienced more than 5 recombination events on average (Figure 2.5D). As suggested by simulations, further shuffling will lead to an increasingly random cassette (Figure 2.11).

2.5 Discussion

To our knowledge, this is the first example of an *in vivo* barcoding scheme with the potential to scale to uniquely label all of the individual cells of an entire tissue or organism. The system, which takes advantage of a recombinase to shuffle fragments within a cassette has several advantageous characteristics. Recombinases similar to Rci (i.e. Cre, FLP, PhiC31) have been successfully employed across a wide variety of organisms,

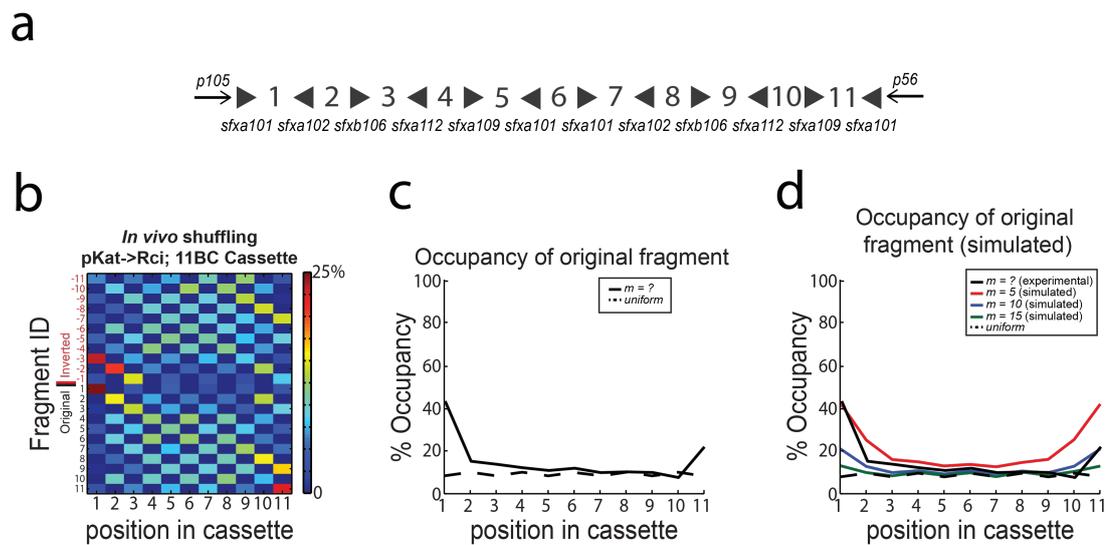


Figure 2.5: High-throughput sequencing of 11BC Rci cassette after *in vivo* shuffling - (a) A 6-fragment extension was concatenated with the 5BC cassette to yield the 11BC cassette. (b) Expression of Rci from the constitutively active pKat promoter results in robust shuffling of a high-diversity 11BC cassette. The colormaps show the distribution of fragment occupancy at each position in the cassette. Colors are scaled from 0 to 25%. (c) The bias at each position was calculated as the number of times the original fragment appeared in its original position divided by the number of cassettes. The dotted black line indicates the expected occupancy of the original fragment at each position in a cassette with completely random occupancy. (d) The bias at each position within a cassette subjected to $m = 5, 10,$ or 15 recombination events was simulated. The dotted black line indicates the expected occupancy of the original fragment at each position in a cassette with completely random occupancy. The solid black line indicates the data observed *in vivo*.

suggesting that our barcoding system could be easily ported to organisms beyond *E. coli*, either by expressing Rci or alternatively, by employing designs that take advantage of asymmetric mutant recombination sites [Colloms et al. 2014] of other recombinases. In addition, the system is highly scalable – addition of a single fragment to a cassette results in an exponential gain in diversity. Moreover, because the input space is small (only a handful of unique segments), each segment can be designed to be maximally orthogonal, thus rendering barcode readout highly robust to DNA sequencing errors. We took advantage of this fact in designing our barcode reconstruction algorithm, which relies on local alignment between the known input fragments and the final imperfect sequence recovered from high-throughput sequencing.

In its current form, however, this paradigm has at least several shortcomings that will need to be addressed before the system can be used at a larger scale. First, the expression of the Rci protein must be controlled in order to induce shuffling at a specific time point and then stopped to prevent further recombination events. In theory this can be accomplished through the use of an inducible promoter (i.e. T7 promoter, Tetracycline-responsive promoter, etc.). In practice, the expression level and recombination efficiency will need to be monitored at different levels of induction to permit solubility, genomic stability, and optimal recombination kinetics. Second, the cassettes, despite their design, were still subject to various poorly understood biases. Further work will be needed on Rci (and other recombinases) to determine the pairwise efficiencies between different recombination sites, the efficiency of recombination as a function of length between recombination sites, and to increase the recombination kinetics to achieve the maximal number of recombination events per unit time. As we saw in our simulations, a higher number of recombination events leads to far greater diversity and less bias. The biases that we detected, particularly in the case of the 5BC cassette, were likely exacerbated by the fact that we introduced a homogeneous barcode into exponentially dividing cells, where the kinetics of cell division likely outpace that of recombination (at least initially). In terminally differentiated cells, such as neurons, this is less of an issue as the expression of the recombinase can be sustained for long periods of times (i.e. weeks or months). However, in the case of dividing cells, recombinase expression must be pulsed for short durations to allow shuffling and then abruptly stopped to permit lineage tracing. An additional limitation of our design is that the barcodes are carried on extrachromosomal plasmids and thus cannot be used

for lineage tracing in organisms that do not allow for plasmid replication and inheritance. Introduction of the cassette into the genome of *E. coli* can be accomplished by homologous recombination and should be straightforward. Genetic engineering tools such as ZFNs, TALENS, or CRISPR will allow introduction of barcode cassettes in other organisms.

There are two factors that need to be considered in terms of the compatibility of our technique with current DNA sequencing technology. The first is read depth, or the number of amplicons that can be read in a single DNA sequencing run. The second is the read length, or the number of bases that can be obtained for a given amplicon which is currently the main limitation that constrains the application of our barcoding system. Over the last several years, both read depth and read length have increased steadily. It is likely that this trend will continue and that sequencing technology will soon be able to economically meet the needs of our barcoding system. Furthermore, additional research on Rci could allow for the optimization of barcode fragment lengths to permit recombination with shorter fragments – thereby decreasing the length of the cassette to be within reach of current next-gen sequencing platforms. Nevertheless, it is useful here to review the current high-throughput sequencing technologies and the compatibility with the barcoding approach outlined here.

Current Illumina technology produces ~ 8 billion reads per flow cell, which is enough to measure millions of barcoded cells at sufficient depth (assuming a uniform 100 reads per barcode). The read-depth of high-throughput sequencing is thus well matched to our current system. The other factor that must be considered is the available read length of current high-throughput sequencing technology. The longest read lengths currently offered by Illumina are for the MiSeq platform, which currently offers $2 \times 300 = 600$ bp reads. This limitation severely constrains the scale of our approach. With 600bp reads we can sample barcode cassettes of length < 600 bp. With our current design of 100bp fragments separated by 31bp *sfx* sites, we can read cassettes with only $n = 4$ barcode fragments (~ 550 bp). This limits the achievable diversity to < 100 unique barcodes. “Hacking” techniques are available that can push the read length of the Illumina technology to longer read lengths; 2×500 has recently been demonstrated [Birol et al. 2013]. Alternatively, stripping a primer after a certain number of reads and rehybridizing with a new primer would allow for multiple subreads from the same amplicon – thus allowing probing of the sequence at various sites for reconstruction of

the full barcode [Illumina 2014, Mellor and Roth 2014]. Moreover, it is likely that this technology will continue its steady rollout of improvements to both read-depth and read-length that will allow for a well-matched economical technology for sequencing barcodes generated by our method.

Alternative sequencing platforms, including Roche 454 or Pacific Biosciences (used here), offer different specifications that may be more applicable currently. The PacBio platform allows $\sim 100k$ reads of $>1kb$, and the newest Roche sequencing platform, the GS FLX Titanium XL+, offers read lengths of up to 1000bp for $\sim 700,000$ reads. This would allow monitoring of at least 7000 unique sequences. At 1000bp our cassette design can reach $n = 7$ barcode fragments for a total diversity of $\sim 18,000$ barcodes. At this read length and depth, our barcoding technology is well matched and would allow tracking of ~ 7000 barcodes.

Within these constraints, our technology could be applied immediately to a dissection of dynamics of a microbial population during various stressors including antimicrobials, limited resources, or niche competition. Specifically, our technology is uniquely positioned to probe serial population bottlenecks, which remain poorly understood [Elena and Lenski 2003]. A population of cells carrying the barcode cassette can be exposed to transient barcode shuffling (Figure 2.6A, B, C). The pool can then be probed by DNA sequencing to test the original distribution of barcodes (Figure 2.6D). At this point, the population can be exposed to various stressors resulting in a population bottleneck that selects <1000 barcoded cells (Figure 2.6E). After recovery (Figure 2.6F), the population can be probed again to measure the resulting distribution of barcodes (Figure 2.6G). The effect of serial bottlenecks can be measured by additional rounds of transient shuffling and exposure to stressors. Other barcoding techniques, such as shotgun cloning, would not allow for the serial tracking of population dynamics because new barcodes would need to be introduced at each stage. Our technique will be advantageous in any situation in which genetic diversity must be introduced at a specific point in time (i.e. lineage tracing), or in cells in which the introduction of genetic material is challenging and/or inefficient.

We have shown that *in vivo* shuffling of a cassette of DNA fragments by a recombinase – Rci – can generate significant diversities for cellular barcoding purposes. Barcoding of individual cells within bacterial or yeast populations should prove to be a useful tool for population geneticists and evolutionary biologists and will allow for a

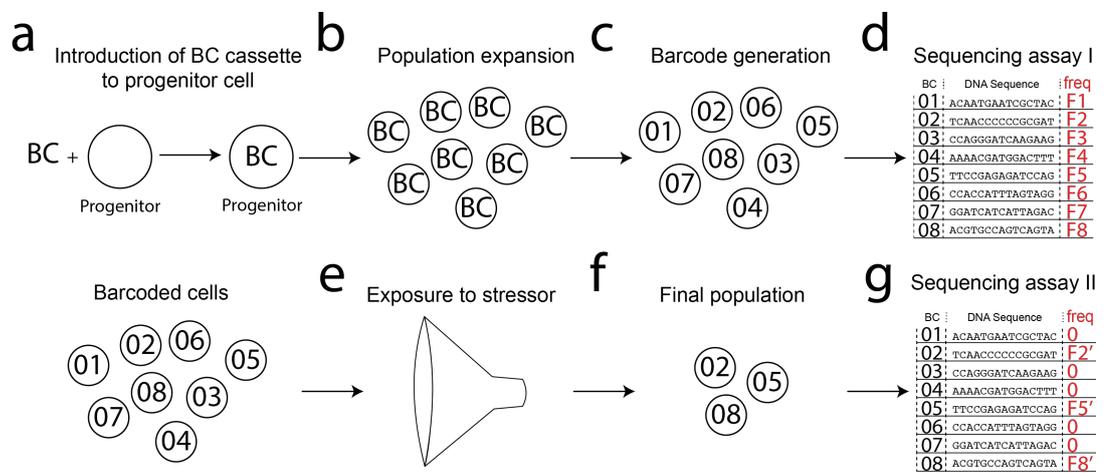


Figure 2.6: Overview of an example *in vivo* barcoding experiment - (a) A barcode cassette is integrated into a progenitor cell. (b) The progenitor cell gives rise to a large population of cells, each harbouring the barcode cassette. (c) Barcodes are generated via transient activity of the recombinase to yield uniquely labelled cells. (d) The barcode pool is then probed by DNA sequencing to test the original distribution of barcodes. (e) The population can be exposed to various stressors resulting in a population bottleneck. (f) This results in a final pool of barcoded cells. (g) The barcode pool is then be assayed again by DNA sequencing to test the final distribution of barcodes.

detailed analysis of population genetics and growth dynamics under various conditions. The system has few moving parts (all that is needed is Rci and a barcode cassette) and is likely to work across a variety of higher organisms with optimization. This system, applied in other organisms could pave the way to the dissection of complex developmental programs, study of heterogeneity within tissues, and/or probing of interactions between cells in a population (i.e. B- and T-cells, neurons, tumours, etc.). *In vivo* barcoding will also pave the way to new explorations in systems biology, from high-throughput monitoring of non-cell-autonomous spread of genetic material to the variability of single-cell transcription profiles. In combination with other molecular tools, *in vivo* barcoding has the potential to provide biologists with unprecedented knowledge of the complex orchestration of single cells within populations.

2.6 Supplementary Material

2.6.1 Supplementary Notes

2.6.1.1 Supplementary Note 1

Simulations (see section 2.3 for details) suggested that the Cre cassettes are subject to considerable biases. Specifically, the ends of barcode fragments are favored for retention (Figure 2.7A). This is explained by the simple observation that there are more combinations of *lox* sites that, when acted on by Cre, will result in the excision of middle fragments (Figure 2.7B). These inherent biases severely limit the practical diversities that can be generated with Cre-based cassettes.

2.6.1.2 Supplementary Note 2

For details on initial experiments employing Rci in mammalian cells, see Appendix A.

2.6.2 Supplementary Figures

Supplementary Figure 1

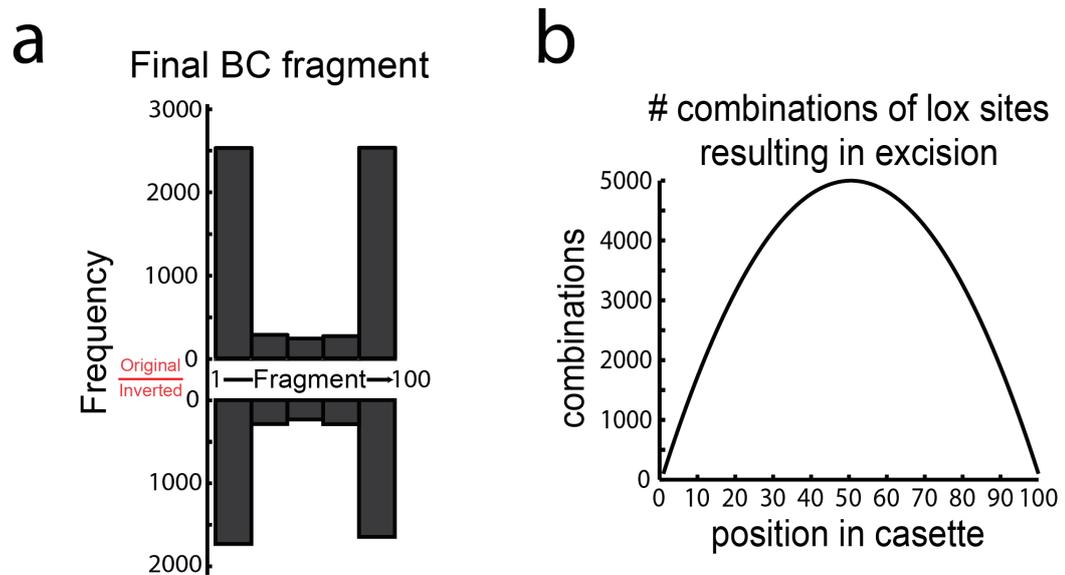


Figure 2.7: Biases of the Cre barcode architecture - (a) Simulated Cre recombination on 10,000 cassettes of length $n = 100$ reveals extreme biases for retaining end fragments. (b) There are many more pairs of *lox* sites that can lead to the excision of more central fragments.

Supplementary Figure 2

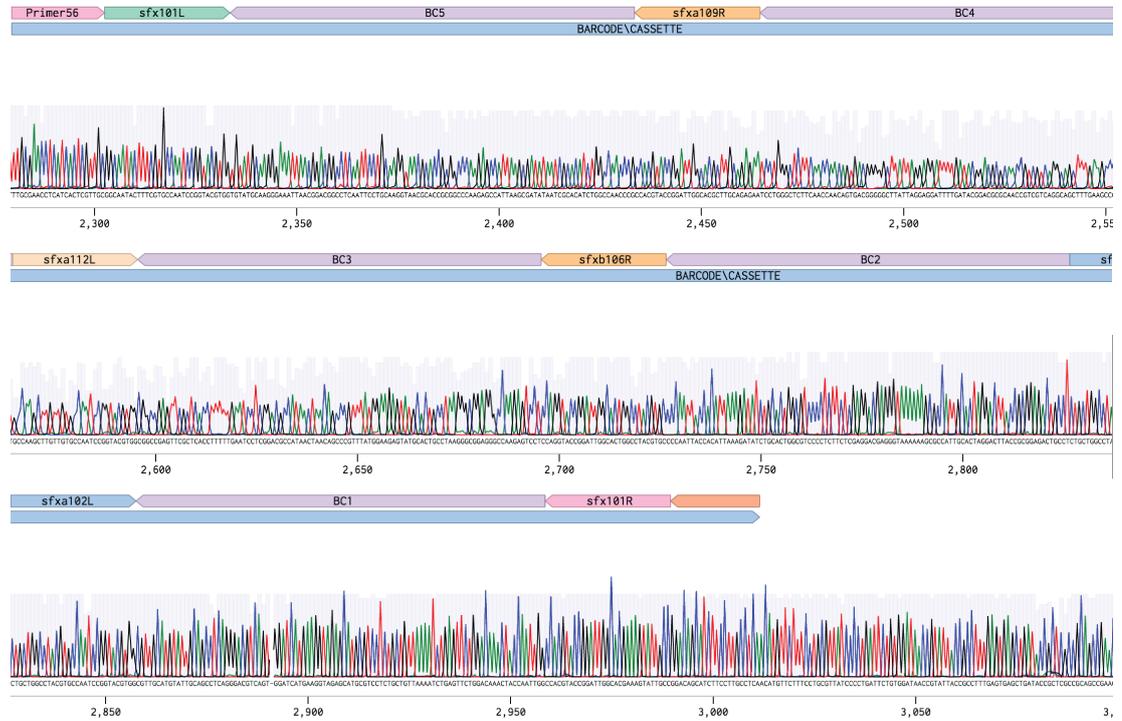


Figure 2.8: 5BC cassette stability during growth - Sanger sequencing of the 5BC cassette after several generations of growth in bacterial cells shows no recombination of the cassette.

Supplementary Figure 3

| position in cassette | | | | | | Likely Recombination Event |
|----------------------|----|----|----|----|------|----------------------------|
| 1 | 2 | 3 | 4 | 5 | freq | |
| -5 | -4 | -3 | -2 | -1 | 1 | <u>sfx101R + sfxa101L</u> |
| -3 | -2 | -1 | 4 | 5 | 1 | <u>sfx101R + sfxa112L</u> |
| 1 | -4 | -3 | -2 | 5 | 1 | <u>sfx102L + sfxa109R</u> |
| 1 | -2 | 3 | 4 | 5 | 1 | <u>sfx102L + sfxb106R</u> |
| 1 | 2 | -5 | -4 | -3 | 1 | <u>sfxb106R + sfx101L</u> |
| 1 | 2 | 3 | -4 | 5 | 3 | <u>sfxa112L + sfxa109R</u> |
| 1 | 2 | 3 | 4 | 5 | 12 | <u>No Recombination</u> |

Figure 2.9: T7 induced Rci expression - All of the reconstructed sequences resulting from shuffling by induced expression of Rci from the T7 promoter can be explained by a single recombination event.

Supplementary Figure 5

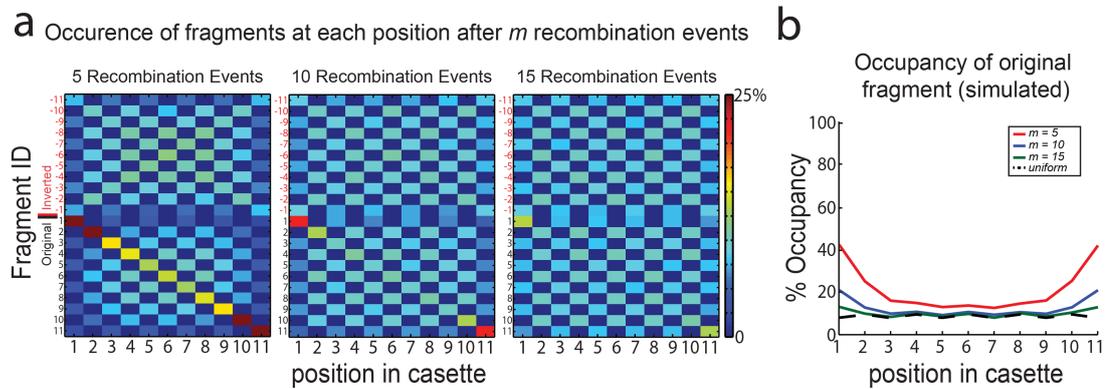


Figure 2.11: Cassettes approach complete randomness as the number of recombination events increase - (a) Simulated cassettes subjected to 5, 10, or 15 random recombination events. The colormaps show the distribution of fragment occupancy at each position in the cassette. Colors are scaled from 0 to 25%. (b) The bias at each position was calculated as the number of times the original fragment appeared in its original position divided by the number of cassettes. The dotted black line indicates the expected occupancy of the original fragment at each position in a cassette with completely random occupancy.

2.6.3 Sequences of relevant genetic elements

2.6.3.1 5BC_Cassette (synthesized by IDT)

GCTTTACCTCGCACTGCCCAGAGTGACATGTTTGCGAACCTCATCACTCG
TTGCGGCAATACTTTCGTGCCAATCCGGTACGTGGTGTATGCAAGGGAAA
TTAACGGACGGCCTCAATTCCCTGCAAGGTAACGCACCGCGGCCAAGAGC
CATTAAAGCGATATAATCGCACATCTGGCCAACCCGCCACGTACCGGATTG
GCACGCTTGACAGAGAATCCTGGGCTCTTCAACCAACAGTGACGGGGGCTT
ATTAGGAGGATTTTGATACGGACGCGCAACCGTTCGTACAGGCAGCTTTGAA
GCCCCGTCTCTCCGGATGCCAAGCTTGTGTGCCAATCCGGTACGTGGCGG
CCGAGTTCGCTCACCTTTTTGAATCCTCGGACGCCATAACTAACAGCCCG
TTTATGGAAGAGTATGCACTGCCTAAGGGCGGAGGGCCAAGAGTCCTCCA
GGTACCGGATTGGCACTGGCCTACGTGCCCAATTACCACATTAAGATA
TCTGCACTGGCGTCCCCTCTTCTCGAGGACGAGGGTAAAAAAGCGCCATT
GCACTAGGACTTACCGCGGAGACTGCCTCTGCTGGCCTACGTGCCAATCC
GGTACGTGGCGTTGCATGTATTGCAGCCTCAGGGACGTCACTGGATCATG
AAGGTAGAGCATGCGTCCCTCTGCTGTTAAAATCTGAGTTCTGGACAACT
ACCAATTGGCCACGTACCGGATTGGCACGAAAGTATTGCCGGACAGCATC
TTCCTTGCCCTCAACATGTCGAACACAGTGGCACGATGCATGAGTCT

2.6.3.2 BCextension (synthesized by GeneWiz)

CGCGCTAGCGGCAATACTTTCGTGCCAATCCGGTACGTGGGATATAGGAT
TGATTTTTCTTCACTGTACGCCGAGAACATAAGAAGCGTCATAACCCTA
TCAGTCCCAAGAATAAGTCCCAGTATTAACGTGTTTCAGCACCCACGTACCG
GATTGGCACGCTTGACAGAGAAATAGGGAAGGCTCATAGCGTTCGAAGTAG
TTCCCAACCAGAACACTCCTGGTGTGAGTCGCTGGAAGCCATCTATAGAAAG
GACGACGCTCGGGCCGGTGTAGCCAAGCTTGTGTGCCAATCCGGTACGT
GGACAACGTCAACGTTAATCTCGTGCATCAGATCCTAGTTGGTTGAGAAC
AGAGCGCGACGTGGGGCTGGGAGCCGATAACCGTAAGTAGGAGAGGGCA
GCCCCAGGTACCGGATTGGCACTGGCCTACGTGCGCCATCGCACCATTAA
ACATAACCGGATCGATATTTAGGTGTGCGGGGCTCCTGCATCAATTAATA
CATCGTTCGAACCAGACCCGTCCATTCATCTCGCCTGCTGGCCTACGTGC
CAATCCGGTACGTGGGCGATTAATCAGCATCCAAGGTTGCACACCGGTTT

AATCGTGGGACATCGTCAAGCGTCAATATCCTAAAGACCTGCGAGGTTTA
GGCATCAGTGCGGGGCCACGTACCGGATTGGCACGAAAGTATTGCCACCT
CGAAAGTGACGATTGTTATTGATCGGTTTGCACCCCGACCAGCCAGGCAT
CTCATCTCCACGGGGATCGAGGGCATCAATCACATTTGACGGATCCGCG

**2.6.3.3 Rci sequence (from NCBI Reference Sequence: NC_013120.1 RE-
GION: complement 7291274066)**

ATGCCGTCTCCACGCATCCGTAAAATGTCCCTGTCACGCGCACTGGATAA
GTACCTGAAAACAGTTTCTGTTCACAAGAAAGGGCATCAACAGGAGTTT
ACCGGAGCAATGTTATCAAGCGATATCCCATTGCTCTTCGGAATATGGAC
GAAATAACAACCGTTGATATTGCTACATACAGAGACGTTTCGTTTAGCAGA
AATAAACCCCCGAACGGGTAAAGCCATTACAGGTAATACTGTACGTCTTG
AACTCGCCCTTCTGTCATCTCTGTTCAATATTGCTCGTGTTGAATGGGGA
ACCTGTCGTAATAACCCGGTTGAACTGGTTCGCAAGCCGAAAGTATCCTC
CGGACGAGATCGCCGGCTAACGTCTTCAGAAGAACGTGCGCTTTCTCGCT
ATTTCCGCGAAAAAATCTGATGTTGTATGTCATTTTCCATCTTGCCCTTG
AAACAGCCATGCGGCAGGGCGAAATACTGGCCTTACGTTGGGAGCACATT
GATTTGCGCCACGGTGTGGCTCATTTACCTGAAACCAAAAACGGTCACTC
ACGGGATGTTCCCTCTGTCCAGACGTGCCCGTAACTTTCTTCAAATGATGC
CCGTTAATCTCCACGGCAATGTTTTTATTACACCGCATCCGGCTTTAAA
AATGCCTGGAGAATAGCCACACAACGACTTCGCATCGAGGACCTGCATTT
TCACGATCTACGGCATGAAGCAATAAGCCGCTTCTTCGAACTGGGTAGCC
TGAATGTAATGGAGATTGCTGCAATATCAGGACATCGTTCCATGAATATG
CTGAAACGGTATACTCATCTTCGTGCATGGCAACTGGTCAGTAAGCTTGA
TGCCCGCCGGCGGCAGACACAAAAGTGGCAGCATGGTTTGTGCCGTATC
CTGCCCATATCACGACTATCGATGAAGAAAATGGGCAGAAAGCGCATCGT
ATTGAGATCGGTGATTTTGATAACCTTCACGTCCTGCCCACAACAAAAGA
GGAAGCAGTTCACCGCGCCAGTGAGGTTTTGTTGCGTACACTGGCCATTG
CAGCACAGAAAGGCGAACGTGTCCCATCTCCCGGAGCGTTACCTGTTAAC
GACCCTGACTACATTATGATTTGCCCTCTGAACCCGGGCAGCACACCGCT
GTAA

2.6.4 Plasmid maps

All plasmids and associated maps are included online at <http://nar.oxfordjournals.org/content/early/2014/07/09/nar.gku604.abstract>, and on Benchling.com: https://benchling.com/ipeikon/ipeikon_published/

- IDP190: 5BC Cassette
- IDP205: T7→Rci; 5BC Cassette
- DIG35: pKat→Rci; 5BC Cassette
- BCextension: 6 fragment barcode extension
- DIG70: T7→Rci; 11BC Cassette
- DIG71: pKat→Rci; 11BC Cassette

2.6.5 Description of Supplementary Files

All files are included online at <http://nar.oxfordjournals.org/content/early/2014/07/09/nar.gku604.abstract>.

- crecode.m : Simulates Cre recombination on a cassette where fragments are separated by two *lox* sites in opposing orientation
- rcicode.m : Simulates Rci recombination on a cassette where fragments are separated by one *sfx* site in alternating orientation
- rcicode2.m : Simulates Rci recombination on a cassette where fragments are separated by two *sfx* sites in opposing orientation
- randDNA.m : Generates random DNA sequences for bootstrapping SW alignment thresholds
- procRCI.m : Processes data from Sanger sequencing to reconstruct barcodes
- procRCI_PB.m : Processes data from PacBio sequencing to reconstruct barcodes

2.6.6 Sequencing data files

All files are included online at <http://nar.oxfordjournals.org/content/early/2014/07/09/nar.gku604.abstract>.

- 5BC_CCS_IDP205.fastq : No Rci expression (T7→Rci; 5BC Cassette)
- 5BC_CCS_DIG35.fastq : Rci expression (pKat→Rci; 5BC Cassette)
- 11BC_CCS_DIG71.fastq : Rci expression (pKat→Rci; 11BC Cassette)

2.7 Acknowledgements

A published version of this chapter appeared in *Nucleic Acids Research* in 2014 [Peikon et al. 2014]. Anthony Zador and I devised the outline of the Rci system. Diana Gizatullina and I performed all of the experiments. Diana was instrumental in the realization of this system. I had extremely useful conversations with Josh Dubnau (CSHL), Alex Koulakov (CSHL), Ton Schumacher (NKI), and Reid Johnson (UCLA) regarding this project.

3

Encoding neural connectivity into DNA

3.1 Abstract

Neural circuit mapping is generally viewed as a problem of microscopy. No current method can achieve high-throughput mapping of circuits with single neuron precision. Here we describe NO-C, a novel approach that transforms connectivity into a form that can be read out by high-throughput sequencing via the tethering of RNA barcodes in synaptic complexes. Improvements of this technique will allow tracing of larger, more complex circuits, or even full brains. A high-throughput method for tracing complete, or even partial, connectivity maps of neural circuits at single neuron resolution could serve as a foundation for neuroscience research.

3.2 Introduction

Neural networks are incredibly sophisticated computational circuits. However, circuit architectures remain largely unknown. Low throughput methods of probing connectivity including electrophysiology have provided strong evidence for structural determinants of neural circuit function [Song et al. 2005]. These efforts have uncovered elements of high-order structure within neural circuits as well as spatially intertwined but non-interconnected networks [Jiang et al. 2013, Yoshimura and Callaway 2005]. While physiological methods have been successful at highlighting circuit connectivity

phenomena, they cannot be easily scaled for the analysis of larger neural circuits or a full nervous system.

High-throughput neural connectivity mapping has traditionally been viewed as a problem of microscopy – electron microscopy (EM) remains the only technique for the direct imaging of synapses. Tracing of processes through EM stacks remains challenging and labor-intensive, but several major efforts are underway to increase the throughput and autonomy of EM [Briggman and Denk 2006, Chklovskii et al. 2010, Kaynig et al. 2013, Kleinfeld et al. 2011]. Complementary methods based on molecular tools aim to make neural circuit tracing amenable to light-microscopy techniques [Feinberg et al. 2008, Kim et al. 2012, Liu et al. 2013, Livet et al. 2007], but these methods are currently limited to sparse connectivity mapping. No current method can achieve high-throughput mapping of the wiring diagram of a neural circuit at single-neuron resolution.

We recently proposed BOINC, a scheme for harnessing the advances of high-throughput sequencing to map neural connectivity [Zador et al. 2012]. Here, we establish one such method and a suite of molecular components for achieving this goal. This method, which we call NO-C – taking specific inspiration from chromosome conformation capture (3C) techniques [Dekker et al. 2002, Kalhor et al. 2012] – relies on the association of cellular RNA barcodes across synaptic partners via tethering of RNA barcodes to complementary pre- and post-synaptic transmembrane proteins (Figure 3.1A). The proteins, in turn, are specifically cross-linked to form a covalent complex that can be immunoprecipitated (Figure 3.1B). The RNA barcode pairs are fused and converted into a single nucleic acid string (Figure 3.1C) and subjected to standard high-throughput sequencing to determine a connectivity matrix (Figure 3.1D).

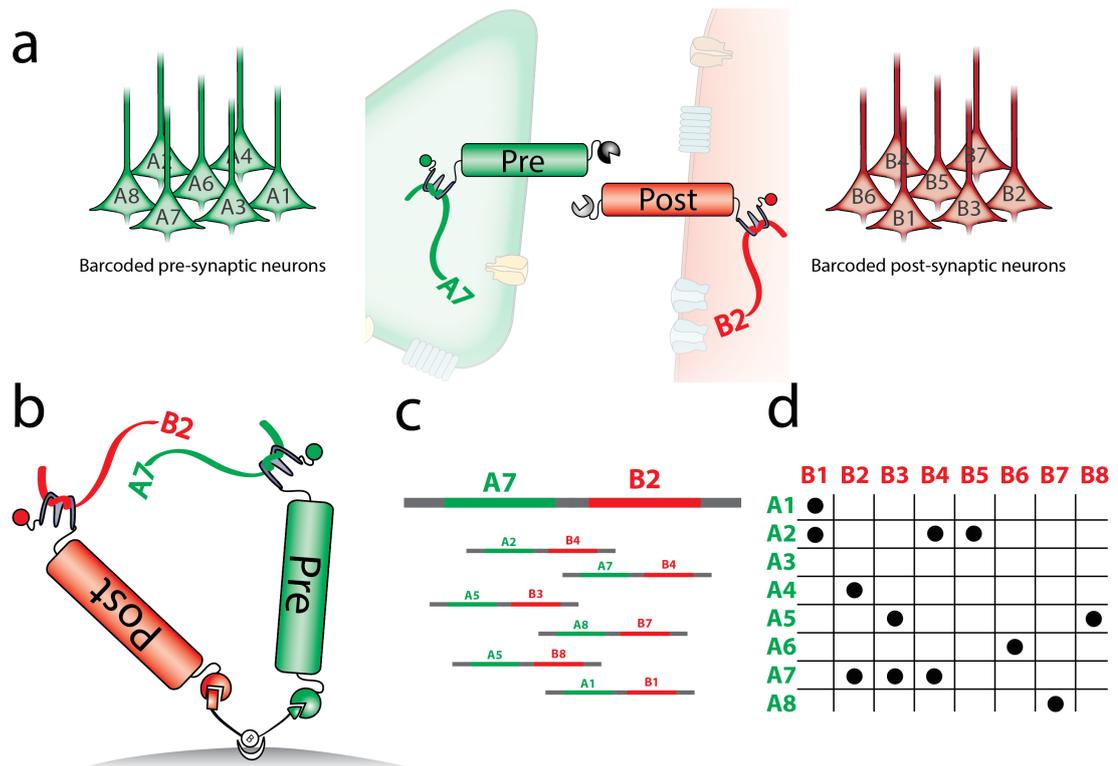


Figure 3.1: Overview of NO-C - (a) A population of pre-synaptic neurons and a population of post-synaptic neurons are separately endowed with an RNA barcode and a modified synaptic protein. The modified synaptic proteins bind the RNA barcodes, thereby tethering barcodes to the pre- or post-synaptic membrane, respectively. The proteins meet at the synapse. (b) The synaptic complex is cross-linked and purified via immunoprecipitation. Each complex contains a pre-synaptic and post-synaptic protein bound to a pre-synaptic and post-synaptic barcode. (c) Barcode pairs representing pairs of connected neurons are joined, amplified, and sequenced. (d) Sequencing data allows the reconstruction of a connectivity matrix.

3.3 Results

We employed the synaptic proteins Neurexin1B (Nrx1B) and Neuroligin1AB (Nlg1AB), which were modified to bind RNA barcodes for tethering at the pre- or post-synapse, respectively. This was accomplished via the fusion of the 22 amino acid RNA binding protein, λ N, to the cytoplasmic domain of Nrx1B and Nlg1AB, and the engineering of RNA barcodes to contain four repeats of the cognate RNA binding motif, the 15 nucleotide boxB hairpin.

To permit tight linking of the barcodes across synapses, we fused CLIP and SNAP to Nrx1B and Nlg1AB, respectively. These self-labeling peptides allow for the specific cross-linking of the two proteins upon addition of a bifunctional (containing reactive moieties for both CLIP and SNAP) small molecule in trans [Gautier et al. 2009]. The small molecule contains domains that allow efficient immunoprecipitation (IP) via biotin, and elution, via a cleavable disulfide bridge (Figure 3.2). To aid in biochemistry and/or visualization, we fused the protein tag Myc to Nrx1B, and HA to Nlg1AB, and introduced coding sequences for GFP and mCherry into the pre-synaptic and post-synaptic barcode transcripts, respectively.

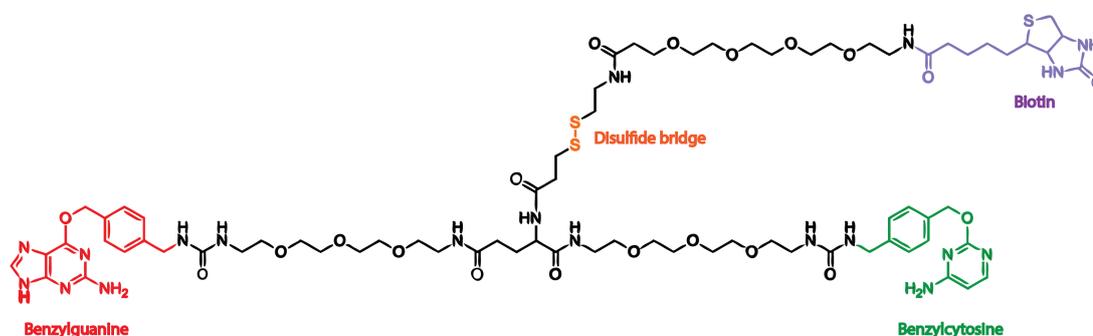


Figure 3.2: BG-PEG-(S-S)-Biotin-PEG-BC cross-linker - This cross-linker is equipped with functional groups BG and BC, which mediate the covalent tagging of SNAP or CLIP respectively. In addition, the molecule contains a biotin moiety for immunoprecipitation, and a cleavable disulfide bridge for non-denaturing elution.

Obtaining proteins that successfully met all of our criteria – bind RNA barcodes specifically and efficiently, traffic to and interact at synapses, and permit specific and efficient covalent cross-linking – required several iterations. For rapid turnaround, the majority of initial debugging was done in human embryonic kidney (HEK) cells. Fusion of CLIP or SNAP alone (to Nr_x1B and Nl_g1AB, respectively) permitted membrane trafficking and SNAP/CLIP tagging at the cell membrane (Figure 3.3A) and efficient cross-linking and immunoprecipitation (Figure 3.3B). Unfortunately, the addition of λ N to the extreme C-terminus of either protein (λ N(c)) prevented proper membrane trafficking. Changing the position of the λ N domain to well-tolerated positions within the C-terminal tails of each protein (λ N(i), see subsection 3.6.3.4) [Dresbach et al. 2004, Fairless et al. 2008, Sara et al. 2005, Slavoff et al. 2011] ameliorated the trafficking issues (Figure 3.3C), but RNA barcode binding (1xBoxB) was poor (Figure 3.3D). The addition of long linkers and four repeats of the λ N [Daigle and Ellenberg 2007] at the same position, permitted proper membrane trafficking (Figure 3.3E) in HEK cells, but expression of these proteins in neurons resulted in poor trafficking (Figure 3.4) and little to no transneuronal interaction as measured by the proximity ligation assay (Figure 3.5, Figure 3.6). We therefore tested a variety of RNA binding domains at different positions in the C-terminal tail until we found a pair of proteins capable of transneuronal interaction (Figure 3.7, Table 3.1, Table 3.2). Ultimately, a single λ N domain, embedded within the C-terminal tail of Nr_x1B or Nl_g1AB, and separated by long linkers on either side, allowed proper trafficking and interaction in neurons. These proteins failed to strongly bind the RNA barcode with 1xBoxB but did bind the RNA barcode with 4xBoxB with high efficiency and specificity, with little to no background (Figure 3.8).

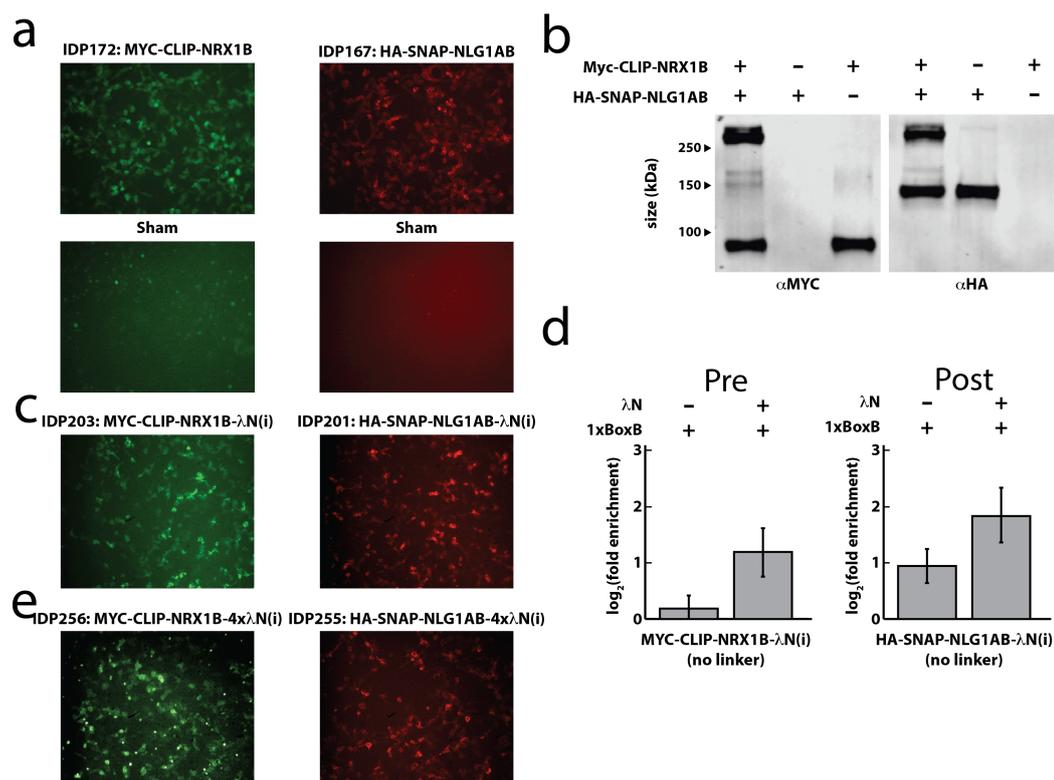


Figure 3.3: Protein design and testing in HEK cells - (a) Fusion of MYC-CLIP to Neurexin or HA-SNAP to Neuroligin permits proper membrane trafficking and CLIP/SNAP tagging in HEK cells. (b) Protein tagging and immunoprecipitation from transiently transfected HEK cells using the cross-linker BG-PEG-Biotin-PEG-BC. (c) Cell-surface tagging of proteins with a single internal λN domain. (d) RNA-IP from transiently transfected HEK cells using the cross-linker BG-PEG-Biotin-PEG-BC. PRE=MYC-Clip-Nrx1B-λN(i). POST=HA-SNAP-Nlg1AB-λN(i). The λN domains were fused directly to the protein without linker sequences. (e) Cell-surface tagging of proteins with and internal 4xλN domain. GREEN=CLIP-488, RED=SNAP-488.

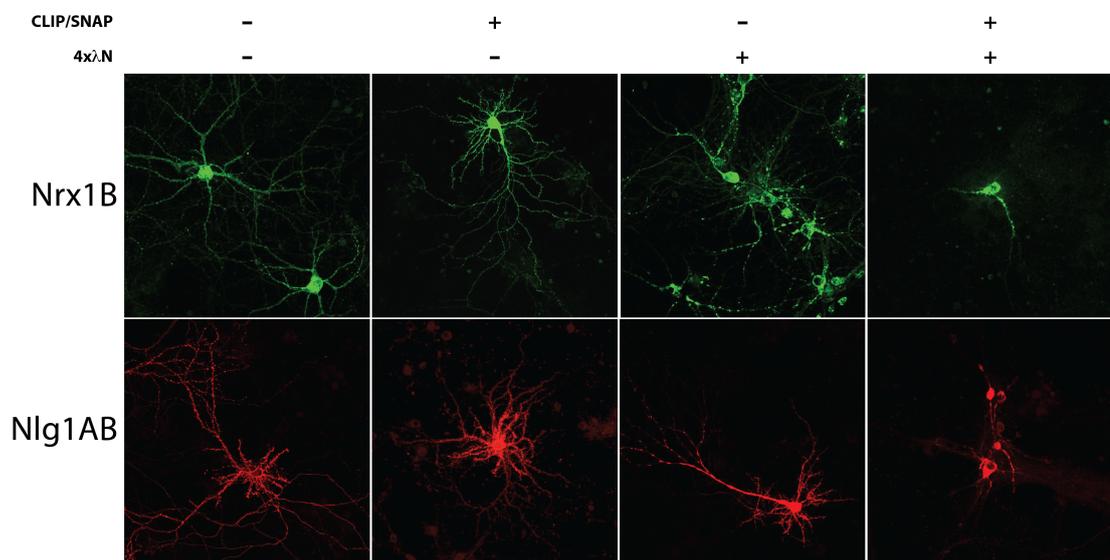


Figure 3.4: Membrane protein tagging in neurons - The addition of CLIP/SNAP domain or the 4xλN domain alone permits trafficking throughout the cell. Simultaneous fusion of both domains prohibits proper trafficking.

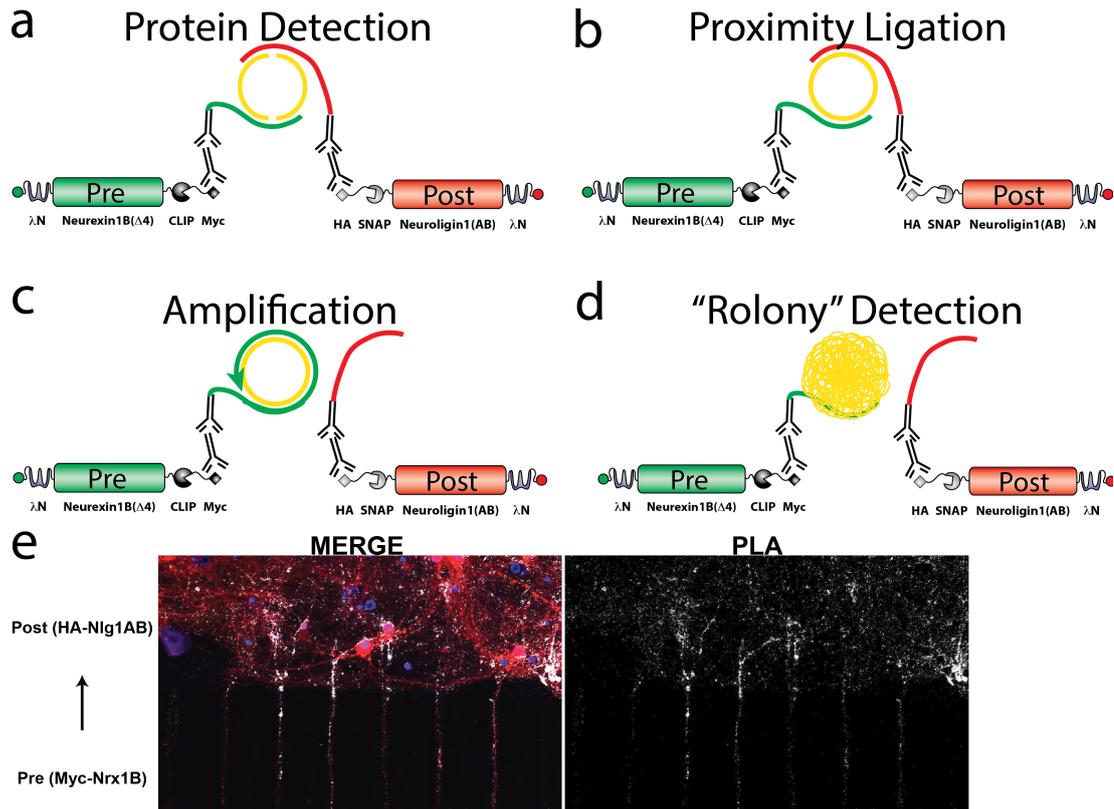


Figure 3.5: Proximity ligation assay for detecting transneuronal interactions -
 (a) Proteins are probed via a primary antibody and a secondary antibody that is covalently linked to an oligonucleotide. If the proteins are in close proximity, the oligonucleotides bridge two additional "test" oligos to form a circle. (b) Ligase seals the nicks in the circle. (c) Rolling circle amplification creates many copies of the DNA. (d) The resulting "rolony" is probed by a fluorescent oligonucleotide with complementarity. (e) This method can be used to probe interactions across neurons, for example detecting the interaction of the synaptic proteins MYC-Nrx1B and HA-Nlg1AB.

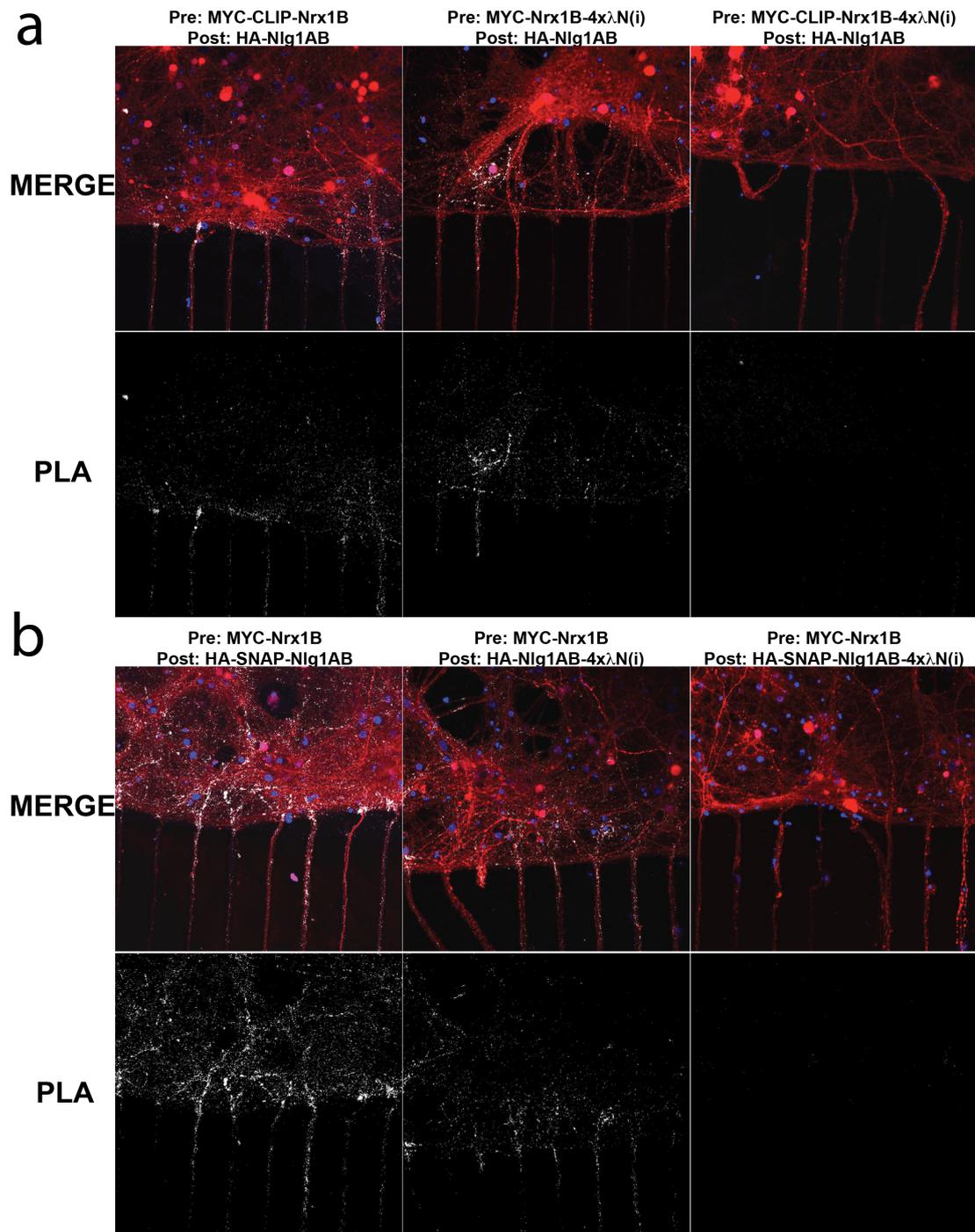


Figure 3.6: Employing PLA to find interacting protein pairs - (a) Pre-synaptic fusion proteins based on MYC-Nrx1B were tested for interaction with the protein HA-Nlg1AB in Xona microfluidic chambers. (b) Post-synaptic proteins based on HA-Nlg1AB were tested for interaction with the protein MYC-Nrx1B in Xona microfluidic chambers. Red=mCherry, White=PLA signal.

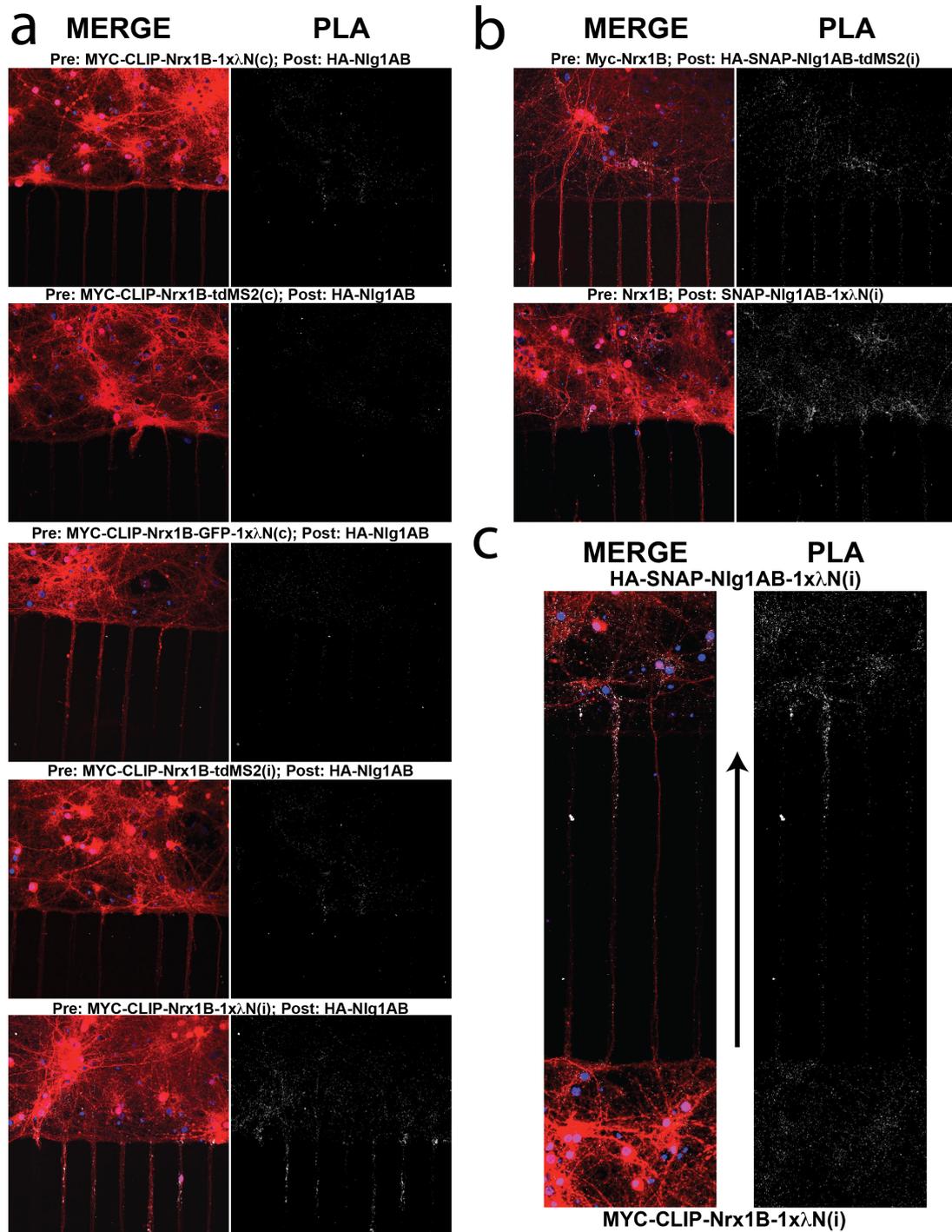


Figure 3.7: Employing PLA to test for compatible RNA binding domains - (a) Various RNA binding domains fused to Myc-CLIP-Nrx1B were tested for their ability to interact with HA-Nlg1AB using PLA. (b) Various RNA binding domains fused to HA-SNAP-Nlg1AB were tested for their ability to interact with Myc-Nrx1B using PLA. (c) Proteins containing a single repeat of the λN domain were tested for interaction with each other. Red=mcherry, Blue=Dapi, White=PLA signal.

Table 3.1: Summary of presynaptic proteins tested by PLA

| Presynaptic Protein | Postsynaptic Protein | PLA Signal? |
|--|-----------------------------|--------------------|
| Myc-Neurexin | HA-Neurologin | YES |
| Myc-CLIP-Neurexin | HA-Neurologin | YES |
| Myc-CLIP-Neurexin-4x λ N(i) | HA-Neurologin | NO |
| Myc-CLIP-Neurexin-tdMS2(i) | HA-Neurologin | NO |
| Myc-CLIP-Neurexin-4x λ N(c) | HA-Neurologin | NO |
| Myc-CLIP-Neurexin-GFP(c)-4x λ N(c) | HA-Neurologin | NO |
| Myc-CLIP-Neurexin-1x λ N(i) | HA-Neurologin | YES |

Table 3.2: Summary of postsynaptic proteins tested by PLA

| Presynaptic Protein | Postsynaptic Protein | PLA Signal? |
|----------------------------|--------------------------------------|--------------------|
| Myc-Neurexin | HA-Neurologin | YES |
| Myc-Neurexin | HA-SNAP-Neurologin | YES |
| Myc-Neurexin | HA-SNAP-Neurologin-4x λ N(i) | NO |
| Myc-Neurexin | HA-SNAP-Neurologin-tdMS2(i) | YES |
| Myc-Neurexin | HA-SNAP-Neurologin-1x λ N(i) | YES |

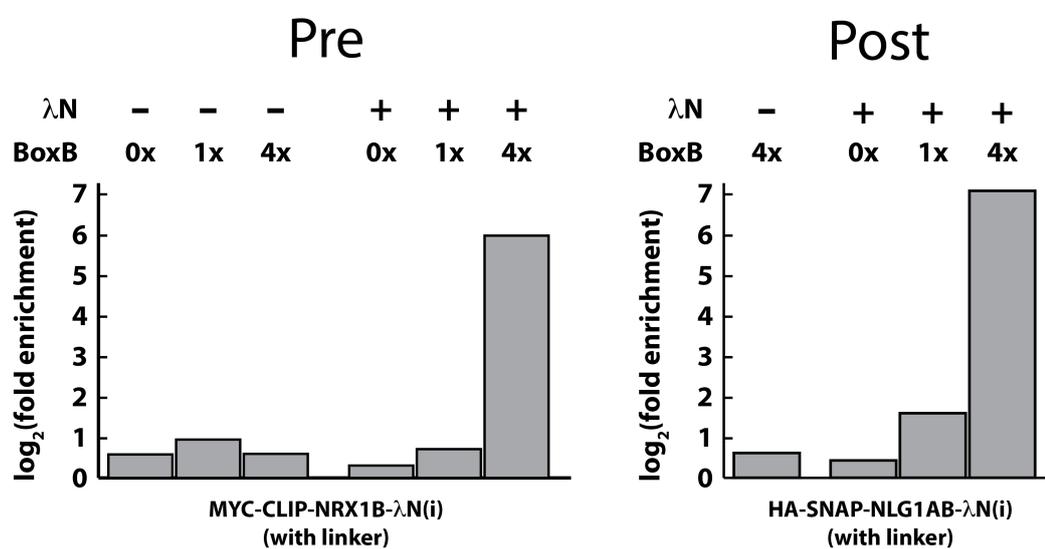


Figure 3.8: RNA-IP with 1xλN fusion proteins - PRE: Myc-CLIP-Nrx1B-λN(i) and POST: HA-SNAP-Nlg1AB-λN proteins were transfected into HEK cells with an RNA barcode containing either 0xBoxB, 1xBoxB, or 4xBoxB. Immunoprecipitation following cell-surface biotin tagging allowed purification of protein-RNA complexes.

3.3 Results

The final proteins, Myc-CLIP-Nrx1B-1x λ N(i) (PRE) and HA-SNAP-Nlg1AB-1x λ N(i) (POST), traffic to the HEK cell membrane (Figure 3.9B), can be joined covalently by addition of the cross-linker before cell-lysis (Figure 3.9C), bind strongly to RNA barcodes with the 4xBoxB motif (Figure 3.9D), and are efficiently eluted under mild reducing conditions (Figure 3.9E).

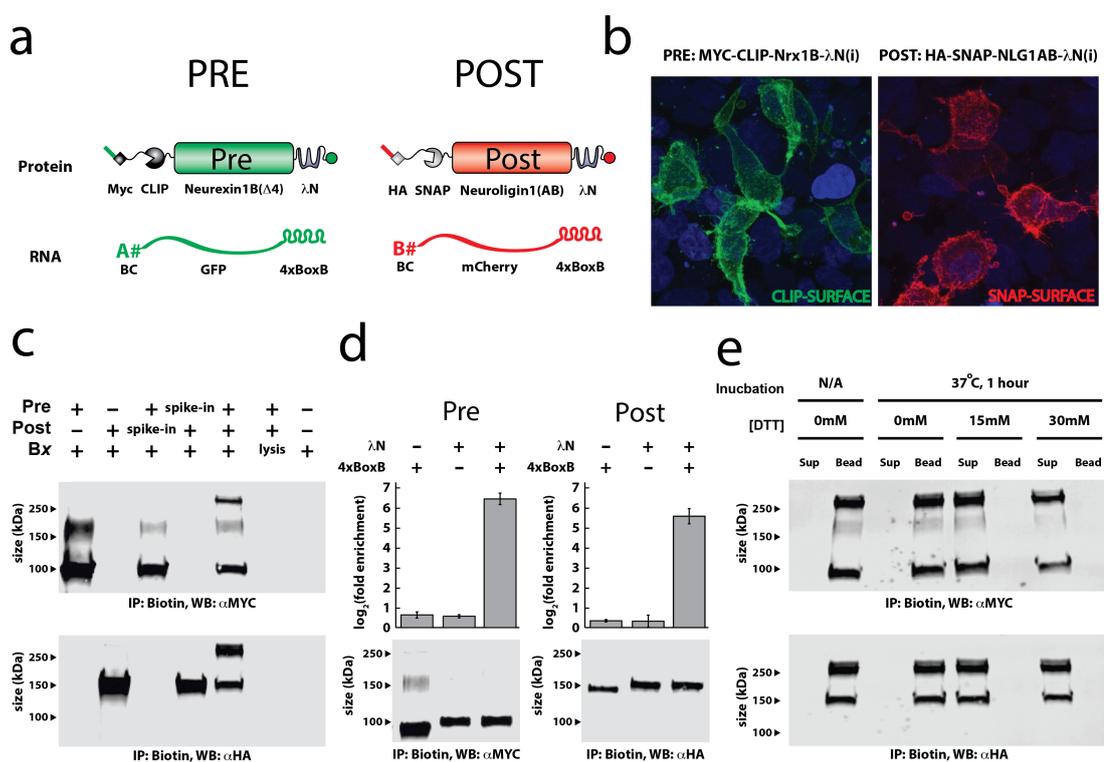


Figure 3.9: Optimization of biochemistry in HEK cells - (a) The pre- and post-synaptic components of the No-C system. (b) Membrane tagging of fusion proteins. Green=CLIP-488, RED=SNAP-488. (c) Immunoprecipitations of proteins from transiently transfected cells, after membrane tagging with BG-PEG-Biotin-PEG-BC. (d) RNA-IP from transiently transfected cells, after membrane tagging with BG-PEG-Biotin-PEG-BC. (e) Cleavage of immunoprecipitated complexes via application of DTT.

After optimizing the biochemistry in HEK cells, we moved to test the system in neurons. First, we confirmed that our protein modifications still permitted proper trafficking of each protein to the cell membrane in neurons (Figure 3.10A,B). To test for the interaction of the proteins between neurons, we grew cultured primary neu-

rons in a dual-chamber microfluidic device [Taylor et al. 2005]. This device allows for the co-culturing of two fluidically isolated populations of cells that are connected via microfluidic grooves that permit only the growth of axons between chambers. We infected one population of cells with barcoded Sindbis virus [Gerlach et al. 2013, Golden et al. 1995, Lu et al. 2011] expressing the pre-synaptic components (Sindbis-PRE) and the other population of cells with barcoded Sindbis virus expressing the post-synaptic components (Sindbis-POST) (Figure 3.10C). Interactions between the two proteins were detected preferentially within the post-synaptic chamber by the proximity ligation assay (PLA, see: Figure 3.10C, Figure 3.5), suggesting preferential trafficking of the proteins to the pre- and post-synaptic compartments, respectively. Using electron microscopy, we confirmed that the PLA signals often occur at synapses (Figure 3.10D).

With these functional properties confirmed, we set out to develop a method for joining RNA barcodes specifically and efficiently. A variety of methods were tested (see Appendix C for details) before settling on emulsion overlap reverse transcription PCR. Briefly, application of the cross-linker forms covalent complexes of the pre- and post-protein, which are in turn tightly bound to barcode RNA. Purified complexes are released from beads, diluted, and emulsified for overlap PCR fusion [DeKosky et al. 2013] (Figure 3.11) using a microfluidic droplet system (Figure 3.12A) such that each droplet contains on average < 1 complex (Figure 3.12B, Figure 3.13).

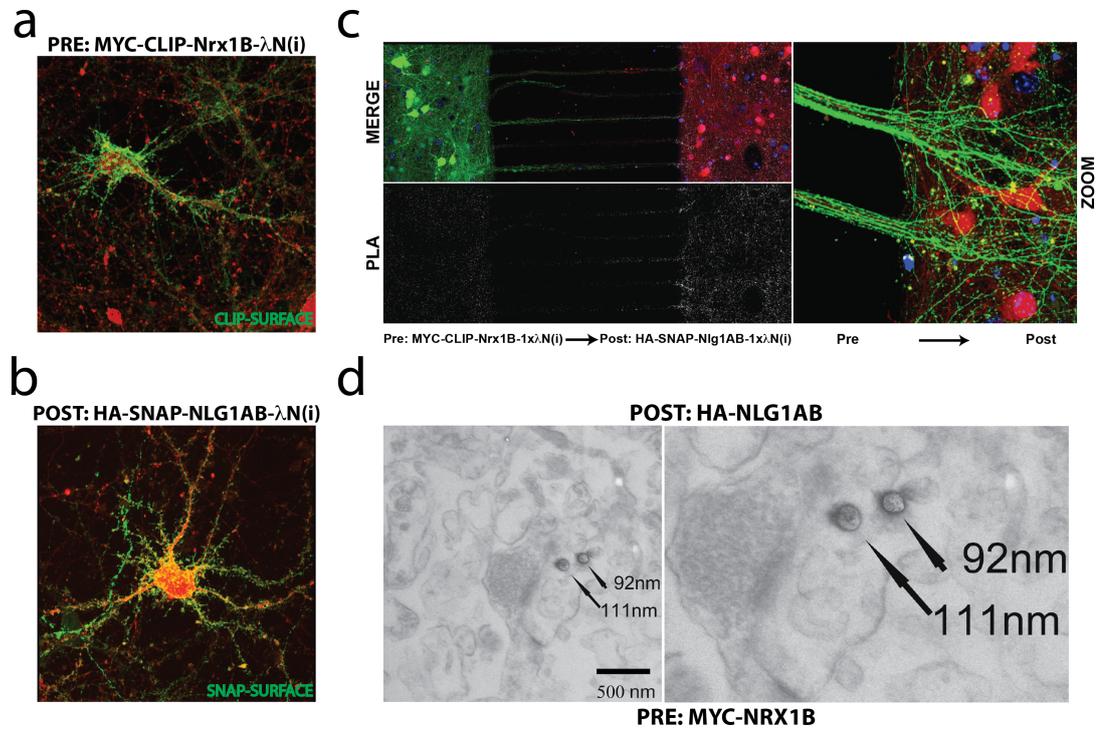


Figure 3.10: Characterization of system components in neurons - (a) Membrane tagging of neurons infected with Sindbis virus expressing Myc-CLIP-Neurexin1B- λ N(i). GREEN=CLIP-488, Red=mCherry. (b) Membrane tagging of neurons infected with Sindbis virus expressing HA-SNAP-Neuroigin1AB- λ N(i). GREEN=SNAP-488, Red=mCherry. (c) Cells grown in a microfluidic chamber system allow the separate infection of two interacting populations of cells. PLA can be used in this system to detect the transneuronal interaction of the pre-synaptic and post-synaptic proteins. GREEN=GFP, RED=mCherry. White=PLA. (d) Transmission electron microscopy of PLA rolonies with incorporated iodo-UTP.

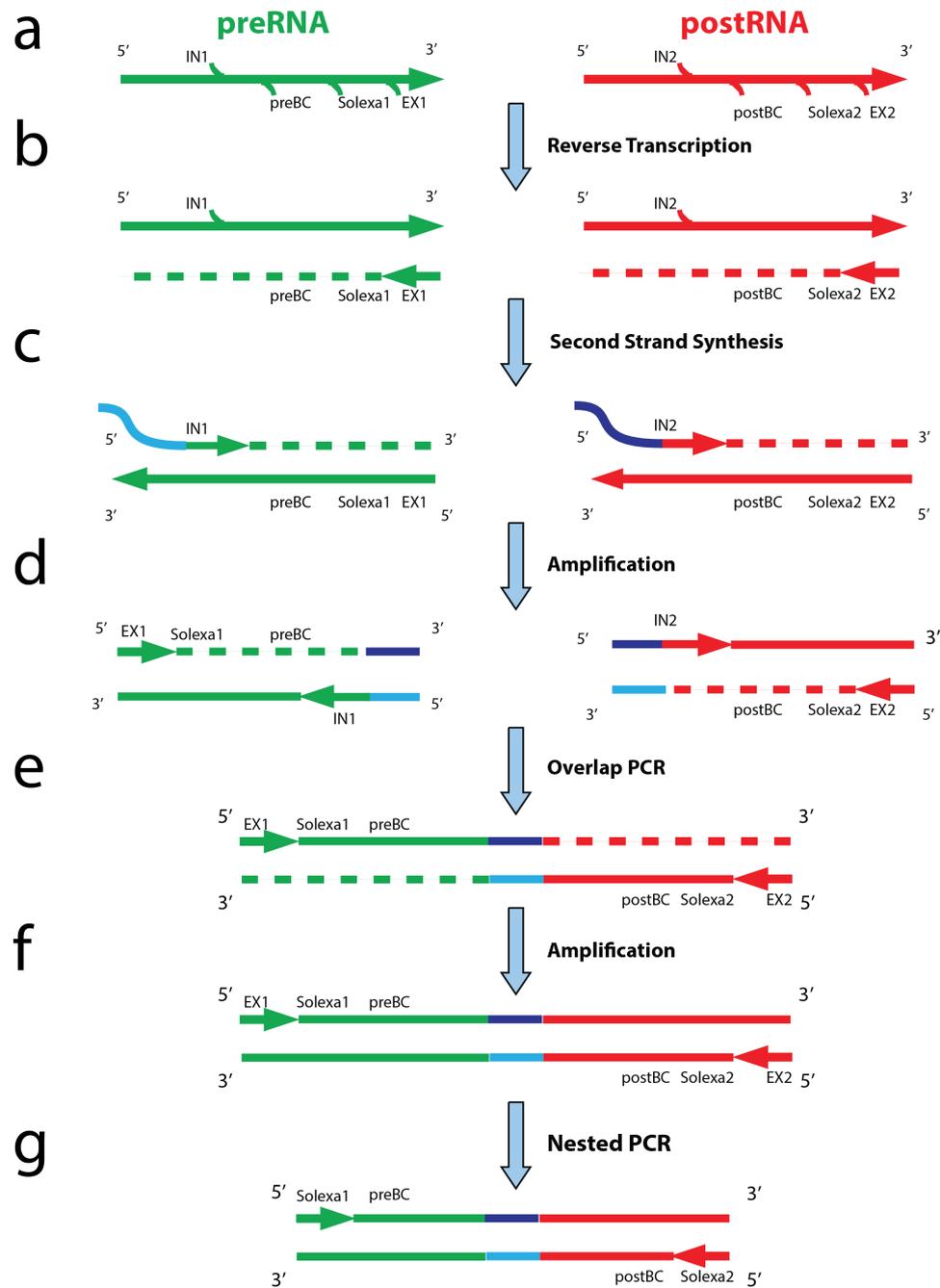


Figure 3.11: Schematic of overlap RT-PCR - (a) Pre-RNA and Post-RNA molecules containing the barcode sequences, as well as Solexa I and Solexa II sequences for high-throughput sequencing. (b) The RNA molecules are reverse transcribed to cDNA. (c) cDNA is subjected to second strand synthesis using a primer that adds a region of homology between the preRNA and postRNA. (d) Barcodes are amplified individually via an external primer and a limiting concentration of internal primers. (e) Internal primers become limiting and the preRNA and postRNA amplicons perform overlapping PCR. (f) The fused barcode pairs are further amplified using the external primers, before being subjected to (g) nested PCR amplification using the sequencing primers.

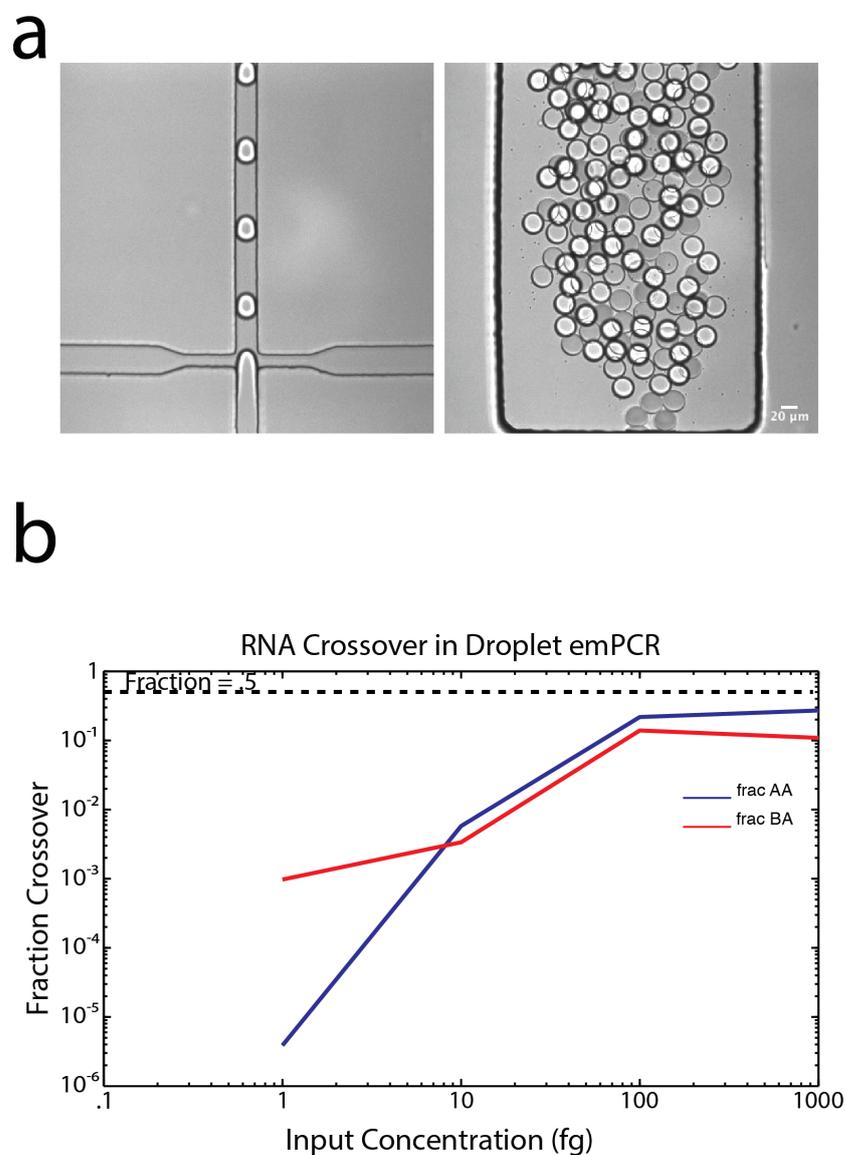


Figure 3.12: Emulsion PCR for barcode joining - (a) Uniform emulsion droplets are formed in a microfluidic device at the interface between the aqueous stream containing PCR mix and the surfactant stream. The emulsion is collected on the microfluidic device before transferring to a PCR tube. (b) Two separate barcode pairs (A-A and B-B) are subjected to uniform emulsions simultaneously to probe the concentrations at which overlap PCR becomes promiscuous. Fraction=.5 is equivalent to an aqueous (non-emulsified) PCR reaction. Here RNA molecules containing both an pre- and post- segment were employed.

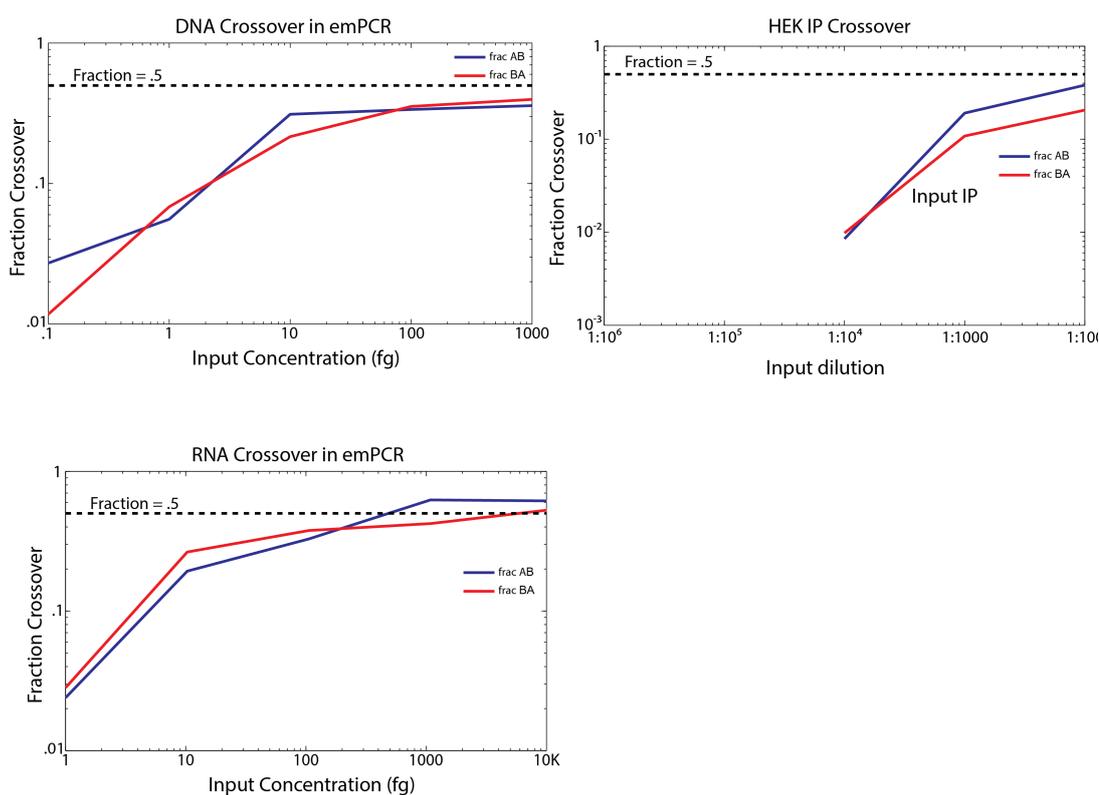


Figure 3.13: Probing emulsion occupancy via qPCR - (a) DNA molecules containing a pre- and post- segment were employed for emulsion PCR. Two separate barcode pairs (A-A and B-B) were emulsified simultaneously and cross-over rates (A-B B-A) were measured via qPCR. (b) RNA was input to the emulsion. (c) Immunoprecipitated complexes from two separate immunoprecipitations (A-A and B-B) were emulsified. Here the emulsions were created using a bench-top vortex and are non-uniform in size.

Sequencing of barcode pairs from complexes formed at the HEK cell membrane gave results which matched the real-time PCR-based quantification. We obtained ~ 100 pre-synaptic barcodes and ~ 100 post-synaptic barcodes, resulting in ~ 1000 barcode pairs. By performing two “zip-coded” experiments in parallel (barcodes are tagged with an additional sequence, a “zip-code” corresponding to the experiment) and mixing the complexes at lysis, we find that there is a low, but significant false-positive rate of $\sim 15\%$ (barcodes from one experiment joined to barcodes of another experiment). This false-positive rate, importantly, only measures false-positives arising from biochemical procedures post-lysis (i.e. protein-RNA dissociation, insufficient dilution for emulsions, etc.). It cannot quantify the false-positive barcode pairs that result from aberrant localization of the proteins or any other *in vivo* effects. The false-positive rate appears to be dominated by the emulsion input concentration (Figure 3.12B, Figure 3.13), and therefore this can be limited by further dilution of the complexes before emulsification.

Although each of the required biochemical components (barcoding, protein-joining, RNA barcode joining) has been established, further work will be needed for tracing a neuronal circuit. First, RNA immunoprecipitations (RNA-IPs) and protein-crosslinking must be validated in neuronal cells. While the HEK cell results show that RNA-IPs are efficient and specific with these proteins, the effect of a different cellular context in neurons must be examined. In addition, the cross-linker (4A long), is significantly shorter than the lower limit of PLA. It remains possible that the cross-linker will need to be extended in order to enable efficient cross-linking in neurons. Cleavage of Neurexin and Neuroligin has been observed *in vivo* [Peixoto et al. 2012, Saura et al. 2011, Suzuki et al. 2012] and this cleavage could be significant if it separates the RNA binding domain from the extracellular CLIP/SNAP domains. Indeed, we have observed that our proteins, when expressed in neurons, are significantly lower in molecular weight than when expressed in other cell types. This correlates with decreased enrichment of the barcode transcripts during RNA-IPs from neurons. Currently, we are testing various methods to limit cleavage of the proteins after *in situ* tagging with the cross-linker. Finally, the expression levels of the proteins from Sindbis must be considered. Over-expression of certain proteins can cause aberrant localization which will result in either false-negative barcode pairs, or worse, false-positives arising from nearby processes of non-synaptic partner cells that express the two-components required for cross-linking.

Further characterization of the proteins and their localization (and importantly their interaction zone) is underway using electron microscopy.

This method should serve as a platform for future development towards encoding neural connectivity, and/or other biological variables – including cellular activity and/or identity – into a form that can be read by high-throughput sequencing, a technology which already operates at the scale of the complexity of neural circuits. Improvements to this method that allow for the co-expression of all of the components within the same cells – for example, by employing two separate RNA binding domains (Figure 3.7B, Table 3.1, Table 3.2) – will allow for bi-directional and local-circuit mapping. Transgenic techniques, combined with an *in vivo* barcoding scheme [Peikon et al. 2014] will allow for scaling this technology to mapping full brains. Spatial positioning of barcoded cells, while lost in the simplest incarnation, can be recovered by a variety of methods [Zador et al. 2012], most notably by performing fluorescence *in situ* sequencing (FISSEQ, see Appendix F) [Lee et al. 2014].

High-throughput methods for mapping neuronal circuits will have a profound impact on neuroscience research. Circuits can be mapped at different developmental time-points, across animals with different genetic abnormalities, after discrete behavioral experiences, and/or after exposure to various pharmacological compounds. By employing DNA sequencing, the field of connectomics may enter an era of “big data” currently being experienced by other fields of biology, which were the first to harness DNA sequencing for deciphering genomes and transcriptomes.

3.4 Materials and Methods

3.4.1 HEK293 cultures

HEK293 cells were grown under standard conditions at 37°C with 5% CO_2 in DMEM supplemented with 10% FBS and 5% Pen-Strep. Cells are cultured in tissue culture plates coated with poly-L-lysine to aid adherence to the plate.

3.4.2 Neuronal cultures

Primary cultures of dissociated mouse hippocampal neurons were prepared from E18 mice. Briefly, brains are isolated and placed into cold HBSS. Hippocampi are isolated and washed 2X with HBSS before being subjected to incubation in HBSS + trypsin + DNase for 15min at 37deg with gentle periodic agitation. Cells are dissociated by pipetting up and down with a pipette and then a fire polished Pasteur pipette. Cells are spun at 1000g for 10min at 4°C, resuspended, counted, and plated at a density of 200,000 cells per well of a 12-well plate. Neurons are cultured in Neurobasal (Life Technologies) + 2% B27 (Life Technologies) + 1% Pen-Strep + 0.5mM L-glutamine. For the first 4 days after dissociation, glutamate is added to the solution at a final concentration of 125 μ M. The cells are incubated at 37°C with 5% CO_2 .

3.4.2.1 Xona microfluidic cultures

Microfluidic chambers (Xona Microfluidics, LLC) are bonded to clean, sterile, and coated (0.5mg/mL poly-D-lysine) by lightly pressing the microfluidic device on the glass. 100,000 dissociated hippocampal neurons are plated on each side of the microfluidic device. Cells are grown under conditions as described above.

3.4.3 Plasmid construction

All plasmids were constructed with Gibson Assembly and the sequences of all final plasmids used in this study are available upon request. Briefly, we performed scarless assembly of up to 6 fragments simultaneously using Gibson assembly. Plasmid backbones containing the EF1a promoter and an SV40 polyA tail were amplified by PCR using primers with complementarity to the inserts. Inserts were prepared by PCR using primers with complementary to the adjacent fragment (either the vector backbone of an additional insert).

3.4.4 Sindbis Production

Sindbis vectors were engineered as aforementioned, using Gibson assembly. Our Sindbis vectors contain either one or two subgenomic promoters, allowing the production of two functional transcripts/proteins in infected cells. To produce virus, $\sim 10\mu\text{g}$ of vector is linearized using the *PacI* site after the subgenomic expression cassette. The linearized plasmid is purified with a phenol:chloroform extraction, followed by a chloroform extraction, followed by an ethanol precipitation overnight at -80°C . The DNA is resuspended to $500\text{ng}/\mu\text{L}$ and $1\mu\text{g}$ is used for *in vitro* transcription from the SP6 promoter using the mMessage mMachine SP6 *in vitro* transcription kit (Life Technologies). The *in vitro* transcribed RNA is directly transfected into BHK cells via lipofection, using Lipofectamine 2000 (Life Technologies) according to the manufacturer's instructions. 24 hours after transfection, the supernatant is isolated and subjected to ultracentrifugation at 30,000rpm for 2hr in an SW50 rotor. The supernatant is removed and the viral pellet is resuspended in $< 200\mu\text{L}$ of the remaining cell culture media.

3.4.4.1 Sindbis titering

All viruses are titered via qPCR. Briefly, a standard was prepared by estimating the number of copies of a Sindbis plasmid from the concentration measured by a spectrophotometer (NanoDrop). After isolation of virus by ultracentrifugation, a $1\mu\text{L}$ aliquot is mixed with RNase1 (Epicenter) according to the manufacturer's instructions (30min at 37°C). The RNase digestion is terminated by the addition of 1mL of Trizol (Life Technologies), and RNA extraction is performed according to the manufacturer's instructions. RNA is precipitated by adding $1\mu\text{L}$ of glycoblue (Life Technologies) and centrifugation for 30min at maximum speed at 4°C . RNA is resuspended in $6\mu\text{L}$ of H_2O , and $2\mu\text{L}$ of RNA are reverse transcribed with a gene-specific primer (j032: 5'-GGGTGCCTTGCTTGAAGTG-3') that primes in the nonstructural protein using SuperScriptIII (Life Technologies) according to the manufacturer's instructions. Serial dilutions of the cDNA (1:10 and 1:100) are prepared and $1\mu\text{L}$ of each is used as the input to a qPCR using primers (j033: 5'-TATCCGCAGTGCGGTTCCAT-3' and j034: 5'-TGTCGCTGAGTCCAGTGTGG-3') that bind to the non-structural proteins. The C_T values are converted to genome copies by comparing to the standard, also subjected to the qPCR procedure.

3.4.5 Barcode Library Production

20ug of plasmid DNA is digested for 2hrs at 37°C with two restriction enzymes that were engineered within the RNA coding sequence (NotI-MluI, New England Biolabs). 10-20U of each restriction enzyme is used per μg of DNA. The distance between the restriction sites is engineered to be ≈ 10 but ≈ 30 nt to obviate the need for gel extraction. Following digestion, the reaction is heat inactivated at 65 or 80°C for 20min (dependent on the restriction enzymes used). The reaction is allowed to cool to room temperature and then 1U/ μg of Alkaline Phosphatase is added and incubated for 1hr at 37°C. The digested DNA is purified via a purification column (Promega Wizard SV PCR Cleanup) according to the manufacturer's instructions.

To prepare the insert, oligonucleotides with a random barcode sequence, flanked by restriction sites are synthesized and PAGE purified. The oligonucleotides are double stranded using a reverse primer and Accuprime Pfx (Invitrogen: Accuprime Pfx Supermix) under the following thermocycling protocol: 95°C for 5min followed by 15 cycles of 95°C for 15sec, 55°C for 30sec, and 68°C for 20sec, followed by a final extension at 68°C for 5min.

After double stranding, the insert is purified via a purification column (Promega Wizard SV PCR Cleanup) according to the manufacturer's instructions. The DNA is then digested at 37°C for 2hrs, and purified again via column.

Ligation is performed at a molar ratio of 1:6 (vector to insert) overnight at 4°C using T4 DNA ligase (New England Biolabs). We typically ligate approximately $1\mu\text{g}$ of DNA to achieve a diversity of 1M barcodes. For different diversities, the ligation reaction can be scaled up or down. The next day, the ligation product is purified with AmpureXP DNA beads (Agencourt, Beckman Coulter) according to the manufacturer's instructions. The DNA is eluted in $40\mu\text{L}$ of H₂O and electroporated into competent *E. coli* (TOP10, Invitrogen), recovered in 50mL SOC for exactly 1hr, and grown overnight in selective LB at 37°C. 1/1000 of the recovered cells are plated for estimation of colony count (i.e. rough estimation of diversity). The next day, DNA is purified via a high-capacity DNA preparation kit (Qiagen Megaprep).

3.5 Tagging with BG/BC derivatives

BG- and BC- functionalized derivatives, including the bifunctional cross-linkers were provided by New England Biolabs (NEB) and used according to the manufacturer's instructions. Briefly, a stock of 1mM tag was prepared by diluting 50nmol of the lyophilized tag in 50 μ L of anhydrous DMSO. Cells expressing cell-surface CLIP- or SNAP-fusion proteins were labeled with 5 μ M tag in DMEM (with 10% FBS). Cells were incubated with the tag for 30min, washed 3X with DMEM, and prepared for imaging and/or immunoprecipitation.

3.5.1 Immunoprecipitation

3.5.1.1 Protein immunoprecipitation

After tagging with biotinylated linkers, cells are lysed in \sim 1mL of lysis buffer PLB3 per 1M cells. PLB3, a previously described lysis buffer [Bottos et al. 2009] contains 20mM TrisHCl, 150mM NaCl, 4mM KCl, 2.5mM MgCl₂, 2.5mM CaCl₂, 10% glycerol, 1% CHAPS, 1% TritonX-100, and protease inhibitors (Roche Complete EDTA-free protease inhibitor). The cells are lysed in the well by pipetting 5-10 times per well and the lysates are incubated on ice for 30 min, pipetted again, and then spun at top speed for 30 min at 4°C to pellet cell debris. A fraction of the lysate is preserved for total protein analysis via western blot or other means. The remainder of the lysate is loaded onto 100 μ L of pre-washed Dynabeads M280 Streptavidin magnetic beads (Life Technologies) per 1M cells and incubated with gentle rotation overnight at 4°C. Beads are washed 6x with PLB3 with ascending salt concentrations (2x with 150mM NaCl, 2x with 300mM NaCl, and 2x with 500mM NaCl).

3.5.1.2 Elution

Protein samples are eluted via addition of 15mM DTT for 1hr at 37°C (if the cleavable cross-linker has been used) or via heating to 95°C in SDS for 5min.

3.5.1.3 Western Blotting

We used antiMYC tag antibody, clone A46 (Millipore) for the detection of the pre-synaptic proteins and antiHA tag antibody, HA.11 (Covance) for the detection of the post-synaptic proteins.

3.5.1.4 RNA immunoprecipitation

RNA-IPs are performed similarly to protein immunoprecipitations (see subsection 3.5.1.1 with slight modifications. For RNA-IP, cells are lysed in PLB3R – PLB3 with the addition of RNase inhibitors (100U/mL lysis buffer RNasin, Promega).

For emulsions, the complexes are eluted from beads under non-denaturing conditions (with DTT) and input into emulsions as described (see subsection 3.5.3.3).

For qPCR, after the 6X washes, the beads are resuspended in 400 μ L Proteinase K buffer (100mM Tris-HCL pH 7.5, 12.5mM EDTA, 150mM NaCl, 1%SDS) with 2 μ L of GlycoBlue and 20 μ L of Proteinase K. The samples are incubated at 65°C for 1 hour with gentle shaking (Eppendor mixer). RNA is extracted with acid phenol:chloroform, then chloroform, and then precipitated overnight in ethanol at -80°C. RNA is pelleted via centrifugation, washed once with 70% EtOH, and then resuspended in 10 μ L of H₂O. The RNA is then subjected to DNase digestion (RQ1 RNase-free DNase, Promega) according to the manufacturer’s instructions. Finally, the RNA is reverse transcribed and the cDNA is used as input for qPCR.

Enrichment in qPCR is calculated by first calculating the ΔC_t values for IP-Total for a housekeeping gene (B-actin) and the barcode containing transcripts – mCherry or GFP. $\Delta\Delta C_t$ is calculated by BC-Actin.

3.5.2 Proximity Ligation Assay

Neurons grown in Xona microfluidic chambers were fixed in 4% PFA in 0.1M PBS for 15 min at 25°C, washed 2X in PBS, and washed 1X in PBS+glycine for 5 min. We then employed the Duolink PLA kit (Sigma) according to the manufacturer’s instructions. For the detection of the pre-synaptic proteins, we used a goat anti-MYC antibody (Abcam ab9132), which recognizes the MYC tag on the extracellular domain of the presynaptic proteins. For detection of the post-synaptic proteins, we used a rabbit anti-HA antibody (abcam9110), which recognizes the HA tag on the extracellular domain of the postsynaptic proteins. Secondary antibodies were obtained from Duolink: MINUS Probe Donkey anti-Rabbit (Sigma DUO92005) and PLUS Probe Donkey anti-Goat (Sigma DUO92003). The Duolink *in situ* far red detection reagents (Sigma DUO92013) were used for detection.

3.5.2.1 Electron Microscopy of PLA signals

Here, we modify the above procedure to allow visualization of the PLA colonies with EM. Neurons grown in Xona microfluidic chambers were fixed in 4% PFA in 0.1M PBS for 15 min at RT, washed 2X in PBS, and washed 1X in PBS+glycine for 5 min. The samples were then blocked in PBS + 1%BSA at 37°C for 1 hour. Samples are then incubated in primary antibodies (1:1000 in PBS+1%BSA) at 25°C for 1 hour, washed 3X for 5 min each in PLA Buffer A (minus Tween20). Samples are then incubated in PLA probes diluted 1:5 in PBS+1%BSA at 37°C for 1 hour, washed 2X for 5 min each in PLA Buffer A (minus Tween20). The samples are then incubated with ligation mix for 30 min at 37°C, washed 2X for 2 min each in PLA Buffer A (minus Tween20). Samples are then incubated with amplification mix containing: .25mg/mL BSA (Life Technologies), 10mM each dATP, dCTP, dGTP (Roche), 5'-iodo-dUTP (Jena Biosciences), .02um compaction oligo (a generous gift from Olaf Ola Sderberg), and .25U/ μ L Phi29 polymerase (NEB) diluted in 1X Phi29 buffer (NEB) in a total volume of 50 μ L at 37°C for 70min. Following amplification, the samples are washed 2X for 10 min each with PLA Buffer B, and washed again for 1 min with 0.01X PLA Buffer B before proceeding to EM fixation.

For EM, cells are fixed by immersion in 3% glutaraldehyde in .1mol/L PB at 4°C overnight (coverslips were fixed in a 60mm petridish). The samples were rinsed 2X in DI water to remove glutaraldehyde and then post-fixed in 1.5% potassium ferrocyanide and 1% glutaraldehyde in DI water for 1 hour. Cells were rinsed once in DI water to remove osmium, dehydrated with a series of increasing % of ethanol, and embedded in 50% acetone-epon araldite resin.

The glass coverslip was removed by placing the block glass slide down in a plastic petri dish containing 1mL of hydrofluoric acid overnight, and the remaining block was rinsed for 10min in a large beaker. The block was dried, mounted on an epoxy stub for micratomoe sectioning. Enface sections were made of the block-face, counter-stained with lead citrate for 1 minute, and imaged with a Hitachi H7500 TEM at 75kv. EM negatives were scanned at 2400dpi with an Epson Perfection V750 pro scanner.

3.5.3 Emulsion Overlap PCR

3.5.3.1 Design and construction of microfluidic device

Microfluidic chips were PDMS/glass hybrids fabricated using soft-lithography as previously described [Brouzes 2012]. Microfluidic channels were treated with a fluorinated tri-chloro silane reagent (heptadecafluoro-1,1,2,2-tetrahydrodecyl)trichlorosilane (Gelest) diluted at 1% weight in FC3280 oil (3M). The microfluidic chips were mounted on an inverted microscope (Motic, AE31) equipped with a Firewire camera (Scout scA640-120fm, Basler), and illuminated by a high power LED (Luxeon) driven by a MOSFET circuit connected to one of the digital output pins of a multifunction data acquisition card (NI PCIe-7841R, National Instruments).

3.5.3.2 Generation of emulsion droplets

Fluids were actuated by a set of pressure controllers (MPV1, Proportion Air), which were controlled by a Labview (National Instruments, TX) application via a microprocessor (Arduino). We generated 13 pL (29 nm diameter) droplets at 1 kHz, using a 15 μm deep x 20 μm wide hydrodynamic focusing nozzle. Droplets were stabilized with a PEG-Krytox based surfactant [Holtze et al. 2008] dissolved at 1% weight in HFE7500 fluorinated oil (3M). Stability through thermocycling was further increased by adding Tetronic 1307 (BASF) at 1% weight to the RT-PCR mixture [Tewhey et al. 2009]. Droplets were collected into 0.2 mL PCR tubes, the bulk oil phase was removed, and the reaction was subjected to thermocycling as described below (subsubsection 3.5.3.3). After completion of the reaction, samples were released by destabilizing the emulsion by addition of one volume of destabilizing agent (Raindance Technologies).

3.5.3.3 One-step overlap reverse-transcription PCR

One-step overlap RT-PCR was performed in emulsions as previously described [DeKosky et al. 2013]. A 100 μL RT-PCR mixture was made containing:

- IP input (variable)
- 25 μL OneStep MasterMix
- 5 μL Primer 759 (10 μM)

- 5 μ L Primer 760 (10 μ M)
- 5 μ L Primer 761 (1 μ M)
- 5 μ L Primer 762 (1 μ M)
- 20 μ L 5% Tetronic
- H₂O to 100 μ L

Briefly, RT was performed for 30 min at 55°C, followed by 2 minutes at 94°C. PCR amplification was performed with the following thermocycling conditions. 1 cycle of: 94°C for 30s, 50°C for 30s, 72°C for 2 min; followed by 4 cycles of: 94°C for 30s, 55°C for 30s, and 72°C for 2 min; followed by 22 cycles of 94°C for 30s, 60°C for 30s, 72°C for 2 min; followed by a final extension step for 7 min at 72°C. After thermal cycling, the emulsion was visually inspected to ensure stability. The emulsions were broken with emulsion destabilizer (Raindance Technologies), and the aqueous phase was collected for subsequent purification. The PCR product was purified using a spin-column clean-up kit (Promega Wizard SV PCR Cleanup Kit).

Table 3.3: Overlap reverse-transcription primers

| Primer ID | Primer Name | Primer Sequence |
|-----------|-------------|---|
| 759 | emPre-RT | 5'-CAGCTCGACCAGGATGGGCA-3' |
| 760 | emPost-RT | 5'-TTCAGCTTGGCGGTCTGGGT-3' |
| 761 | OE-PostF | 5'-TATTCCCATGGCGCGCCGCTGGTTCGGTACGGTAACGGA-3' |
| 762 | OE-PreF | 5'-GGCGCGCCATGGGAATACGGACGATGCCGTCCCTCGTA-3' |

3.5.4 Nested PCR & DNA sequencing

A nested PCR amplification was performed (Accuprime Pfx Supermix) in a total volume of 100 μ L using 10 μ L of eluted cDNA as template with 100nM primers under the following conditions: 2 min 94°C; followed by 30 cycles of 94°C for 30s, 62°C for 30s, and 72°C for 30s; followed by a final extension at 72°C for 7 min. The final product was purified by gel electrophoresis and sequenced using the PE50 protocol on the MiSeq platform (Illumina, San Diego, CA).

3.6 Supplementary Material

3.6.1 RNA design

The pre-synaptic and post-synaptic RNA are based on **GFP** and **mCherry** coding sequences, respectively. In the 5'UTR of the preRNA we placed an anchor sequence (**preHandle**), a 100bp **spacer** sequence, a **NotI** restriction site (for BC cloning), the reverse complement of a **qPCR tag** denoting the library batch (A or B), a 30nt **barcode**, the reverse complement sequence of the Illumina sequences **P5-SBS3T**, and a **MluI** restriction site for barcode cloning. In the 5'UTR of the postRNA we placed an anchor sequence (**postHandle**), a 100bp **spacer** sequence, a **MluI** restriction site (for BC cloning), the reverse complement of a **qPCR tag** denoting the library batch (A or B), a 30nt **barcode**, the reverse complement sequence of the Illumina sequences **P7-SBS8**, and a **NotI** restriction site for barcode cloning. In the 3'UTR of both sequences, we placed 4 repeats of the boxB hairpin motif (**4xBoxB**).

3.6.2 Sequences of RNAs

3.6.2.1 preRNA coding sequence

```
CGGACGATGCCGTCCTCGTAGTTCGGGCATGTACTGGAGCCGA
GAGGTAACCTCATTATAATCGTTCGCTATTCAGGGATTGACCAA
CACCGGAAAACATCTCACTTGAAGTAATATATACGACAGAGTCCG
CGGCCGC A ACCGGTGGACACGTCTTATGTGNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGA
AAGAGTGTAGATCTCGGTGGTCGCCGTATCATTACCGGTGCCAC
CATGGTGAGCAAGGGCGAGGAGCTGTTACCGGGGTGGTGCC
ATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCA
GCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCT
GACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCT
GGCCACCCCTCGTGACCACCCTGACCTACGGCGTGACAGTGCTTC
AGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTC
CGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCA
AGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGA
GGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGAC
```

TTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACA
ACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAG
AACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGG
ACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCC
ATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAG
CACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGAT
CACATGGTCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCT
CGGCATGGACGAGCTGTACAAGTAAAGATCTTACCTAATTGCCG
TCGTGAGGTACGACCACCGCTAGCTGTACAGCCCTGAAAAGG
GCTCGAGCCCTGAAAAGGGCAATTGCCCTGAAAAGGGCGTC
GACGCCCTGAAAAGGGCGGAATTCATGTCCAATTTACTGACTG
TACACCAAATTTGCCTGC

3.6.2.2 postRNA coding sequence

GCTGGTTCGGTACGGTAACGGAGAGTCCGTCCCCTCTTATCCTCG
GCGTTGTGTGTCAAATGGCGTAGATCTGGATTGACTCTATGACG
GTATCTGCTGATCGGTAGGGAGACCGAGAATCTATCGGGCTAA
CGCGTAACCTCCGAAAATGCTGGCACNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNAGATCGGAAGAGCGGTTTCAGCAGGAATG
CCGAGACCGATCTCGTATGCCGTCTTCTGCTTGGCGGCCGCACT
AATCTAGGCCACCATGGTGAGCAAGGGCGAGGAGGATAACATG
GCCATCATCAAGGAGTTCATGCGCTTCAAGGTGCACATGGAGG
GCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGA
GGGCCGCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTG
ACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCC
TCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCCGCCG
ACATCCCCGACTACTTGAAGCTGTCTTCCCCGAGGGCTTCAAG
TGGGAGCGCGTGATGAACTTCGAGGACGGCGGCGTGTTGACCG
TGACCCAGGACTCCTCCCTGCAGGACGGCGAGTTCATCTACAAG
GTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAAT
GCAGAAGAAGACCATGGGCTGGGAGGCCTCCTCCGAGCGGATG
TACCCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGC
TGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAGAC

CACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACA
ACGTCAACATCAAGTTGGACATCACCTCCCACAACGAGGACTAC
ACCATCGTGGAACAGTACGAACGCGCCGAGGGGCCCACTCCA
CCGGCGGCATGGACGAGCTGTACAAGTCTTAGAGATCTTACCTA
ATTGCCGTCGTGAGGTACGACCACCGCTAGCTGTACAGCCCTGA
AAAAGGGCTCGAGCCCTGAAAAAGGGCAATTGCCCTGAAAAAG
GGCGTCGACGCCCTGAAAAAGGGCGGAATTCATGTCCAATTTAC
TGACTGTACACCAAATTTGCCTGC

3.6.3 Protein design

We started with the interacting synaptic proteins – the presynaptic protein Neurexin1B (Uniprot Accession Number: P0DI97, isoform 1b) and the post-synaptic protein Neuroligin1AB (Uniprot Accession Number: Q99K10, isoform 1).

3.6.3.1 Neurexin1B Amino Acid sequence

MYQRMLRCGADLGSPGGSGGGAGGRLALIWIWVPLTSLGLLGVAW
GASSLGAHHIHHFHGSSKHHSVPIAIYRSPASLRGGHAGTTYIFSKG
GGQITYKWPPNDRPSTRADRLAIGFSTVQKEAVLVRVDSSSLGDY
LELHIHQGKIGVKFNVTDDIAIEESNAIINDGKYHVVRFRTRSGGNA
TLQVDSWPVIERYPAGNNDNERLAIARQRIPYRLGRVVDWLLDK
GRQLTIFNSQATIIIGGKEQGQPFQGLSLGLYNGKVLNMAAENDA
NIAIVGNVRLVGEVPSSMTTESTATAMQSEMSTSIMETTTTLATSTA
RRGKPPTKEPISQTTDDILVASAECPSDDEDIDPCEPSSGGLANPTR
VGGREPYPGSAEVIRESSSTTGMVVGIVAAAALCILILLYAMYKYRN
RDEGSYHVDESRNYISNSAQSNQAVVKEKQPSSAKSANKNKNKND
KEYYV*

3.6.3.2 Neuroligin1AB Amino Acid sequence

MALPRCMWPNYVWRAMMACVVHRGSGAPLTLCLLGCLLQTFHV
LSQKLDDVDPLVTTNFGKIRGIKKELNNEILGPVIQFLGVPYAAPPT
GEHRFQPPEPPSPWSDIRNATQFAPVCPQNIIDGRLPEVMLPVWFT
NNLDVVSSYVQDQSEDCLYLNIVPTEDGPLTKKHTDDLGDNDGAE

DEDIRDSGGPKPVMVYIHGGSYMEGTGNLYDGSVLASYGNVIVITV
 NYRLGVLGFLSTGDQAAKGNYGLLDLIQALRWTSENIGFFGGDPLR
 ITVFGSGAGGSCVNLLTLSHYSEGNRWSNSTKGLFQRAIAQSGTALS
 SWAVSFQPAKYARILATKVGCVSDTVELVECLQKKPYKELVDQD
 VQPARYHIAFGPVIDGDVIPDDPQILMEQGEFLNYDIMLGVNQGEG
 LKFVENIVDSDDGVSASDFDFAVSNFVDNLYGYPEGKDVLRETIKF
 MYTDWADRHNPETRRKTLLALFTDHQWVAPAVATADLHSNFGSP
 TYFYAFYHHCQTDQVPAWADAAHGDEVVYVLGIPMIGPTELFPCN
 FSKNDVMLSAAVVMTYWTNFAKTGDPNQVPVQDTKFIHTKPNRFEE
 VAWTRYSQKDQLYLHIGLKPRVKEHYRANKVNLWLELVPHLHNLN
 DISQYTSTTTKVPSTDITLRPTRKNSTPVTSAFPTAKQDDPKQQPSP
 FSVDQRDYSTELSVTIAVGASLLFLNILAFAALYKDKRRHDVHRR
 CSPQRTTTNDLTHAPEEEIMSLQMKHTDLDECESIHPHEVVLRTA
 CPPDYTLAMRRSPDDIPLMTPNTITMIPNTIPGIQPLHTFNTFTGGQ
 NNTLPHPHPHSHSTTRV*

3.6.3.3 Protein-protein interaction

The first problem that needed to be solved, was to engineer a tight interaction, if possible covalent, between the two proteins, specifically at synapses. In addition, we desired an interaction that would not preclude the co-expression of the protein partners in the same cell – a requisite for complete circuit tracing. The reconstitution of GFP results in a tight, albeit non-covalent, complex of two proteins spanning the synaptic cleft. This interaction, better known as GRASP [Feinberg et al. 2008, Kim et al. 2012, Yamagata and Sanes 2012] could be used to associate two synaptic partners in our context. Therefore, we employed the selective cross-linking (SCROSS) system, as described [Gautier et al. 2009]. We fused **MYC-CLIP** and **HA-SNAP** to Nr1B and Nlg1AB, respectively. In both cases, fusion was directly after the signal peptide sequence (after amino acid G46 in Nr1B and K47 in Nlg1AB).

3.6.3.4 RNA-binding

In order to bind RNA, we needed to find positions within the cytoplasmic tails of each protein that would tolerate a fusion protein. We fused **Linker- λ N-Linker** after AA S423 in Nr1B and after T776 in Nlg1AB.

The full coding sequence and AA sequence of each protein are given below. Annotated plasmid maps are available upon request.

3.6.3.5 Coding Sequence: Myc-CLIP-Nrx1B- λ N(i)

ATGTACCAGAGGATGCTCCGGTGCGGCGCCGATCTGGGATCGC
CCGGGGGCGGCAGTGGCGGCGGCGCAGGGGGGCGCCTGGCCCT
GATCTGGATAGTCCCGCTCACCTCGGCGGCCTCCTAGGAGTGG
CCTGGGGGGGAATTCGAACAAAACTCATCTCAGAAGAGGATCT
ACAATTCATGGACAAAGACTGCGAAATGAAGCGCACCCCTG
GATAGCCCTCTGGGCAAGCTGGAAGTGTCTGGGTGCGAACAGG
GCCTGCACCGTATCATCTTCCTGGGCAAAGGAACATCTGCCGCC
GACGCCGTGGAAGTGCTGCCCCAGCCGCGGTGCTGGGCGGAC
CAGAGCCACTGATCCAGGCCACCGCCTGGCTCAACGCCTACTTT
CACCAGCCTGAGGCCATCGAGGAGTTCCCTGTGCCAGCCCTGCA
CCACCCAGTGTTCCAGCAGGAGAGCTTTACCCGCCAGGTGCTGT
GGAAACTGCTGAAAGTGGTGAAGTTCGGAGAGGTCATCAGCGA
GAGCCACCTGGCCGCCCTGGTGGGCAATCCCGCCGCCACCGCC
GCCGTGAACACCGCCCTGGACGGAAATCCCGTGCCCATTTCTGAT
CCCCTGCCACCGGGTGGTGCAGGGCGACAGCGACGTGGGGCCC
TACCTGGGCGGGCTCGCCGTGAAAGAGTGGCTGCTGGCCCACG
AGGGCCACAGACTGGGCAAGCCTGGGCTGGGTGTCGACGGTAC
CGCATCCAGTTTGGGAGCGCACCCACATCCACCATTTCCATGGCA
GCAGCAAGCATCATTGAGTGCCTATTGCAATCTACAGGTCACCA
GCATCCTTGCGAGGCGGACACGCTGGGACGACGTATATCTTTAG
CAAAGGTGGTGGACAAATTACATATAAGTGGCCTCCTAATGACC
GACCCAGTACACGAGCAGACAGGCTGGCCATCGGGTTTAGCAC
TGTTTCAGAAGGAAGCAGTGTGTTGGTGCCTGTGGACAGTTCCTCA
GGCCTGGGTGACTACCTTGAGCTGCACATACACCAAGGAAAAAT
TGGAGTTAAATTTAATGTTGGGACAGATGACATCGCCATTGAGG
AGTCCAATGCAATCATTAAATGATGGAAAATACCATGTAGTGCCT
TTCACGAGGAGTGGTGGCAATGCCACGTTACAGGTGGACAGCT
GGCCAGTCATCGAACGATACCCTGCAGGGCGTCAGCTCACAATC
TTCAATAGCCAAGCAACCATAATAATTGGCGGGAAAGAGCAGG

GCCAGCCCTTCCAGGGCCAGCTCTCTGGGCTTTACTACAATGGC
 TTGAAAGTTCTGAATATGGCGGCAGAGAACGATGCCAACATCG
 CCATAGTGGGAAATGTGAGGCTGGTCCGGTGAAGTGCCTTCCTCT
 ATGACAACCTGAGTCGACAGCCACTGCCATGCAGTCCGAGATGTC
 CACCTCAATCATGGAGACCACCAACCCTGGCTACCAGCACAG
 CTCGACGAGGAAAGCCCCCACAAGGAGCCTATCAGCCAGAC
 CACGGATGATATCCTTGTGGCCTCGGCAGAGTGTCCCAGCGACG
 ATGAGGACATTGACCCCTGTGAGCCGAGCTCAGGTGGGTTAGC
 CAACCCACCCGAGTGGGCGGCCGCGAACCATAACCAGGCTCG
 GCAGAGGTGATCCGGGAGTCTAGCAGTACCACTGGCATGGTGG
 TGGGGATTGTCGCAGCAGCAGCTCTGTGCATCCTCATCCTCCTC
 TATGCCATGTACAAGTACAGGAACCGGGATGAAGGGTTCGAGCG
 CTCCACCGCTCGACGGAGCCGGAGCTGGCGCTGGAGCTGGAGC
 CGGAGCTGGCGGGCTAGCCACC**ATGGACGCACAAACACGACGA**
CGTGAGCGTCGCGCTGAGAAACAAGCTCAATGGAAAGCTGCAA
ACCTCGAGCCACCGCTCGACGGAGCCGGAGCTGGCGCTGGAGC
 TGGAGCCGGAGCTGGCGGTCTAGCCACCCACCGAGCGCTTACC
 ACGTGGATGAGAGTCGAAACTACATCAGTAACTCAGCACAGTCC
 AATGGGGCTGTGGTCAAGGAGATGCAACCCAGCAGTGCTAAAA
 GCGCCAACAAAAACAAGAAGAACAAGGATAAGGAGTATTATGT
 CTAG

3.6.3.6 AA Sequence: Myc-CLIP-Nrx1B- λ N(i)

MYQRMLRCGADLGSPGGGSGGGAGGRLALIWIVPLTLGGLLGVAV
 GEF**EQ**KLISEEDL**QF**MDKDCMKRTTLDSP**LG**KLELSGCE**Q**GLHR**I**
 FLGKGTSAA**DA**VEVPAPAAVLGGPE**LI**QATAWLNAYFHQPEAIEE
 FPVPALHHPVF**Q**QESFTRQVLW**K**LLKVV**K**FGEVISESHLAALVGNP
 AATAAVNTALDGNPVPILIPCHR**V**VQGDS**D**VG**P**YLGGLAVKEWLL
 AHEGHRLG**K**PGLG**VD**GTASS**L**GAHHI**H**HFHGSSKHHSVPIAIYRSPA
 SLRGGHAGTTYIFSKGGGQITYKWPPNDRPSTRADRLAIGFSTVQK
 EAVLVRVDSS**S**GLGDYLELHIHQGKIGVKFN**V**GTDDIAIEESNAI**ND**
 GKYHVVR**F**TRSGGNATLQ**V**DSWPVIERYPAGRQLTIFNSQATI**IG**G
 KEQG**Q**PF**Q**GLSGLYYNGLKVLNMAAENDANIAIVGNVRLVGEVP

SSMTTESTATAMQSEMSTSIMETTTTLATSTARRGKPPTKEPISQTT
DDILVASAECPSDDEDIDPCEPSSGGLANPTRVGGREPYPGSAEVIR
ESSSTTGMVVGIVAAAALCILILLYAMYKYRNRDEGSSAPPLDGAGA
GAGAGAGAGGLATMDAQTRRRERRAEKQAQWKAANLEPPLDGA
GAGAGAGAGAGGLATPPSA YHVDESRNYISNSAQSNGA VVKEMQP
SSAKSANKNKKNKDKEYYV*

3.6.3.7 Coding Sequence: HA-SNAP-Neuroigin1AB-λN(i)

ATGGCACTTCCCAGATGCATGTGGCCAAATTATGTTTGGAGAGC
TATGATGGCATGTGTGGTCCACAGGGGATCCGGTGCCCCATTGA
CTCTCTGCTTGTGGGATGTTTGCTACAGACTTTTCACGTACTC
TCTCAAAGATATCCATACGACGTTCCGGACTACGCAGAATTCAT
GGACAAAGACTGCGAAATGAAGCGCACCACCCTGGATAGCCCT
CTGGGCAAGCTGGAAGTGTCTGGGTGCGAACAGGGCCTGCACC
GTATCATCTTCCTGGGCAAAGGAACATCTGCCGCCGACGCCGTG
GAAGTGCCTGCCCCAGCCGCCGTGCTGGGCGGACCAGAGCCAC
TGATGCAGGCCACCGCCTGGCTCAACGCCTACTTTCACCAGCCT
GAGGCCATCGAGGAGTTCCCTGTGCCAGCCCTGCACCACCCAGT
GTTCCAGCAGGAGAGCTTTACCCGCCAGGTGCTGTGGAAACTG
CTGAAAGTGGTGAAGTTCGGAGAGGTCATCAGCTACAGCCACC
TGGCCGCCCTGGCCGGCAATCCCGCCGCCACCGCCGCCGTGAAA
ACCGCCCTGAGCGGAAATCCCGTGCCCATTTCTGATCCCCTGCCA
CCGGGTGGTGCAGGGCGACCTGGACGTGGGGGGCTACGAGGGC
GGGCTCGCCGTGAAAGAGTGGCTGCTGGCCCACGAGGGCCACA
GACTGGGCAAGCCTGGGCTGGGTGTCGACTTGGATGATGTAGA
CCCATTTGGTTACTACTAACTTTGGCAAGATTAGGGGAATTAAGA
AAGAACTCAATAATGAAATTTTGGGTCTGTCATTCAGTTTCTT
GGGGTTCCATATGCCGCTCCACCAACAGGAGAACATCGTTTCCA
GCCTCCAGAACCACCATCTCCCTGGTCTGACATCCGGAACGCCA
CTCAGTTTGCTCCTGTATGTCCCAGAAATATCATTGATGGCAGA
TTGCCTGAAGTTATGCTTCCTGTGTGGTTCACTAATAACTTGGA
TGTGGTTTCATCATACTCAAGACCAGAGTGAAGACTGTCTAT
ACTTAAACATCTATGTCCCAACTGAAGATGGTCCCCTTACAAAG

AAACACACAGATGATTTAGGTGATAATGACGGTGCTGAAGATG
AAGATATTCGGGACAGTGGGGGTCCCAAACCAGTGATGGTGTA
CATCCATGGCGGCTCTTACATGGAAGGTAAGTGGAAATCTGTATG
ATGGGAGTGTCTTGGCAAGCTATGGCAATGTGATCGTCATCACA
GTCAACTATCGGCTTGGGGTACTTGGCTTCTTGAGCACAGGGGA
TCAGGCTGCCAAAGGAAACTACGGGCTCCTTGACCTCATCCAGG
CCCTAAGATGGACCAGCGAGAACATTGGGTTCTTTGGTGGTGAC
CCCTTGCGAATCACTGTGTTTTGGATCAGGCGCTGGGGGTTTCATG
TGTCAACCTGCTGACTTTATCCCATTATTCTGAAGGTAACCGTT
GGAGCAATTCAACCAAAGGACTTTTTCAACGAGCAATAGCTCAG
AGTGGAACAGCCCTTCCAGCTGGGCTGTTAGTTTCCAGCCTGC
AAAATACGCTAGAATTCTGGCCACAAAAGTTGGCTGCAATGTTT
CAGATACAGTAGAGTTAGTAGAATGCCTGCAGAAGAAGCCTTA
CAAAGAACTTGTTGATCAAGATGTTCAACCAGCCCGATACCACA
TAGCCTTTGGACCTGTGATCGATGGTGATGTAATACCAGATGAC
CCTCAGATACTGATGGAACAAGGAGAGTTCCTCAACTATGATAT
AATGTTGGGAGTTAACCAAGGGGAAGGGTTGAAGTTTGTGAA
AATATAGTAGATAGTGATGATGGTGTATCAGCCAGTGATTTTGA
CTTTGCTGTTTCTAATTTTGTGATAATTTATATGGATATCCTGA
AGGCAAAGATGTTTTGAGAGAAACCATTAAATTCATGTATACTG
ACTGGGCTGATCGCCATAATCCTGAAACTAGAAGGAAGACATTG
TTGGCTTTGTTTACGGACCATCAATGGGTAGCACCTGCTGTGGC
CACAGCAGACCTTCACTCGAACTTTGGCTCACCTACATACTTCT
ATGCCTTTTATCATCATTGCCAAACAGATCAAGTTCCAGCTTGG
GCTGATGCAGCTCATGGGGATGAGGTTCCCTATGTGTTGGGAAT
CCCCATGATTGGCCCTACAGAGTTATTTCCCTTGCAATTTCTCCA
AGAATGATGTGATGTTGAGTGCAGTAGTAATGACATACTGGAC
GAATTTTGCTAAAACCTGGTGACCCAAATCAACCAGTTCCTCAAG
ACACAAAATTCATCCATACCAAACCAACCGCTTTGAAGAAGTA
GCATGGACCAGATATTCAGAGAAAGACCAGCTTTATCTCCATAT
TGGATTAAAACCGAGAGTTAAAGAGCATTACAGAGCCAATAAG
GTAAATCTCTGGCTGGAGCTGGTACCTCATCTGCATAATCTCAA
TGACATTTCTCAGTATACCTCGACAACAACCTAAAGTGCCATCCA

CGGACATCACTCTCAGACCTACAAGGAAGAATTCCACACCAGTC
ACATCAGCCTTTCCCACTGCCAAACAGGATGATCCCAAGCAACA
ACCAAGCCCCTTCTCGGTGGATCAGAGGGACTACTCCACAGAGC
TAAGTGTCACTATCGCAGTGGGGGCCTCTCTGCTGTTTCTCAAC
ATCTTGGCTTTTGCAGCCCTGTACTACAAGAAGGATAAGAGGAG
ACATGATGTCCACCGGAGGTGCAGCCCTCAGCGCACGACCACCA
ACGACCTAACCCATGCTCCAGAAGAGGAAATTATGTCTCTCCAA
ATGAAGCACACTGACTTGGATCACGAGTGTGAGTCCATCCATCC
ACATGAGGTGGTTCTTCGGACCAGCGCTCCACCGCTCGACGGAG
CCGGAGCTGGCGCTGGAGCTGGAGCCGGAGCTGGCGGGCTAGC
CACCATGGACGCACAAACACGACGACGTGAGCGTGCAGCTGAG
AAACAAGCTCAATGGAAAGCTGCAAACCTCGAGCCACCGCTCGA
CGGAGCCGGAGCTGGCGCTGGAGCTGGAGCCGGAGCTGGCGGT
CTAGCCACCCACCGAGCGCTGCCTGTCCCCCAGATTATACTCT
AGCTATGAGGAGGTCACCTGATGATATCCACTAATGACACCTA
ACACCATCACAATGATTCCCAACACTATAACCAGGGATTAGCCC
TTACATACATTCAACACATTTACTGGAGGACAGAATAATACT
GCCCCATCCCCACCCACACCCCCATTACATTCAACAACCAGGG
TATAG

3.6.3.8 AA Sequence: HA-SNAP-Neuroigin1AB- λ N(i)

MALPRCMWPNYVWRAMMACVVHRGSGAPLTLCLLGLLQTFHVL
SQKYPYDVPDYAEFMDKDCMKRTTLDSP LGKLELSGCEQGLHRII
FLGKG TSAADAVEVPAPAAVLGGPEPLMQATAWLNAYFHQPEAIE
EFPVPALHHPVFQQESFTRQVLWKLLKVVKFGEVISYSHLAALAGN
PAATAAVKTALSGNPVPIIPCHR VVQGDLDVGGYEGGLAVKEWL
LAHEGHRLGKPG LGVDLDDVDPLVTTFNGKIRGIKKELNNEILGPV
IQFLGVPYAAPPTGEHRFQPPEPPSPWSDIRNATQFAPVCPQNIIDG
RLPEVMLPVWFTNNLDVSSYVQDQSEDCLYLN IYVPTEDGPLTKK
HTDDLGDNDGA EDEDIRDSGGPKPVMVYIHGGSYMEGTGNLYDGS
VLASYGNVIVITVNYRLGVLGFLSTGDQAAKGN YGLLDLIQALRWT
SENIGFFGGDPLRITVFGSGAGGSCVNLLT LSHYSEGNRWSNSTKGL
FQRAIAQSGTALSSWAVSFQPAKYARILATKVGCNVSDTVELVECL

QKKPYKELVDQDVQPARYHIAFGPVIDGDVIPDDPQILMEQGEFLN
 YDIMLGVNQGEGLKFVENIVDSDDGVSASDFDFAVSNFVDNLYGYP
 EGKDVLRETIKFMYTDWADRHNPEPTRRKTLLALFTDHQWVAPAV
 ATADLHSNFGSPTYFYAFYHHCQTDQVPAWADAAHGDEVVYVLG
 IPMIGPTELFPCNFVSKNDVMLSAVVMYWTNFAKTGDPNQVVPQD
 TKFIHTKPNRFEEVAWTRYSQKDQLYLHIGLKPRVKEHYRANKVN
 LWLELVPHLHNLNDISQYTSTTTKVPSTDITLRPTRKNSTPVTSAFP
 TAKQDDPKQQPSPFSVDQRDYSVELSVTIAV GASLLFLNILAFAALY
 YKKDKRRHDVHRRCSQRRTTNDLTHAPEEEIMSLQMKHTDLDE
 CESIHPHEVVLRTSAPPLDGAGAGAGAGAGAGGLATMDAQTRRRE
 RRAEKQAQWKAANLEPPLDGAGAGAGAGAGAGGLATPPSAACPP
 DYTLMRRSPDDIPLMTPNTITMIPNTIPGIQPLHTFNTFTGGQNNT
 LPHPHPHPHSHSTTRV*

3.6.4 Protein-RNA interaction

The protein-RNA interaction, λ N-boxB binding, was efficient only when 4 repeats of the boxB sequence were employed (Figure 3.8). Initially, we maintained the stoichiometry between the λ N repeats and boxB repeats for simplicity. It may indeed be the case that some of our initial constructs in which a single λ N domain was placed within the cytoplasmic tail of Nr_x1B or Nlg1AB without a flexible linker were sufficient to bind 4xBoxB RNAs. The presence of the linker may be deleterious in some settings and thus, may be removed if necessary (assuming that the 1x λ N does indeed bind 4xboxB when fused directly within the cytoplasmic tail – i.e. no linkers).

3.7 Acknowledgements

Justus Keschull, Diana Gizatullina, and I performed all of the experiments. The majority of the biochemistry was performed by myself and Diana Gizatullina. Justus Keschull performed most of the imaging and protein-protein interaction (PLA) techniques in neurons. Ivan Correa (NEB) synthesized all of the BG- and BC- tags that were not commercially available. Eric Brouzes (SBU) developed the microfluidic droplet system which we used for overlap RT-PCR. Anthony Zador was instrumental in developing the idea for the project, and in designing experiments and analyzing data.

3.7 Acknowledgements

Thanks to my lab mates Hassana Oyibo, Peter Znamenskiy, Qiaojie Xiong, Gang Cao, and Huiqing Zhan who provided lots of helpful advice on this project. And thanks to Ed Boyden, Fei Chen, and Steve Bates for useful conversations and brainstorming sessions. This work started as a collaboration with Dario Bressan and Gregory Hannon – thanks to both of them for their contributions early on. In addition, I had many useful troubleshooting conversations with Josh Dubnau (CSHL), Scott Silverman (Uillinois), and Alex Koulakov (CSHL).

4

Discussion

The wiring of a neural circuit largely influences its function. To date, no high-throughput method of tracing neural circuits exists. We have developed a set of tools for converting neural connectivity into a form that is amenable to high-throughput DNA sequencing, thereby converting the problem from one of imaging to one of molecular engineering. Despite our progress, challenges remain for the full realization of this goal. Here, we outline the shortcomings of this technique and discuss routes to addressing each. Of course, knowledge of the wiring diagram alone may, in many cases, be necessary but insufficient to fully describe the circuit. Additional layers of information including position, cell-type, neuromodulatory state, synaptic strength, and neural activity are also important variables to be measured. In many cases, this information can be encoded in DNA and integrated into the existing framework.

4.1 Scaling to whole brains

In its present form, the technique we have outlined lacks the ability to scale to trace a full connectome. There are several reasons for this. First, viral expression of proteins and/or RNA only allows for sparse tagging of neurons. Transgenic animals which can generate barcodes *in vivo* (see chapter 2) as well as express the synaptic proteins will be required. Second, the system we have developed currently allows only for unidirectional tracing. Third, the proteins we have developed are based on the endogenous proteins Neurexin1B and Neuroligin1AB – which have some nonzero affinity for each other that

may prevent co-expression and proper trafficking. Finally, our RNA joining method is based on emulsions which are inefficient due to the requirement for limiting dilutions.

4.1.1 Transgenics

A transgenic animal expressing all of the components necessary for the system will be required for scaling to a full brain. To avoid any changes during development, the expression of the components should be made inducible (i.e. by employing the Tetracycline inducible promoter). In addition, an *in vivo* barcoding system (see chapter 2) will be necessary. Our current *in vivo* barcoding system has shown promise, but does not yet function efficiently in mammalian cells. Additional *in vivo* barcoding systems based on non-homologous end joining – i.e. by employing CRISPRs [Cong et al. 2013, Mali et al. 2013] to induce DNA damage – will allow for shorter barcodes, but will need to overcome significant biases.

4.1.2 Enabling bi-directional tracing

To enable bidirectional tracing, both the pre- and post-synaptic components must be expressed within the same cell. Moreover, the pre- and post-synaptic barcode must be the same, or the correspondence must be known. Currently, the pre- and post-synaptic barcodes are expressed from two different viruses, and have sequence elements that are unique to pre or post for joining and identification.

In the simplest form, bidirectional tracing could be enabled by employing orthogonal RNA binding proteins on each of the synaptic protein partners (i.e. PRE: λ N and POST:MS2, see Table 3.1, Table 3.2). To endow the cell with two forms of the RNA barcode, alternative splicing could be used to select one of two mutually exclusive exons carrying the pre-specific or post-specific sequence elements.

4.1.3 Preventing endogenous interactions

Although many neurons express both Neurexin1B and Neuroligin1AB, it is unclear if overexpression of the interacting proteins will result in cytoplasmic interaction that prohibits proper sorting to the pre- and post-synaptic terminals respectively. In addition, because of the roles of these proteins in formation and stabilization of synapses, their interaction with endogenous proteins may disrupt proper synapse formation/stability

– most likely resulting in new synapses. For these reasons, it will likely be advantageous to remove the endogenous interaction domains from each protein, which can be accomplished by introducing well-characterized point mutations [Leone et al. 2010].

4.1.4 Efficient RNA joining

Currently, we employ overlap extension RT-PCR in emulsion for RNA barcode joining. This method is efficient in emulsion droplets containing a complex. However, it suffers from the requirement for single occupancy per droplet. Large volumes of emulsion would be needed to scale this system to full brains – something which is feasible but not ideal. Other RNA joining methods (see C) based on proximity ligation or proximity-based amplification will be more efficient but will require extensive optimization. Simply moving from emulsion to overlap extension RT-PCR on a solid surface (similar to bridge amplification techniques performed by Illumina) should greatly improve the ease of scaling the technology.

Employing an alternative RNA polymerase (i.e. PolII or PolIII transcripts) will allow for easier manipulations of the resulting barcodes (no 5'cap, no 3'polyA tail). However, the maximum transcript length that such polymerases can support is not well tested a major consideration for in vivo barcode generation. Moreover, subcellular localization of non pol-II transcripts could pose problems for tethering of RNA barcodes to the synapse and must be tested. Nonetheless, the ability to express short barcodes without 5'cap or 3'polyA tail enables efficient RNA joining methods based on splint ligation.

4.2 Porting to other organisms

In theory, the tools developed here can be modified for use in other organisms including *D. melanogaster* or *C. elegans*. While it may be possible to employ the proteins developed for mouse for tracing in other organisms, a preferable solution is to engineer a species-specific version based on each species' Neurexin and Neuroligin proteins. Of course, each species will come with its own set of issues including, for example, the requirement to remove the cuticle to infuse the synthetic cross-linkers in *C. elegans*.

4.3 Missing information

In its simplest form the wiring diagram lacks many types of useful information including spatial position of the soma of each barcoded cell, the cell-type or expression profile of each cell, the activity of the cell, the potential neuromodulatory states of the circuit, the presence of gap junctions, etc. We would like to recover as much of this information as possible, and be able to overlay this information on the circuit wiring diagram.

4.3.1 Spatial information

It is important to note that, for the purposes of an electrical circuit, the positions of the nodes of the circuit in space (or on the circuit board) – given a wiring diagram – are irrelevant. This is likely true in the mammalian brain as well. For example, a neuron in auditory cortex responds to a sound not because it is located in the auditory cortex, but because it is connected – through a series of synapses – to neurons which sense sound frequencies (i.e. the cochlea). In essence, the spatial position of a cell (i.e. auditory cortex) is a proxy for its connectivity (i.e. receives information from the cochlea), and thus predicts a neurons function (i.e. responsive to sound). In an idealized case, the complete wiring diagram and identity of the nodes obviates the need for spatial information. In practice, however, our wiring diagrams will be incomplete and will, at least initially, lack cellular identity and other important information. Therefore, in order to be able to connect with decades of neuroscience research, it will be advantageous to retain the spatial position of each barcoded soma and – if possible – the position of each synapse.

In simpler organisms, such as *C. elegans* it may be possible to barcode every cell with a “static” barcode that is the same across individuals. This could be accomplished, for example, by employing a cell-type specific promoter (or combination of promoters) to drive the individual barcode transcripts carried on an extrachromosomal array, or integrated into the genome. In higher-organisms this will not be possible. Spatial information could be preserved by virally introducing “anchor” barcodes throughout the brain that correspond to spatial position of the injection. Moreover, the brain can be cubed prior to homogenization and the barcode-pairs in a given region can be “zip-coded” by spatial position.

Finally, by converting the method such that RNA joining and sequencing takes place *in situ* via fluorescent in situ sequencing (FISSEQ) [Lee et al. 2014], the spatial position of each cell and/or synapse can be preserved (see subsection 4.4.1, Appendix F).

4.3.2 Cell types

Knowledge of cell-type is perhaps one of the most important missing pieces of information in the current instantiation of our system. In an electrical circuit, knowledge of the wiring diagram is necessary, but not sufficient to explain the behavior of the circuit. The identity of the individual nodes – capacitor, resistor, transistor, etc. – must be known or able to be inferred from other measurements. Using Cre-driver lines [Taniguchi et al. 2011], the components of our system can be expressed in specific cell-types. While this approach will be useful to understand statistics of connectivity (i.e. how often do SOM cells synapse on PV cells), it will not allow for simultaneous reconstruction of a full circuit containing many heterogenous cell-types. Cell-types could be labeled by different fluorescent proteins, and then registered to the connectome determined by *in situ* sequencing (see subsection 4.4.1).

An ideal solution would provide information about the transcript profile of each barcoded cell in the wiring diagram. To this end, we have explored preliminary solutions for joining barcodes to cellular transcripts, via hijacking of the spliceosome (see E). Other methods, including FISSEQ (see Appendix F, subsection 4.4.1) will allow for sequencing of endogenous RNAs and barcode RNAs in the same preparation.

4.3.3 Cell physiology

Combining NO-C with cellular physiology is non-trivial, but could be theoretically accomplished in a variety of ways. Activity-dependent promoters [Kawashima et al. 2014] including *c-fos* could allow for expression of the individual components in active cells. Alternatively, functional imaging combined with light-activated promoters [Koneremann et al. 2013] would allow for expression in cells of interest as identified by functional properties.

FISSEQ could also be combined with functional imaging. Methods to register the positions and identities of neurons monitored *in vivo* during two-photon calcium

imaging with the same neurons in an acute brain slice have recently been developed [Ko et al. 2011].

4.3.4 Neuromodulation

Neuromodulation allows for drastic changes of neural circuit activity. In simple nervous systems, such as that of *C. elegans*, neuromodulators have been implicated in the multiplexed function of circuits. In higher organisms, such as mice, the extent of neuromodulation is only beginning to emerge. Without the information about neuromodulation, and how it shapes neuronal activity, the wiring diagram alone will be insufficient to describe all neural circuit function [Bargmann 2012, Bargmann and Marder 2013]. Therefore, a way to recover this information would be desirable.

The use of neuromodulatory sensors *in vivo* including glutamate-sensing fluorescent reporter (GluSnFRs) [Hires et al. 2008] or TANGO [Inagaki et al. 2012] will allow for imaging of neuromodulation – perhaps in conjunction with functional imaging. This information can then be incorporated into a wiring diagram uncovered by imaging (i.e. by FISSEQ). Employing a modified version of TANGO which cleaves an exogenous transcription factor (i.e. Gal4) for activation of a specific promoter could allow for the re-encoding of neuromodulatory activity into RNA expression. Of course, this would need to be done in parallel for a variety of neuromodulators, and the transcripts would need to somehow be joined to the cellular barcode. Despite the challenges, it remains at least theoretically plausible that this type of information could be incorporated into the wiring diagram.

4.3.5 Gap Junctions

Besides synaptic connectivity, some neurons communicate through gap junctions. This important mode of information transfer will be missed by the current instantiation of our technique. However, modifications of connexins involved in gap-junction formation could allow for similar barcode tethering/joining to occur at gap-junctions and enable this information to be encoded into DNA for recovery.

4.4 Future directions

4.4.1 *In situ* sequencing

To enable many of the imaging based approaches aforementioned, and to retain spatial information of the cell-body and synapse, we have begun preliminary work to trace connectivity based on a modified version of PLA which relies on FISSEQ for readout. The components of the system are identical, except that CLIP/SNAP are no longer necessary. Briefly, the synaptic proteins drag RNA barcodes to the synapse, forming a transneuronal complex of a barcode-pair spanning the synaptic cleft. The tissue is fixed, sliced, and barcodes are joined *in situ* by proximity ligation. At this point, the joined barcode pairs are sequenced by FISSEQ to determine the identity of each barcode pair. Additional sequencing rounds reveal the somatic position of each individual barcode, and endogenous RNA transcripts. Because all of the sequencing is done *in situ* the spatial position of all of the sequences is preserved, thereby allowing measurement of the spatial position of the cell bodies, synapses, and endogenous transcripts. In addition to enabling the recovery of detailed spatial information, FISSEQ makes it possible to incorporate other imaging-based information into the data-set as discussed above (subsection 4.3.2, subsection 4.3.3, subsection 4.3.4).

4.5 Conclusions

The general principle of encoding synaptic connectivity into a form that can be read out by DNA sequencing can be applied in many other domains of neurobiology. Tethering RNA barcodes to well localized proteins involved in synaptogenesis, synaptic stability, and/or synaptic pruning could allow for a dynamic look at the connectome. For example, tethering barcodes to AMPA receptors may allow for identification of recently potentiated circuits in the full wiring diagram. In every case, the protein of choice will need to be selected wisely, and assays will need to be developed (similar to our PLA assay) to determine proper functioning of the engineered protein. Nevertheless, we believe that the general approach outlined here will be extended to incorporate a host of useful information into the DNA-based connectivity measurement.

References

- Alexander, R. C., Baum, D. A., and Testa, S. M. (2005). 5 transcript replacement in vitro catalyzed by a group I intron-derived ribozyme. *Biochemistry*, 44(21):7796–7804. 127
- Alivisatos, A., Chun, M., Church, G., Greenspan, R., Roukes, M., and Yuste, R. (2012). The brain activity map project and the challenge of functional connectomics. *Neuron*, 74(6):970–974. 10
- Andreas, S., Schwenk, F., Kter-Luks, B., Faust, N., and Khn, R. (2002). Enhanced efficiency through nuclear localization signal fusion on phage c31-integrase: activity comparison with cre and FLPe recombinase in mammalian cells. *Nucleic Acids Research*, 30(11):2299–2306. 100
- Ayre, B. G., Khler, U., Turgeon, R., and Haseloff, J. (2002). Optimization of transsplicing ribozyme efficiency and specificity by in vivo genetic selection. *Nucleic Acids Research*, 30(24):e141–e141. 140
- Bargmann, C. I. (2012). Beyond the connectome: How neuromodulators shape neural circuits. *BioEssays*, 34(6):458–465. 4, 90
- Bargmann, C. I. and Marder, E. (2013). From the connectome to brain function. *Nature Methods*, 10(6):483–490. 90
- Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., Yuen, M. M. S., Keeling, C. I., Brand, D., Vandervalk, B. P., Kirk, H., Pandoh, P., Moore, R. A., Zhao, Y., Mungall, A. J., Jaquish, B., Yanchuk, A., Ritland, C., Boyle, B., Bousquet, J., Ritland, K., MacKay, J., Bohlmann, J., and Jones, S. J. M. (2013). Assembling the 20 gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, 29(12):1492–1497. 33
- Bock, D. D., Lee, W.-C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzell, A. W., Yurgenson, S., Soucy, E. R., Kim, H. S., and Reid, R. C. (2011). Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 471(7337):177–182. 5
- Bohland, J. W., Wu, C., Barbas, H., Bokil, H., Bota, M., Breiter, H. C., Cline, H. T., Doyle, J. C., Freed, P. J., Greenspan, R. J., Haber, S. N., Hawrylycz, M., Herrera, D. G., Hilgetag, C. C., Huang, Z. J., Jones, A., Jones, E. G., Karten, H. J., Kleinfeld, D., Ktetter, R., Lester, H. A., Lin, J. M., Mensh, B. D., Mikula, S., Panksepp, J., Price, J. L., Safdieh, J., Saper, C. B., Schiff, N. D., Schmähmann, J. D., Stillman, B. W., Svoboda, K., Swanson, L. W., Toga, A. W., Van Essen, D. C., Watson, J. D., and Mitra, P. P. (2009). A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. *PLoS Comput Biol*, 5(3):e1000334. 2
- Bottos, A., Destro, E., Rissone, A., Graziano, S., Cordara, G., Assenzio, B., Cera, M. R., Mascia, L., Bussolino, F., and Arese, M. (2009). The synaptic proteins neuroligins and neuroligins are widely expressed in the vascular system and contribute to its functions. *Proceedings of the National Academy of Sciences*, 106(49):20782–20787. 69
- Briggman, K. L. and Denk, W. (2006). Towards neural circuit reconstruction with volume electron microscopy techniques. *Current Opinion in Neurobiology*, 16(5):562–570. 47
- Brouzes, E. (2012). Droplet microfluidics for single-cell analysis. *Methods in Molecular Biology (Clifton, N.J.)*, 853:105–139. 72
- Bullard, D. R. and Bowater, R. P. (2006). Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage t4. *Biochemical Journal*, 398(Pt 1):135–144. 117
- Chklovskii, D. B., Vitaladevuni, S., and Scheffer, L. K. (2010). Semi-automated reconstruction of neural circuits using electron microscopy. *Current opinion in neurobiology*, 20(5):667–675. 5, 47
- Colloms, S. D., Merrick, C. A., Olorunniji, F. J., Stark, W. M., Smith, M. C. M., Osbourn, A., Keasling, J. D., and Rosser, S. J. (2014). Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination. *Nucleic Acids Research*, 42(4):e23–e23. 32, 106
- Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/cas systems. *Science*, 339(6121):819–823. 86, 134
- Daigle, N. and Ellenberg, J. (2007). n-GFP: an RNA reporter system for live-cell imaging. *Nature Methods*, 4(8):633–636. 50, 110
- Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558):1306–1311. 47
- DeKosky, B. J., Ippolito, G. C., Deschner, R. P., Lavinder, J. J., Wine, Y., Rawlings, B. M., Varadarajan, N., Giesecke, C., Drner, T., Andrews, S. F., Wilson, P. C., Hunicke-Smith, S. P., Willson, C. G., Ellington, A. D., and Georgiou, G. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology*. 59, 72
- Douglas, R. J. and Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, 27(1):419–451. 153
- Dresbach, T., Neeb, A., Meyer, G., Gundelfinger, E. D., and Brose, N. (2004). Synaptic targeting of neuroligin is independent of neuroligin and SAP90/PSD95 binding. *Molecular and Cellular Neuroscience*, 27(3):227–235. 50
- Ekstrand, M. I., Enquist, L., and Pomeranz, L. E. (2008). The alpha-herpesviruses: molecular pathfinders in nervous system circuits. *Trends in Molecular Medicine*, 14(3):134–140. 6

REFERENCES

- Elena, S. F. and Lenski, R. E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, 4(6):457–469. 34
- Fairless, R., Masius, H., Rohlmann, A., Heupel, K., Ahmad, M., Reissner, C., Dresbach, T., and Missler, M. (2008). Polarized targeting of neuexins to synapses is regulated by their c-terminal sequences. *The Journal of Neuroscience*, 28(48):12969–12981. 50
- Feinberg, E. H., VanHoven, M. K., Bendesky, A., Wang, G., Fetter, R. D., Shen, K., and Bargmann, C. I. (2008). GFP reconstitution across synaptic partners (GRASP) defines cell contacts and synapses in living nervous systems. *Neuron*, 57(3):353–363. 5, 47, 77
- Feng, L., Lintula, S., Ho, T. H., Anastasina, M., Paju, A., Haglund, C., Stenman, U.-H., Hotakainen, K., Orpana, A., Kainov, D., and Stenman, J. (2012). Technique for strand-specific gene-expression analysis and monitoring of primer-independent cDNA synthesis in reverse transcription. *BioTechniques*, 52(4):263–270. 130, 131
- Gautier, A., Nakata, E., Lukinavicius, G., Tan, K.-T., and Johnsson, K. (2009). Selective cross-linking of interacting proteins using self-labeling tags. *Journal of the American Chemical Society*, 131(49):17954–17962. 49, 77
- Gerlach, C., Rohr, J. C., Peri, L., Rooij, N. v., Heijst, J. W. J. v., Velds, A., Urbanus, J., Naik, S. H., Jacobs, H., Beltman, J. B., Boer, R. J. d., and Schumacher, T. N. M. (2013). Heterogeneous differentiation patterns of individual CD8+ t cells. *Science*, 340(6132):635–639. 14, 59, 154
- Geschwind, D. H. and Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology*, 17(1):103–111. 12
- Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, 6(5):343–345. 130
- Golden, J. A., Fields-Berry, S. C., and Cepko, C. L. (1995). Construction and characterization of a highly complex retroviral library for lineage analysis. *Proceedings of the National Academy of Sciences*, 92(12):5704–5708. 14, 59, 154
- Gyohda, A., Furuya, N., Kogure, N., and Komano, T. (2002). Sequence-specific and non-specific binding of the rci protein to the asymmetric recombination sites of the r64 shufflon. *Journal of Molecular Biology*, 318(4):975–983. 23, 102, 106
- Gyohda, A. and Komano, T. (2000). Purification and characterization of the r64 shufflon-specific recombinase. *Journal of Bacteriology*, 182(10):2787–2792. 23
- Gyohda, A., Zhu, S., Furuya, N., and Komano, T. (2006). Asymmetry of shufflon-specific recombination sites in plasmid r64 inhibits recombination between direct sfx sequences. *Journal of Biological Chemistry*, 281(30):20772–20779. 102, 103, 106
- Herschlag, D. (1991). Implications of ribozyme kinetics for targeting the cleavage of specific RNA molecules in vivo: more isn't always better. *Proceedings of the National Academy of Sciences*, 88(16):6921–6925. 143
- Hill, S. L., Wang, Y., Riachi, I., Schrmann, F., and Markram, H. (2012). Statistical connectivity provides a sufficient foundation for specific functional connectivity in neocortical neural microcircuits. *Proceedings of the National Academy of Sciences*, 109(42):E2885–E2894. 3
- Hires, S. A., Zhu, Y., and Tsien, R. Y. (2008). Optical measurement of synaptic glutamate spillover and reuptake by linker optimized glutamate-sensitive fluorescent reporters. *Proceedings of the National Academy of Sciences*, 105(11):4411–4416. 90
- Hoess, R., Wierzbicki, A., and Abremski, K. (1985). Formation of small circular DNA molecules via an in vitro site-specific recombination system. *Gene*, 40(2-3):325–329. 22
- Holtze, C., Rowat, A. C., Agresti, J. J., Hutchison, J. B., Angil, F. E., Schmitz, C. H. J., Kster, S., Duan, H., Humphry, K. J., Scanga, R. A., Johnson, J. S., Pisignano, D., and Weitz, D. A. (2008). Biocompatible surfactants for water-in-fluorocarbon emulsions. *Lab on a Chip*, 8(10):1632–1639. 72
- Hooks, B. M., Mao, T., Gutnisky, D. A., Yamawaki, N., Svoboda, K., and Shepherd, G. M. G. (2013). Organization of cortical and thalamic input to pyramidal neurons in mouse motor cortex. *The Journal of Neuroscience*, 33(2):748–760. 153
- Illumina (2014). Performing primer rehybridization on the cBot. 34
- Inagaki, H. K., Ben-Tabou de Leon, S., Wong, A. M., Jagadish, S., Ishimoto, H., Barnea, G., Kitamoto, T., Axel, R., and Anderson, D. J. (2012). Visualizing neuromodulation in vivo: TANGO-mapping of dopamine signaling reveals appetite control of sugar sensing. *Cell*, 148(3):583–595. 90
- Inoue, H., Hayase, Y., Iwai, S., and Ohtsuka, E. (1987). Sequence-dependent hydrolysis of RNA using modified oligonucleotide splints and RNase h. *FEBS Letters*, 215(2):327–330. 120
- Jan, C. H., Friedman, R. C., Ruby, J. G., and Bartel, D. P. (2011). Formation, regulation and evolution of caenorhabditis elegans 3UTRs. *Nature*, 469(7328):97–101. 117
- Jenett, A., Rubin, G. M., Ngo, T.-T. B., Shepherd, D., Murphy, C., Dionne, H., Pfeiffer, B. D., Cavallaro, A., Hall, D., Jeter, J., Iyer, N., Fetter, D., Hausenfluck, J. H., Peng, H., Trautman, E. T., Svirskas, R. R., Myers, E. W., Iwinski, Z. R., Aso, Y., DePasquale, G. M., Enos, A., Hulamm, P., Lam, S. C. B., Li, H.-H., Laverty, T. R., Long, F., Qu, L., Murphy, S. D., Rokicki, K., Safford, T., Shaw, K., Simpson, J. H., Sowell, A., Tae, S., Yu, Y., and Zugates, C. T. (2012). A GAL4-driver line resource for drosophila neurobiology. *Cell Reports*, 2(4):991–1001. 133
- Jiang, X., Wang, G., Lee, A. J., Stornetta, R. L., and Zhu, J. J. (2013). The organization of two new cortical interneuronal circuits. *Nature Neuroscience*, 16(2):210–218. 46
- Johnson, R. C. (2002). Bacterial site-specific DNA inversion systems. *Mobile DNA II. ASM Press, Washington, DC*, pages 230–271. 23, 103
- Johnson, R. C., Ball, C. A., Pfeiffer, D., and Simon, M. I. (1988). Isolation of the gene encoding the hin recombinational enhancer binding protein. *Proceedings of the National Academy of Sciences*, 85(10):3484–3488. 23

REFERENCES

- Jones, J. T., Lee, S. W., and Sullenger, B. A. (1996). Tagging ribozyme reaction sites to follow trans-splicing in mammalian cells. *Nature medicine*, 2(6):643–648. 140
- Jones, J. T., Lee, S.-W., and Sullenger, B. A. (1997). Trans-splicing reactions by ribozymes. In Turner, P. C., editor, *Ribozyme Protocols*, number 74 in *Methods in Molecular Biology*, pages 341–348. Humana Press. 145
- Jones, J. T. and Sullenger, B. A. (1997). Evaluating and enhancing ribozyme reaction efficiency in mammalian cells. *Nature Biotechnology*, 15(9):902–905. 140
- Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1):90–98. 47, 115, 116
- Kawashima, T., Okuno, H., and Bito, H. (2014). A new era for functional labeling of neurons: activity-dependent promoters have come of age. *Frontiers in Neural Circuits*, 8. 89
- Kaynig, V., Vazquez-Reina, A., Knowles-Barley, S., Roberts, M., Jones, T. R., Kasthuri, N., Miller, E., Lichtman, J., and Pfister, H. (2013). Large-scale automatic reconstruction of neuronal processes from electron microscopy images. *arXiv:1303.7186 [cs, q-bio]*. arXiv: 1303.7186. 47
- Keravala, A., Lee, S., Thyagarajan, B., Olivares, E. C., Gabrovsky, V. E., Woodard, L. E., and Calos, M. P. (2008). Mutational derivatives of PhiC31 integrase with increased efficiency and specificity. *Molecular Therapy*, 17(1):112–120. 102
- Kikumori, T., Cote, G. J., and Gagel, R. F. (2001). Promiscuity of pre-mRNA spliceosome-mediated trans splicing: A problem for gene therapy? *Human Gene Therapy*, 12(11):1429–1441. 146
- Kim, J., Zhao, T., Petralia, R. S., Yu, Y., Peng, H., Myers, E., and Magee, J. C. (2012). mGRASP enables mapping mammalian synaptic connectivity with light microscopy. *Nature Methods*, 9(1):96–102. 5, 47, 77
- Kita, T. and Kita, H. (2012). The subthalamic nucleus is one of multiple innervation sites for long-range corticofugal axons: A single-axon tracing study in the rat. *The Journal of Neuroscience*, 32(17):5990–5999. 109
- Kleinfeld, D., Bhariok, A., Blinder, P., Bock, D. D., Briggman, K. L., Chklovskii, D. B., Denk, W., Helmstaedter, M., Kaufhold, J. P., Lee, W.-C. A., Meyer, H. S., Micheva, K. D., Oberlaender, M., Prohaska, S., Reid, R. C., Smith, S. J., Takemura, S., Tsai, P. S., and Sakmann, B. (2011). Large-scale automated histology in the pursuit of connectomes. *The Journal of Neuroscience*, 31(45):16125–16138. 47
- Ko, H., Hofer, S. B., Pichler, B., Buchanan, K. A., Sjström, P. J., and Mrsic-Flogel, T. D. (2011). Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91. 90
- Kohler, U., Ayre, B. G., Goodman, H. M., and Haseloff, J. (1999). Trans-splicing ribozymes for targeted gene delivery. *Journal of Molecular Biology*, 285(5):1935–1950. 143
- Konermann, S., Brigham, M. D., Trevino, A., Hsu, P. D., Heidenreich, M., Le Cong, Platt, R. J., Scott, D. A., Church, G. M., and Zhang, F. (2013). Optical control of mammalian endogenous transcription and epigenetic states. *Nature*, advance online publication. 89
- Kording, K. P. (2011). Of toasters and molecular ticker tapes. *PLoS Comput Biol*, 7(12):e1002291. 10
- Koulakov, A. A., Hromdka, T., and Zador, A. M. (2009). Correlated connectivity and the distribution of firing rates in the neocortex. *The Journal of Neuroscience*, 29(12):3685–3694. 3
- Kubo, A., Kusukawa, A., and Komano, T. (1988). Nucleotide sequence of the *rci* gene encoding shufflon-specific DNA recombinase in the IncII plasmid r64: homology to the site-specific recombinases of integrase family. *Molecular & general genetics: MGG*, 213(1):30–35. 15, 23
- Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNARNA interactions in yeast. *Proceedings of the National Academy of Sciences*, 108(24):10010–10015. 115, 116, 117
- Kurschat, W. C., Muller, J., Wombacher, R., and Helm, M. (2005). Optimizing splinted ligation of highly structured small RNAs. *RNA*, 11(12):1909–1914. 117
- Kvitsiani, D., Ranade, S., Hangya, B., Taniguchi, H., Huang, J. Z., and Kepecs, A. (2013). Distinct behavioural and network correlates of two interneuron types in prefrontal cortex. *Nature*, 498(7454):363–366. 133
- Lagunavicius, A., Merkiene, E., Kiveryte, Z., Savaneviciute, A., Zimbaite-Ruskulienė, V., Radzvilavicius, T., and Janulaitis, A. (2009). Novel application of phi29 DNA polymerase: RNA detection and analysis in vitro and in situ by target RNA-primed RCA. *RNA*, 15(5):765–771. 129
- Lan, N., Howrey, R. P., Lee, S. W., Smith, C. A., and Sullenger, B. A. (1998). Ribozyme-mediated repair of sickle beta-globin mRNAs in erythrocyte precursors. *Science (New York, N. Y.)*, 280(5369):1593–1596. 142
- Langer, S. J., Ghafoori, A. P., Byrd, M., and Leinwand, L. (2002). A genetic screen identifies novel noncompatible loxP sites. *Nucleic Acids Research*, 30(14):3067–3077. 106
- Lapham, J. and Crothers, D. M. (1996). RNase h cleavage for processing of in vitro transcribed RNA for NMR studies and RNA ligation. *RNA (New York, N. Y.)*, 2(3):289–296. 120
- Lapham, J., Yu, Y. T., Shu, M. D., Steitz, J. A., and Crothers, D. M. (1997). The position of site-directed cleavage of RNA using RNase h and 2'-o-methyl oligonucleotides is dependent on the enzyme source. *RNA*, 3(9):950–951. 120
- Le, Y., Gagnet, S., Tombaccini, D., Bethke, B., and Sauer, B. (1999). Nuclear targeting determinants of the phage p1 cre DNA recombinase. *Nucleic Acids Research*, 27(24):4703–4709. 100
- Lee, G. and Saito, I. (1998). Role of nucleotide sequences of loxP spacer region in cre-mediated recombination. *Gene*, 216(1):55–65. 20, 106

REFERENCES

- Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., Terry, R., Jeanty, S. S. F., Li, C., Amamoto, R., Peters, D. T., Turczyk, B. M., Marblestone, A. H., Inverso, S. A., Bernard, A., Mali, P., Rios, X., Aach, J., and Church, G. M. (2014). Highly multiplexed subcellular RNA sequencing in situ. *Science*, 343(6177):1360–1363. 10, 65, 89, 154, 161, 162
- Leone, P., Comoletti, D., Ferracci, G., Conrod, S., Garcia, S. U., Taylor, P., Bourne, Y., and Marchot, P. (2010). Structural insights into the exquisite selectivity of neuroligin/neurexin synaptic interactions. *The EMBO Journal*, 29(14):2461–2471. 87
- Lichtman, J. W. and Denk, W. (2011). The big and the small: Challenges of imaging the brains circuits. *Science*, 334(6056):618–623. 5
- Liu, D. S., Loh, K. H., Lam, S. S., White, K. A., and Ting, A. Y. (2013). Imaging trans-cellular neuroligin-neurexin interactions by enzymatic probe ligation. *PLoS ONE*, 8(2):e52823. 5, 47
- Livet, J., Weissman, T. A., Kang, H., Draft, R. W., Lu, J., Bennis, R. A., Sanes, J. R., and Lichtman, J. W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62. 5, 14, 20, 47, 109, 153
- Lu, R., Neff, N. F., Quake, S. R., and Weissman, I. L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature Biotechnology*, 29(10):928–933. 14, 59, 154
- Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013). RNA-guided human genome engineering via cas9. *Science*, 339(6121):823–826. 86, 134
- Marblestone, A. H., Daugharthy, E. R., Kalhor, R., Peikon, I. D., Kebschull, J. M., Shipman, S. L., Mishchenko, Y., Dalrymple, D. A., Zamft, B. M., Kording, K. P., Boyden, E. S., Zador, A. M., and Church, G. M. (2013). Connectomics: The economics of large-scale neural connectomics. *bioRxiv*. 9, 10
- Marblestone, A. H., Daugharthy, E. R., Kalhor, R., Peikon, I. D., Kebschull, J. M., Shipman, S. L., Mishchenko, Y., Lee, J. H., Kording, K. P., Boyden, E. S., Zador, A. M., and Church, G. M. (2014). Rosetta brains: A strategy for molecularly-annotated connectomics. *arXiv:1404.5103 [q-bio]*. arXiv: 1404.5103. 10, 159
- Mellor, J. and Roth, F. (2014). Personal communication. 34
- Missirlis, P. I., Smailus, D. E., and Holt, R. A. (2006). A high-throughput screen identifying sequence and promiscuity characteristics of the loxP spacer region in cre-mediated recombination. *BMC Genomics*, 7(1):73. 20, 106
- Moore, J. T., Uppal, A., Maley, F., and Maley, G. F. (1993). Overcoming inclusion body formation in a high-level expression system. *Protein Expression and Purification*, 4(2):160–163. 27
- Motmans, K., Thirion, S., Raus, J., and Vandevyver, C. (1997). Isolation and quantification of episomal expression vectors in human t cells. *BioTechniques*, 23(6):1044–1046. 102
- Movshon, J. A. and Newsome, W. T. (1996). Visual response properties of striate cortical neurons projecting to area MT in macaque monkeys. *The Journal of Neuroscience*, 16(23):7733–7741. 4
- Murray, I. A., Stickel, S. K., and Roberts, R. J. (2010). Sequence-specific cleavage of RNA by type II restriction enzymes. *Nucleic Acids Research*, 38(22):8257–8268. 118
- Naik, S. H., Peri, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R. J., and Schumacher, T. N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature*, 496(7444):229–232. 14
- Nishigaki, K., Taguchi, K., Kinoshita, Y., Aita, T., and Husimi, Y. (1998). Y-ligation: An efficient method for ligating single-stranded DNAs and RNAs with t4 RNA ligase. *Molecular Diversity*, 4(3):187–190. 117
- Oh, S. W., Harris, J. A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A. M., Mortrud, M. T., Ouellette, B., Nguyen, T. N., Sorensen, S. A., Slaughterbeck, C. R., Wakeman, W., Li, Y., Feng, D., Ho, A., Nicholas, E., Hirokawa, K. E., Bohn, P., Joines, K. M., Peng, H., Hawrylycz, M. J., Phillips, J. W., Hohmann, J. G., Wornoutka, P., Gerfen, C. R., Koch, C., Bernard, A., Dang, C., Jones, A. R., and Zeng, H. (2014). A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214. 2, 108, 153
- Oviedo, H. V., Bureau, I., Svoboda, K., and Zador, A. M. (2010). The functional asymmetry of auditory cortex is reflected in the organization of local cortical circuits. *Nature Neuroscience*, 13(11):1413–1420. 153
- Peikon, I. D., Gizatullina, D. I., and Zador, A. M. (2014). In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic Acids Research*, page gku604. 45, 65, 159
- Peixoto, R. T., Kunz, P., Kwon, H., Mabb, A., Sabatini, B., Philpot, B., and Ehlers, M. (2012). Transsynaptic signaling by activity-dependent cleavage of neuroligin-1. *Neuron*, 76(2):396–409. 64
- Pfeffer, C. K., Xue, M., He, M., Huang, Z. J., and Scanziani, M. (2013). Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature neuroscience*. 10
- Pinskiy, V., Tolpygo, A. S., Jones, J., Weber, K., Franciotti, N., and Mitra, P. P. (2013). A low-cost technique to cryo-protect and freeze rodent brains, precisely aligned to stereotaxic coordinates for whole-brain cryosectioning. *Journal of Neuroscience Methods*, 218(2):206–213. 161
- Raymond, C. S. and Soriano, P. (2007). High-efficiency FLP and c31 site-specific recombination in mammalian cells. *PLoS ONE*, 2(1):e162. 100
- Roth, G. and Dicke, U. (2005). Evolution of the brain and intelligence. *Trends in Cognitive Sciences*, 9(5):250–257. 8, 20, 153
- Ryall, B., Eydallin, G., and Ferenci, T. (2012). Culture history and population heterogeneity as determinants of bacterial adaptation: the adaptomics of a single environmental transition. *Microbiology and Molecular Biology Reviews*, 76(3):597–625. 1

REFERENCES

- Sara, Y., Biederer, T., Atasoy, D., Chubykin, A., Mozhayeva, M. G., Sdhof, T. C., and Kavalali, E. T. (2005). Selective capability of SynCAM and neuroligin for functional synapse assembly. *The Journal of Neuroscience*, 25(1):260–270. 50
- Saura, C. A., Servin-Morilla, E., and Scholl, F. G. (2011). Presenilin/-secretase regulates neurexin processing at synapses. *PLoS ONE*, 6(4):e19430. 64
- Schumacher, T. N. (2014). Personal communication. 14
- Shadlen, M. N. and Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *The Journal of Neuroscience*, 18(10):3870–3896. 3
- Siegel, R. W., Jain, R., and Bradbury, A. (2001). Using an in vivo phagemid system to identify non-compatible loxP sequences. *FEBS Letters*, 505(3):467–473. 20, 106
- Slavoff, S. A., Liu, D. S., Cohen, J. D., and Ting, A. Y. (2011). Imaging protein-protein interactions inside living cells via interaction-dependent fluorophore ligation. *Journal of the American Chemical Society*, 133(49):19769–19776. 50
- Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197. 18
- Song, S., Sjöström, P. J., Reigl, M., Nelson, S., and Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol*, 3(3):e68. 4, 46
- Srinivas, N., Ouldrige, T. E., ulc, P., Schaeffer, J. M., Yurke, B., Louis, A. A., Doye, J. P. K., and Winfree, E. (2013). On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Research*, 41(22):10641–10658. 125
- Stark, M. R., Pleiss, J. A., Deras, M., Scaringe, S. A., and Rader, S. D. (2006). An RNA ligase-mediated method for the efficient creation of large, synthetic RNAs. *RNA*, 12(11):2014–2019. 117
- Sullenger, B. A. and Cech, T. R. (1994). Ribozyme-mediated repair of defective mRNA by targeted trans-splicing. *Published online: 13 October 1994; | doi:10.1038/371619a0*, 371(6498):619–622. 140
- Suzuki, K., Hayashi, Y., Nakahara, S., Kumazaki, H., Prox, J., Horiuchi, K., Zeng, M., Tanimura, S., Nishiyama, Y., Osawa, S., Sehara-Fujisawa, A., Saftig, P., Yokoshima, S., Fukuyama, T., Matsuki, N., Koyama, R., Tomita, T., and Iwatsubo, T. (2012). Activity-dependent proteolytic cleavage of neuroligin-1. *Neuron*, 76(2):410–422. 64
- Taniguchi, H., He, M., Wu, P., Kim, S., Paik, R., Sugino, K., Kvitsani, D., Fu, Y., Lu, J., Lin, Y., Miyoshi, G., Shima, Y., Fishell, G., Nelson, S. B., and Huang, Z. J. (2011). A resource of cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. *Neuron*, 71(6):995–1013. 89
- Taylor, A. M., Blurton-Jones, M., Rhee, S. W., Cribbs, D. H., Cotman, C. W., and Jeon, N. L. (2005). A microfluidic culture platform for CNS axonal injury, regeneration and transport. *Nature Methods*, 2(8):599–605. 59
- Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H., Kotsopoulos, S. K., Samuels, M. L., Hutchison, J. B., Larson, J. W., Topol, E. J., Weiner, M. P., Harismendy, O., Olson, J., Link, D. R., and Frazer, K. A. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology*, 27(11):1025–1031. 72
- Thyagarajan, B., Olivares, E. C., Hollis, R. P., Ginsburg, D. S., and Calos, M. P. (2001). Site-specific genomic integration in mammalian cells mediated by phage c31 integrase. *Molecular and Cellular Biology*, 21(12):3926–3934. 102
- Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., and Chklovskii, D. B. (2011). Structural properties of the caenorhabditis elegans neuronal network. *PLoS Comput Biol*, 7(2):e1001066. 3, 5
- Wei, Y. and Koulikov, A. A. (2012). An exactly solvable model of random site-specific recombinations. *Bulletin of Mathematical Biology*, 74(12):2897–2916. 22
- Wetterstrand, K. (2013). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). 8
- White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1986). The structure of the nervous system of the nematode caenorhabditis elegans. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 314(1165):1–340. 3, 5
- Wickersham, I. R., Lyon, D. C., Barnard, R. J., Mori, T., Finke, S., Conzelmann, K.-K., Young, J. A., and Callaway, E. M. (2007). Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron*, 53(5):639–647. 6
- Yamagata, M. and Sanes, J. R. (2012). Transgenic strategy for identifying synaptic connections in mice by fluorescence complementation (GRASP). *Frontiers in Molecular Neuroscience*, 5:18. 5, 77
- Yamashita, T., Pala, A., Pedrido, L., Kremer, Y., Welker, E., and Petersen, C. H. (2013). Membrane potential dynamics of neocortical projection neurons driving target-specific signals. *Neuron*, 80(6):1477–1490. 153
- Yoshimura, Y. and Callaway, E. M. (2005). Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nature Neuroscience*, 8(11):1552–1559. 4, 46
- Zador, A. M., Dubnau, J., Oyibo, H. K., Zhan, H., Cao, G., and Peikon, I. D. (2012). Sequencing the connectome. *PLoS Biol*, 10(10):e1001411. 2, 12, 14, 47, 65
- Zaug, A. J., Grosshans, C. A., and Cech, T. R. (1988). Sequence-specific endoribonuclease activity of the tetrahymena ribozyme: enhanced cleavage of certain oligonucleotide substrates that form mismatched ribozyme-substrate complexes. *Biochemistry*, 27(25):8924–8931. 127

5

Declaration

Declaration

I herewith declare that I have produced this work.

This thesis work was conducted from 2009 to 2014 under the supervision of Dr. Anthony M. Zador at Cold Spring Harbor Laboratory.

Ian Peikon

Appendices

Appendix A

Rci in mammalian cells

A.1 Introduction

We previously showed that Rci can be employed in bacterial cells, including *E. coli*, to shuffle artificial cassettes for the purpose of cellular barcoding (see chapter 2). However, our long-term goal is to be able to produce barcodes in mammalian cells, such as mouse neurons. Here, we tested whether Rci functions in mammalian cells.

A.2 Expression of Rci in mammalian cells

A.2.1 Cellular localization of Rci

In order to test the cellular localization of Rci, we created a plasmid, IDP013 (Figure A.1A), which expresses Rci fused at its N-terminal to GFP. Expression of this plasmid in mammalian cells (HEK293 cells) via transient transfection revealed nuclear localization of Rci, despite the lack of detection of a known nuclear localization signal (NLS) (Figure A.1B). Other recombinases, namely the widely employed Cre-recombinase, reach the nucleus of mammalian cells without modification [Le et al. 1999]. In many cases, however, the addition of a nuclear localization signal (NLS) enhances efficiency that can be dependent on its fusion position [Andreas et al. 2002, Raymond and Soriano 2007]. Therefore, we created several variants of the Rci construct, in which the SV40 NLS was fused to the N- or C-terminal.

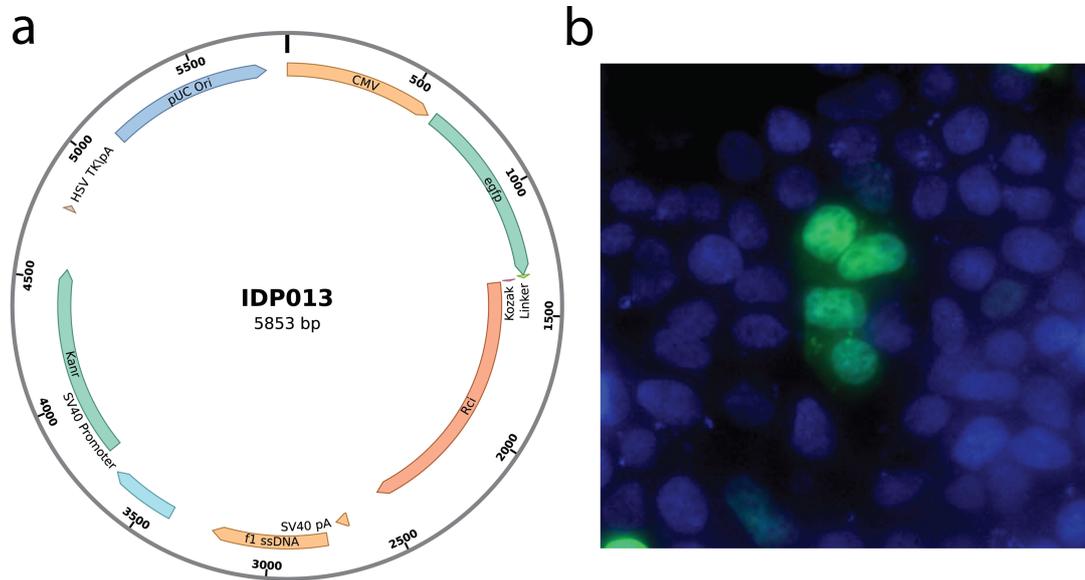


Figure A.1: Rci expression in mammalian cells - (a) Rci is expressed as a fusion protein with an N-terminal eGFP. (b) eGFP-Rci is localized to nuclei. Green=eGFP-Rci, Blue=Dapi.

A.3 Functional testing of Rci in mammalian cells

A.3.1 Rci-mediated inversion

To test if Rci can mediate inversion in mammalian cells we set up a highly sensitive, PCR assay. Briefly, DNA sequences were designed in which PCR can proceed only if inversion has taken place. This is accomplished by placing one primer (P1) binding site outside of the recombination cassette and an additional primer (P3) binding site within the recombination cassette – i.e. flanked by *sfx* sites in opposing orientation (Figure A.2A). Before recombination, the two primer binding sites are in the sense orientation. However, after inversion, the second primer binding site is inverted to antisense orientation, permitting PCR to occur using the primer pair (Figure A.2A). Additional primers (P1, P4, P101) are included in the design for detection (P3-P101, P3-P4) or excision (P2-P3, discussed in subsection A.3.2).

A.3.1.1 Rci-mediated inversion on a plasmid substrate

To test for Rci-mediated inversion, we co-expressed several variants of Rci with a plasmid substrate containing the *sfx* sites *sfxaa201* in opposing orientation (i.e. for inversion). The site *sfxaa201*, which is a symmetric *sfx* site (except for the directional core), was shown in bacteria to be the most efficient *sfx* site and was therefore used for initial testing [Gyohda et al. 2006]. It should be noted that this site was shown in bacteria to permit excision events [Gyohda et al. 2006], and thus the *sfxaa201* site was only used initially to test the function of Rci under the best case conditions in mammalian cells. We co-transfected the inversion test plasmid with *sfxaa201* sites incubated for 24 or 48 hours, and isolated DNA using a column-based plasmid extraction [Motmans et al. 1997]. Approximately 10ng of each sample was loaded into a PCR reaction to probe the presence of the test cassette as well as inversion. We found that Rci mediates inversion on a plasmid substrate with or without an NLS (Figure A.2B). Recombination is inhibited by additional tags, specifically those on the C-terminal, as expected based on the posited role of the C-terminus of Rci in forming tetramers [Gyohda et al. 2002; 2006].

A.3.1.2 Rci-mediated inversion on a genomic substrate

To test for Rci-mediated inversion on a genomic substrate, we first created a cell-line harboring the test cassette (Figure A.2A). To accomplish this, we took advantage of a HEK cell line, which harbors a PhiC31 landing site within the genome [Keravala et al. 2008, Thyagarajan et al. 2001]. The test cassette was cloned into an integration plasmid designed to integrate specifically at the genomic landing site via PhiC31-mediated integration. Co-transfection of the plasmid harboring the cassette and a plasmid expressing a mutant version of PhiC31 resulted in integration of the cassette into the genome, and allowed for selection of a stable cell-line [Thyagarajan et al. 2001]. After isolation of a stable cell-line, we expressed variants of Rci via transient transfection, incubated for 48 hours, and isolated genomic DNA using a column-based genomic extraction (Promega). Approximately 10ng of isolated DNA from each sample was loaded into a PCR reaction to probe the presence of the test cassette as well as inversion (Figure A.2C). We found that Rci mediated inversion on the genomic substrate with or

A.3 Functional testing of Rci in mammalian cells

without an additional SV40 NLS. No recombination was detected in the absence of Rci expression (Figure A.2A).

A.3.2 Rci-mediated deletions

Rci is a site-specific invertase and has not been shown to mediate excision events in bacterial cells. Unlike other site-specific invertases like Gin and Hin [Johnson 2002], no required co-factors have been discovered that confer this exquisite specificity. However, Rci function has not previously been described in mammalian cells, and thus the specificity of Rci for inversion over excision must be empirically determined in this setting. We designed a PCR-based assay, analogous to that for inversion, in order to probe the ability of Rci to mediate excision in mammalian cells Figure A.3. We tested several different *sfx* sites that had previously been identified and shown to function in bacterial cells [Gyohda et al. 2006]. Importantly, these are wild-type sequences which allow Rci to selectively mediate inversion. However, it is unknown if this specificity remains in mammalian cells. In the inversion constructs, *sfx* sites are placed in opposing orientation, whereas in the excision constructs they are placed in the same orientation Table A.1.

Table A.1: Constructs used for inversion and excision testing of Rci

| Construct | <i>sfx</i> site 1 | <i>sfx</i> site 2 | orientation |
|-----------|-------------------|-------------------|-----------------|
| IDP081 | <i>sfxa101</i> | <i>sfxa101</i> | inverted repeat |
| IDP082 | <i>sfxa101</i> | <i>sfxd103</i> | inverted repeat |
| IDP083 | <i>sfxa101</i> | <i>sfxb107</i> | inverted repeat |
| IDP084 | <i>sfxa101</i> | <i>sfxa101</i> | direct repeat |
| IDP085 | <i>sfxd103</i> | <i>sfxd103</i> | inverted repeat |
| IDP086 | <i>sfxd103</i> | <i>sfxb107</i> | inverted repeat |
| IDP087 | <i>sfxd107</i> | <i>sfxd107</i> | inverted repeat |

We observed recombination regardless of the specific *sfx* sites employed. In addition, we observe excision when *sfx* sites are placed in the same orientation (IDP084). It remains possible that the excision events detected are low-frequency events that are detected only because of the highly-sensitive PCR assay. Further work will be needed to quantitatively assess the efficiency of inversion and excision in both bacterial and mammalian systems.

A.3 Functional testing of Rci in mammalian cells

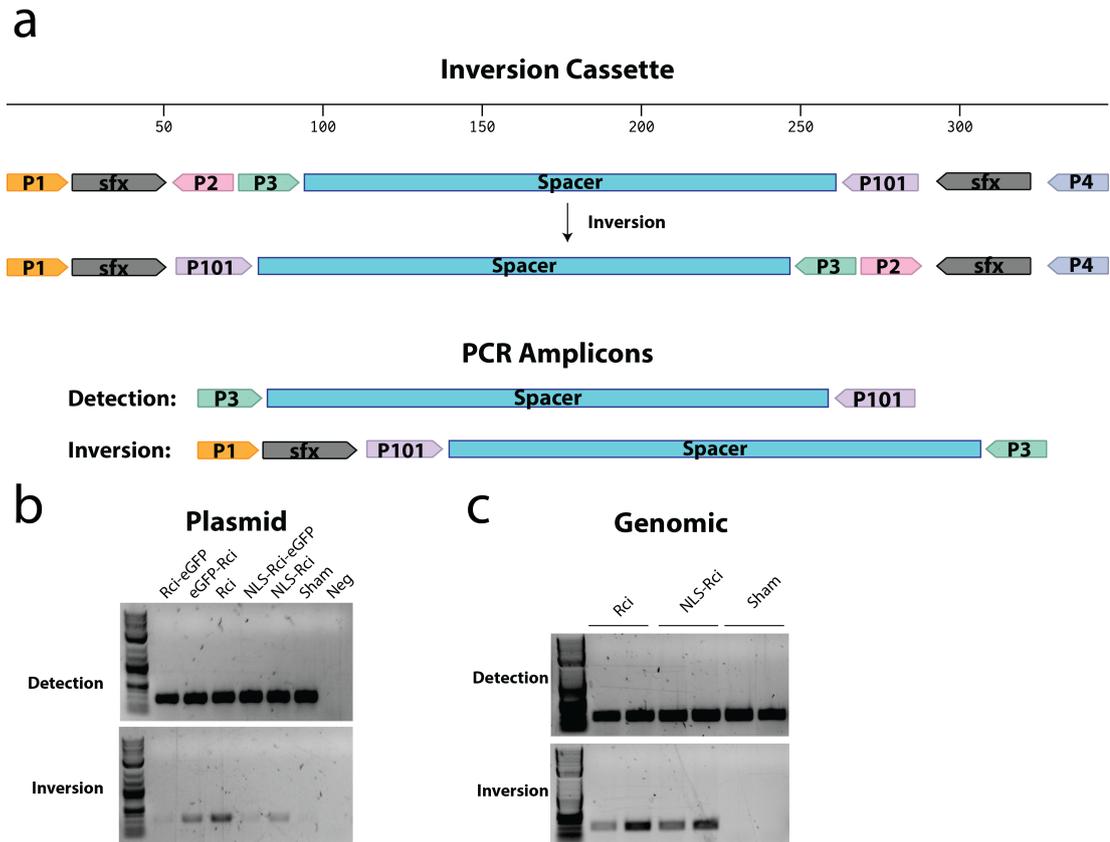


Figure A.2: Rci-mediated inversion in mammalian cells - (a) Inversion cassettes consist of several primers (P1-P4, P101) and two *sfx* sites in opposing orientation. Primers P3 and P101 are always in opposing orientation and can be used for detection via PCR. Primers P1 and P3 begin in the same orientation and cannot be used for PCR. Rci-mediated inversion flips primer P3 relative to P1 and allows PCR-detection of inversion events. (b) co-transfection of Rci expression plasmids and the inversion cassette results in inversion. (c) A genomic inversion substrate is inverted after the transient expression of Rci via transient transfection.

A.3 Functional testing of Rci in mammalian cells

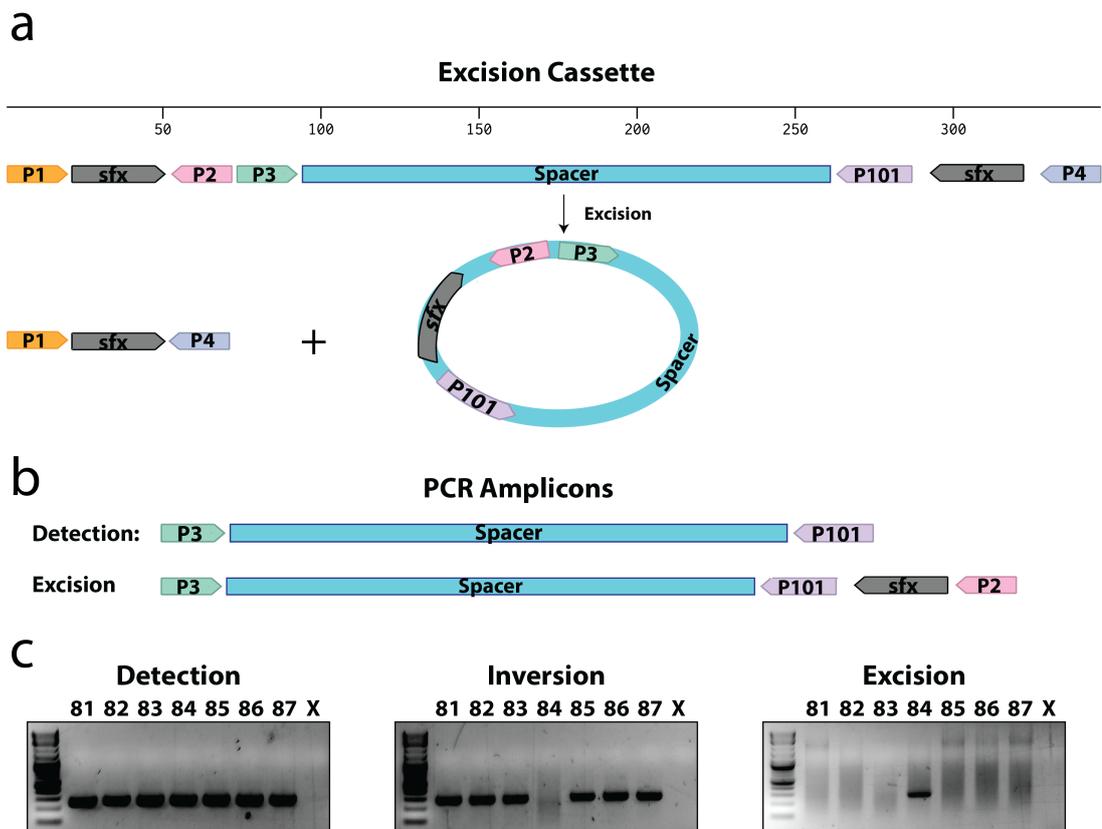


Figure A.3: Rci-mediated excision in mammalian cells - (a) The excision cassette. (b) Amplicons formed in the PCR assay. The excision amplicon is dependent on excision to properly orient primers P3 and P2 for PCR. (c) Co-transfection with Rci results in recombination in every scenario. Rci mediates excision only when the sites are in direct repeat.

A.4 Discussion

Here we have shown preliminary evidence that Rci indeed functions in mammalian cells. Our highly sensitive PCR assay, however, cannot discriminate rare from high-efficiency events. It will be necessary to quantitatively measure the inversion and excision efficiency of Rci in mammalian cells in order to assess the feasibility of employing Rci for mammalian barcoding (see chapter 2).

Our initial attempts to shuffle a barcode cassette in mammalian cells using Rci, was largely unsuccessful. We found very low frequency (1/80 genomes assayed) recombination and observed, anecdotally, toxicity associated with Rci expression in cells that harbor the genomic cassette. This may be due to genomic instability caused by repeated double-stranded break events caused by recombination. It is possible that this could be overcome with inducible expression of Rci, and/or expression of Rci only in non-dividing cells (such as neurons). In addition, Rci recognizes highly degenerate sequences, therefore decreasing the effective length of the *sfx* site to roughly 19 nucleotides [Gyohda et al. 2002; 2006]. At this length, it is probable that many pseudosites exist within the human genome, by chance alone, perhaps providing a substrate for Rci binding/recombination that could result in large chromosomal translocations and subsequently, cell death. Other recombinases, such as PhiC31 and/or Cre recombinase, work well in mammalian cells, and could perhaps be modified for shuffling schemes. Indeed many mutant recombination sites for both PhiC31 and Cre have been developed with interesting characteristics [Colloms et al. 2014, Langer et al. 2002, Lee and Saito 1998, Missirlis et al. 2006, Siegel et al. 2001], and it has been proposed that the asymmetric nature of the *sfx* site confers the specificity for inversion with respect to Rci-mediated recombination [Gyohda et al. 2006]. To date, no such site has been isolated for PhiC31 or Cre, but targeted high-throughput screening may reveal asymmetric sites with preferences for inversion. Alternative methods that take advantage of new tools which allow the precise targeting of double-stranded DNA breaks that stimulate DNA repair processes (CRISPR, TALENs, ZFNs) could prove fruitful in the near future. By stimulating non-homologous end joining, it may be possible to direct the non-templated addition of DNA nucleotides to a specific locus in order to create a *de novo* DNA barcode.

The ability to barcode mammalian cells *in vivo* will unlock the ability to track cells through development, cell-cell interactions, and the mobility of cells within the body (i.e. metastatic cells). Combined with high-throughput measurements (i.e. sequencing), barcoding will allow for an unprecedented look at the heterogeneity of individual cells within complex tissues.

A.5 Acknowledgements

We thank Michele Calos for providing the plasmids required for creating stable cell lines using PhiC31. Diana Gizatullina and I performed all experiments. Anthony Zador, Diana Gizatullina, and I designed all of the experiments.

Appendix B

Mapping long-range projections with high-throughput sequencing

B.1 Introduction

There is increasing interest in mapping long-range connectivity in the mammalian nervous system – the “projectome.” The Allen Brain Projection Atlas, for example, provides the brain-wide pattern of projections from each of about 1000 injection sites in the mouse brain [Oh et al. 2014]. Such information is very useful for designing experiments for investigating how information is transmitted from one brain area to another. Here we propose to develop a relatively simple, inexpensive, and high-yield method for establishing the projection pattern of single neurons to a given target in the mouse brain.

Most existing methods used for probing long-range projections, including those used to generate the Allen Projection Atlas [Oh et al. 2014], provide information only about the aggregate projections of neurons from one region to another. These methods do not readily resolve single neurons. Thus these methods can tell us that neurons in area X projects to areas Y1, Y2, and Y3, but they cannot in general discover, for example, that one subset of neurons projects mainly to Y1 and Y2, whereas another subset projects mainly to Y2 and Y3. To the extent that the information represented in one area is partitioned among distinct output “streams,” determining these distinct output targets is a crucial first step in formulating hypotheses about the functional role of these different streams.

Historically, the only method for determining projection patterns at single neuron resolution has been to label neurons very sparsely and image the labeled axonal tree. Such studies are very labor intensive. A recent study [Kita and Kita 2012] on the projections of layer V corticofugal neurons illustrates the state of the art. Tracer injections into 28 rats yielded 25 reconstructed projections. Although throughput could potentially be increased somewhat by “Brainbow” [Livet et al. 2007] in which neurons express unique combinations of different colored fluorophores, the number of colors resolvable in a single specimen is limited, so in practice the increase in throughput for assessing long-range connections is modest. We recently proposed using FISSEQ (see F) of cell-identifying barcodes expressed throughout neuronal processes for tracing of long-range projections with single neuron resolution. While this technology has the potential to scale to full brains, it remains prohibitively costly and time-consuming for most laboratories. Here we outline a simple, efficient, and cost-effective method for tracing long-range projections at single-neuron resolution using standard high-throughput sequencing techniques that will be easily accessible for labs with standard molecular biology equipment.

Our goal is to develop an efficient method for mapping neuronal projections based on DNA-sequencing that can be adopted by neuroscience laboratories with only basic molecular biology expertise. Therefore, it is our objective to simplify the method to the form of a simple protocol, or “kit” in order to make the technology as widely available as possible. By converting projection mapping to a DNA-sequencing problem, the projections of hundreds or thousands of individual neurons can be traced simultaneously, in a single experiment. The core of our approach is to use viruses to barcode a population of neurons near an injection site with a random short sequence of RNA. The barcode, which is read out by sequencing, acts as a unique identifier for each neuron. The barcode, either through passive diffusion or active transport, reaches the distal projections of cells. In this way, in contrast to all existing approaches, we can identify distant targets of a neuron without tracing its entire axon. Here we propose to develop a relatively simple, inexpensive and high-yield iteration of this method for establishing the projection pattern of single neurons in the mouse brain and employ our method to two areas Auditory Cortex (A1) and Locus coeruleus (LC) to demonstrate the potential of the technology across brain regions with different properties.

B.2 Results

The basic strategy is depicted in Figure B.1. First, we inject a barcoded virus (Figure B.1A) into a brain region of interest (Figure B.1B). A barcode of length 30nt has a theoretical diversity of $4^{30} = 10^{18}$, far more than the number of neurons infected in a typical experiment ($10^2 - 10^4$). In practice the actual diversity of $< 10^7$ we routinely obtain is limited by several bottlenecks encountered during viral production, but is still far in excess of the number of neurons in a typical experiment. The barcoded virus thus acts as a unique identifier for the neuron it infects. The barcode is transported, either by passive diffusion or by engineered binding to a modified pre-synaptic protein, to the distal projections (Figure B.1C). The animal is then sacrificed, and target brain regions of interest (ROIs) are dissected. The resolution of the projection map is determined by the size of the dissected target ROI. Depending on the dissection methods used, target ROIs of $\sim 500\mu\text{m}$ can readily be obtained, and with some optimization ROIs as small as $100\mu\text{m}$ or even smaller are feasible. Finally the barcode mRNA is extracted from the ROI, reverse transcribed, PCR amplified and subjected to multiplexed high throughput sequencing (Figure B.1D). The sequencing information is disentangled to form a map of which neurons project to which ROIs (Figure B.1E) and the strength of the projection (in number of neurons) to each ROI (Figure B.1F).

B.2.1 Tracing of neurons originating in auditory cortex

Barcoded Sindbis virus was injected into the left auditory cortex of a mouse. RNA transport to axons was facilitated by co-expression of a modified form of the pre-synaptic protein Neurexin1B, engineered to interact specifically and efficiently with the RNA barcode by virtue of the λN functional domain, which binds to a short RNA sequence, boxB, that has been inserted in 4 copies into the RNA barcode transcript [Daigle and Ellenberg 2007]. This protein, represents the pre-synaptic half of our design for mapping neuronal connectivity via high-throughput sequencing (see chapter 3). The protein binds the RNA with high affinity as shown by RNA-IP and traffics properly to synapses as measured by the proximity ligation assay (see chapter 3 for details).

Forty-eight hours after infection, animals were sacrificed and acute slices were prepared for dissection of target ROIs. Seven target regions of interest were dissected, and barcodes from each region were extracted and sequenced. These regions included:

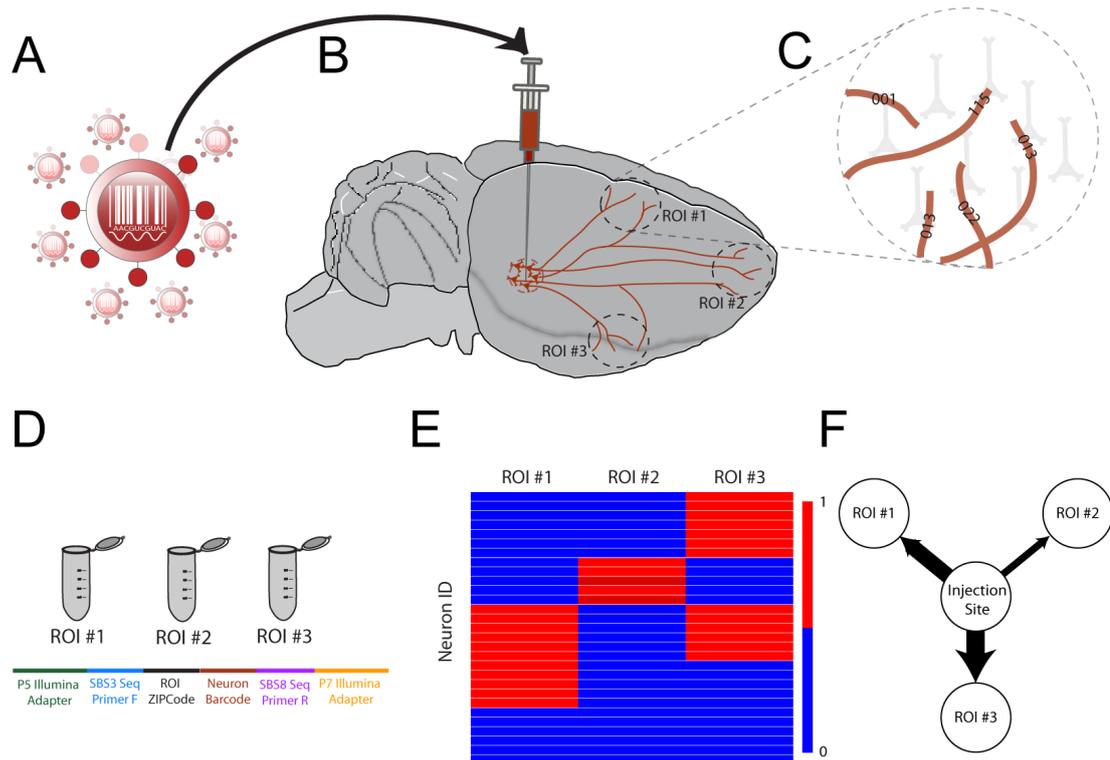


Figure B.1: Employing DNA sequencing for projection mapping - (a) A high-diversity barcode library is packaged into a virus (i.e. Sindbis, Lentivirus, Adeno Associated Virus). (b) Barcoded virus is infected focally into the brain region whose projection pattern is to be analyzed. (c) Viral RNA is transported into projecting axons in distant brain regions. (d) Brain regions of interest (ROIs) are dissected and RNA is extracted. Sequencing libraries are prepared with ROI barcodes added during library construction. (e) A binary projection map is generated based on sequencing data and clustered to reveal the higher-order structure of the projection pattern with single neuron resolution. (f) A strength of projection diagram can be drawn based on the number of injected neurons projecting to each target ROI.

(ipsilateral striatum anterior, ipsilateral striatum posterior, ipsilateral thalamus, ipsilateral inferior colliculus, contralateral inferior colliculus, contralateral auditory cortex, and contralateral striatum. The overall pattern of projection (Figure B.2A) was consistent with the known anatomy. For example, one of the most numerous projections was to the ipsilateral auditory striatum. The single neuron projection data (Figure B.2B) highlighted some surprising results. For example, we observe very few neurons which project to the contralateral striatum. Further work will be needed to understand the single-neuron projection data in more depth and is discussed in chapter 4.

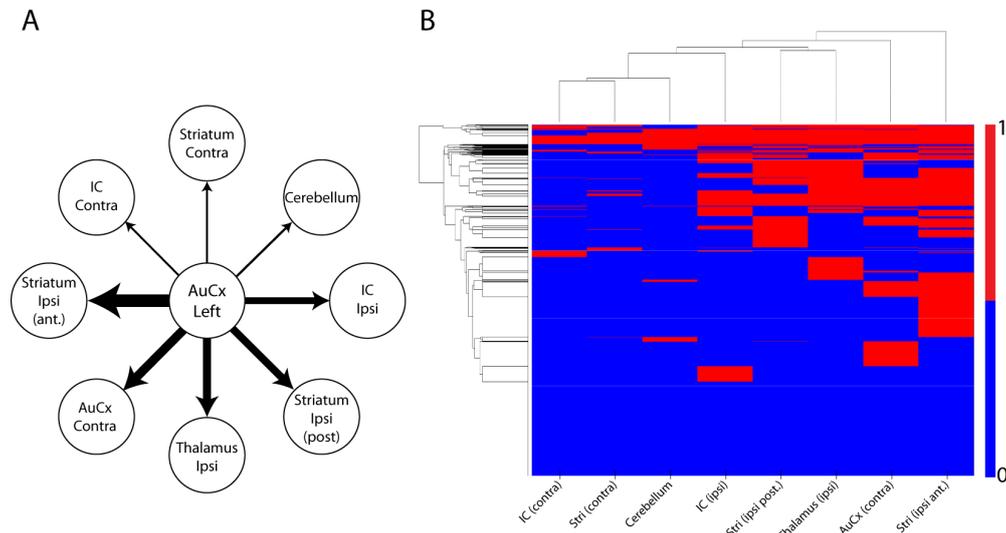


Figure B.2: Reconstruction of projectome from left auditory cortex - The left auditory cortex was injected with barcoded Sindbis virus. (a) Reconstructed projection strengths from auditory cortex to various regions. Line thickness represents the number of neurons projecting to each area. (B) Single neuron clustergram. Each row represents an individual neuron (barcode) and each column denotes a region of interest. The data are clustered to reveal structure of the projection pattern at single neuron resolution.

B.3 Discussion

This pilot experiment demonstrates the basic feasibility of the approach. Several aspects require optimization. PCR is very sensitive, so we expect the false negative rate to be low. On the other hand, false positives (i.e. false discovery of barcodes in area

to which the neurons do not project) arising from RNA contamination during tissue dissection are a significant concern; we expect they can be minimized by thorough washing and RNAase treatment prior to membrane lysis. RNA contamination from fibers of passage is another concern which can be addressed with other methods.

These conclusions are preliminary; further control experiments must be performed to rule out alternative interpretations. First, we must minimize the rate of false positives, i.e. the rate at which barcodes that appear in a given ROI arise from contamination from another region. Such contamination can arise during tissue dissection and RNA extraction. We have spent considerable time optimizing the protocol to minimize such contamination (using as a negative control the fact that the auditory cortex does not project to the cerebellum), and now estimate it to be quite low. Second, we must minimize the rate of false negatives. False negatives can occur due to failure of barcode RNA to reach the synapse, inefficient RNA reverse transcription, or insufficient sequencing depth.

As discussed above, our bulk sequencing methods do not allow for the precise knowledge of the position of the soma of each individual neuron. Imaging of the infection site (i.e. if RNA barcodes are within the 5'UTR of the transcript of an XFP) will reveal the extent of the labeling, but will not resolve the correspondence between spatial position of a soma and its barcode. By sequencing barcoded RNAs in situ in a brain slice – FISSEQ will allow us to determine the laminar position of each barcoded soma (see F). FISSEQ may also make it possible to relate projection information with physiology data including functional calcium imaging within the same preparation, e.g. by extending the techniques of (Ko et al. 2011).

B.4 Methods

B.4.1 Injections

Stereotactic injections of barcoded Sindbis virus were performed to introduce the components into the brain. All injections were performed in 5-week old male CBA/CaJ mice in left auditory cortex under anesthesia (ketamine medetomidine). Following injections, mice were returned to their cages for 48 hours before sacrificing.

B.4.2 RNA isolation

Brains were dissected in HEPES buffer at room temperature and individual regions of interest were placed immediately in Trizol (Life Technologies), homogenized, and frozen on dry ice. RNA isolation was performed according to the Trizol manufacturer's protocol.

B.4.3 RT-PCR

Reverse transcription was performed with gene-specific primers that add a Solexa II adapter to the cDNA sequence. PCR was performed using SolexaI and SolexaII primers (P5-SBS3T and P7-SBS8).

B.5 Acknowledgements

This work was done in collaboration with several other graduate students (former and current). Peter Znamenskiy and I performed the first experiments tracing fibers in auditory cortex. Pedro Garcia de Silva and Justus Keschull have since performed subsequent experiments done in locus coeruleus and have been extremely helpful in discussions.

Appendix C

RNA ligation methods

C.1 Introduction

In the context of our project to map neural connectivity via DNA sequencing (see chapter 3), the immunoprecipitated RNA barcode pairs representing synaptic partners (Figure 2.1b) must be covalently linked for subsequent reverse-transcription PCR and DNA sequencing (Figure 2.1c). This is, in concept, very similar to ligation methods that have previously been developed for RNA and/or DNA [Kalhor et al. 2012, Kudla et al. 2011]. Initially, we pursued several alternative RNA-linking methodologies in parallel before ultimately settling on overlap extension emulsion reverse transcription PCR (OEemRTPCR). Here we describe the other methods that were considered and their limitations. Finally, we consider additional methods which, with further development, may be more optimal than OEemRTPCR.

C.2 General considerations

C.2.1 Maintaining the Protein-RNA complexes

Because of the set-up employed in our project (see chapter 3), there are several constraints which must be considered. First, the RNA and protein are tightly associated by a non-covalent bond. Therefore, any ligation method employed must be compatible with the protein-RNA interaction. In addition, the membrane proteins bound to the RNA must be kept in solution. This requires the use of detergent and glycerol subsection 3.5.1.1 in all buffers – which may be incompatible with certain enzymatic

reactions.

C.2.2 Stoichiometry of ligation reactions

Stoichiometry is crucial when performing a ligation. An excess of the joining molecule be it a bridge oligo or a bifunctional ribozyme will lead to each target containing a bridge, thus prohibit joining. This poses a significant challenge because, once on beads, the exact quantities of all of the species present are difficult to ascertain.

C.2.3 Pol-II transcripts

It is important to note that many of the issues we faced could potentially be circumvented by the use of an alternative RNA polymerase. However, the maximum transcript length that such polymerases can support is not well tested a major consideration for in vivo barcode generation (see chapter 2 for a description of the constraints). Moreover, subcellular localization of non Pol-II transcripts could pose problems for tethering of RNA barcodes to the synapse and must be tested empirically. Finally, the use of alternative promoters to drive the expression of RNA would prohibit the use of certain viruses, namely Sindbis. We use Sindbis for purely strategical reasons (fast turn-around time, high expression), and thus substitution with other viruses that tolerate alternative promoters is possible. In any case, because we employed Pol-II, all of our barcode transcripts are 5'capped and polyadenylated. These modifications prohibit many direct biochemical ligation methods, as will be discussed.

C.3 ssRNA ligation

Perhaps the simplest method of joining two RNAs relies on the ssRNA ligase, T4 RNA Ligase 1 (T4RNL1). T4RNL1 catalyzes the nonspecific intra- or intermolecular ligation of two single stranded nucleic acid chains (ssRNA or ssDNA) terminating in 3'OH (acceptor) and 5'P (donor), respectively. ssRNA ligation has been successfully employed previously to join two closely apposed RNA molecules [Kudla et al. 2011]. However, T4RNL1 has strong preferences for intramolecular ligation and severely reduced ligation efficiencies for intermolecular ligation. Intermolecular ligation is often plagued by unwanted side products (i.e. intramolecular circularization and/or concatemers [Kalhor

et al. 2012, Kurschat et al. 2005]. The efficiency of intermolecular ligation can be increased by creating a pseudo intramolecular reaction via a bridge oligo [Stark et al. 2006] or by exploiting complementarity between the two RNA molecules to be joined [Kudla et al. 2011, Nishigaki et al. 1998]. These techniques, however, require synthetic blocking groups on the ends of the ssRNA substrates in order to prevent concatenation and/or circularization – technically challenging to achieve with transcripts prepared *in vivo*. In addition, the presence of a long polyA tail makes the splinting of PolII transcripts non-trivial – requiring some sort of predictable trimming to < 7 A's [Jan et al. 2011]. Finally, T4RNL1-mediated RNA ligation is traditionally employed for the ligation of short RNA molecules. Long RNA molecules have significant secondary structure that will increase the likelihood of unwanted side-reactions. This problem can be somewhat ameliorated by heating and snap cooling, or heating in the presence of complementary oligos, but our ability to heat the sample is limited by the requirement to maintain the protein-RNA interaction. For these reasons, we did not pursue ssRNA ligation.

C.4 Splint ligation

In splint ligation, two RNA molecules are bridged by an oligonucleotide provided *in trans* (Figure C.1). Ligation by T4 DNA ligase (T4DNL) or T4 RNA ligase 2 (T4RNL2) occurs if and only if a 5'P and 3'OH are perfectly aligned by the bridge oligo, thus largely prohibiting side-reactions including circularization and concatemerization [Kurschat et al. 2005]. This method is largely employed for the ligation of synthetic RNA molecules, produced either via chemical synthesis or *in vitro* transcription.

C.4.1 Splinted ligation of short oligonucleotides

We tested the two ligases, T4DNL and T4RNL2 for the ligation of two fluorescently labeled oligonucleotides. Briefly we mixed a 5'Cy3 labeled acceptor oligonucleotide (Cy3) and a 5'Phosphorylated, 3'Cy5 labeled acceptor (Cy5) with a splint and either ligase. We found this reaction to be highly efficient using both T4DNL and T4RNL2 (Figure C.2). Previous research has shown that T4RNL2 is the most efficient of the bacteriophage T4 ligases at sealing an RNA-RNA nick bridged by a DNA strand [Bullard and Bowater 2006], and therefore, we continued with T4RNL2.

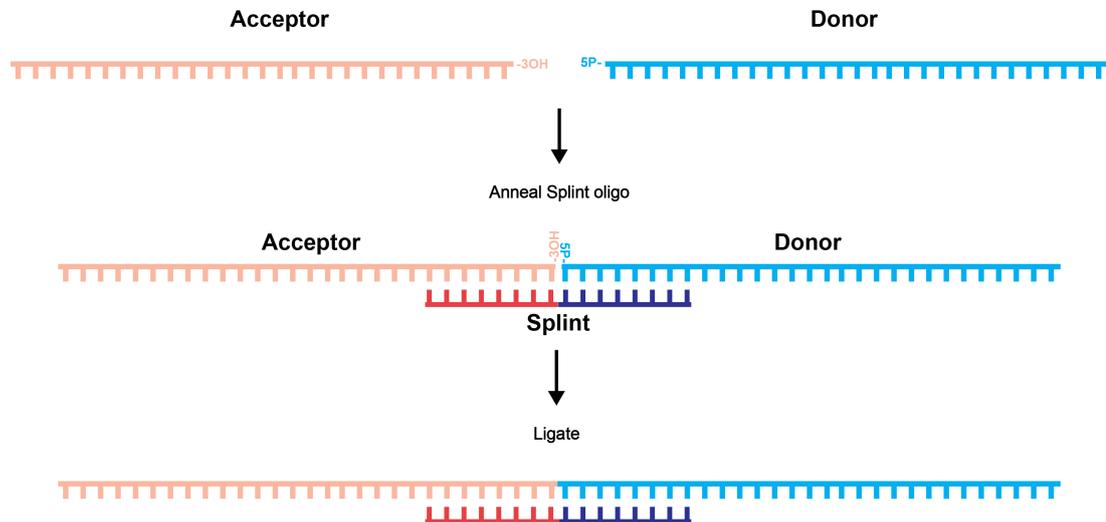


Figure C.1: Splint Ligation of RNA - An acceptor molecule with a free 3'OH and a donor molecule with a free 5'P are splinted by a bridge oligonucleotide. A ligase seals the nick between the splinted acceptor and donor RNA molecules.

C.4.2 Adapting splinted ligation for Pol-II transcripts

To adapt this method for our purposes – to ligate two long RNA molecules transcribed *in vivo* by Pol-II – we had to solve several problems. Importantly, because of the exquisite base-pairing specificity of T4DNL and T4RNL2, it is an absolute requirement that the sequences at the ends of both RNA molecules are known. In addition, there is a requirement for a free 3'OH on the acceptor RNA molecule and a free 5'P on the donor molecule (Figure C.1). Because our barcodes are expressed as Pol-II transcripts, the ends are not suitable for splint ligation directly. First, the terminal sequences are not known with certainty, particularly the 3' end at which the length of polyA tails are highly variable. In addition, the 5' ends of Pol-II transcripts are capped, and therefore lack the necessary 5'P for ligation.

In order to generate ends suitable for RNA ligation, we desired a method for the site-specific cleavage of RNA transcripts that results in the appropriate ends (i.e. 3'OH and 5'P). It has been reported that certain restriction enzymes can cleave RNA in an RNA-DNA hybrid [Murray et al. 2010] but these methods were inefficient in our hands. Instead, we used RNaseH mediated cleavage, in which RNaseH is directed to cleave

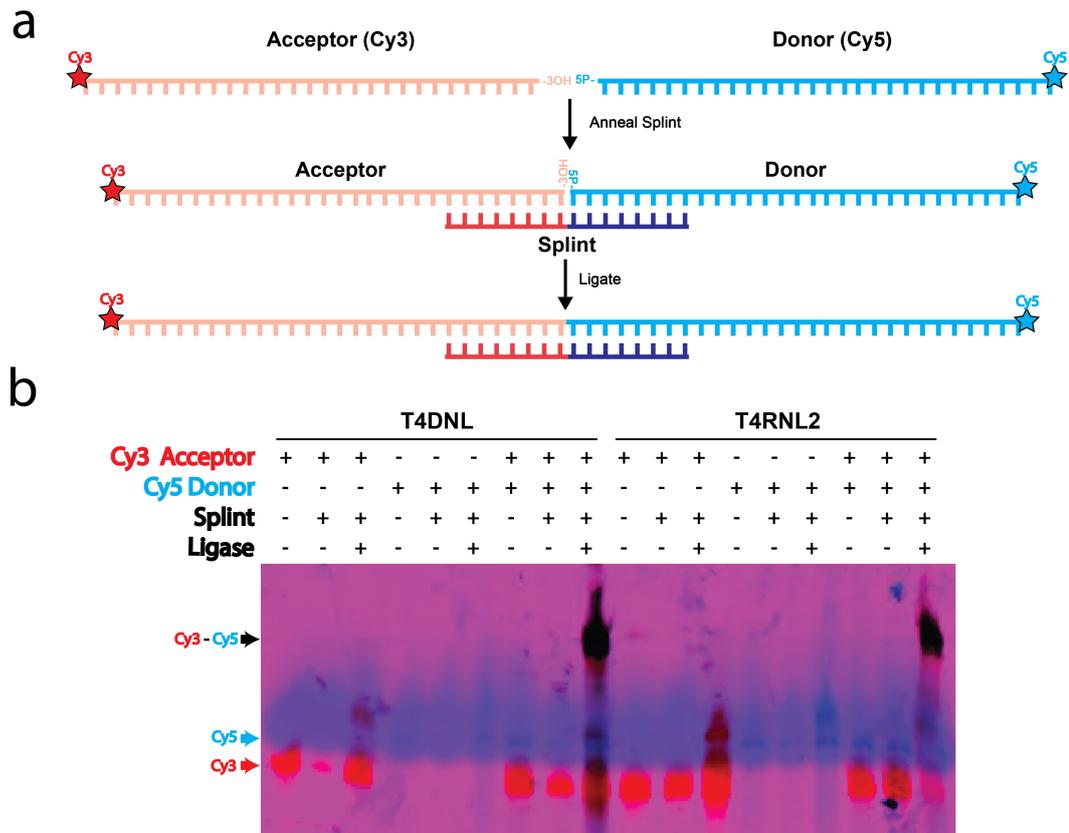


Figure C.2: Splinted ligation of short oligonucleotides - (a) A Cy3-labeled acceptor oligonucleotide is ligated with a Cy5 labeled donor oligonucleotide in the presence of a splint. Perfect base pairing is required at the junction for ligation. (b) Ligation occurs only in the presence of both oligos, a splint, and either T4 DNA ligase (T4DNL) or T4 RNA ligase 2 (T4RNL2).

specifically by a 2'Ome/DNA hybrid oligo [Inoue et al. 1987, Lapham and Crothers 1996].

C.4.2.1 RNaseH cleavage

RNaseH is an endoribonuclease which cleaves the phosphodiester bond of RNA molecules that are hybridized to complementary DNA. If the RNA is hybridized to a modified nucleotide, such as 2'O-methyl (2'Ome) RNA nucleotides, RNaseH is unable to cleave the RNA strand. However, hybrid 2'Ome/DNA oligonucleotides, when properly designed, direct the RNaseH cleavage at a single dinucleotide junction in the target RNA strand (Figure C.3). Regions of the RNA that are hybridized to the 2'Ome nucleotides are protected from enzymatic cleavage. The RNA bases that are paired with complementary DNA bases are, however, available for cleavage. The footprint of the RNaseH protein permits the cleavage only at a single dinucleotide junction (when a 4nt DNA segment is present in the targeting oligo). Importantly, the exact position at which the cleavage occurs is dependent on the source of the RNaseH [Lapham et al. 1997] and must be considered.

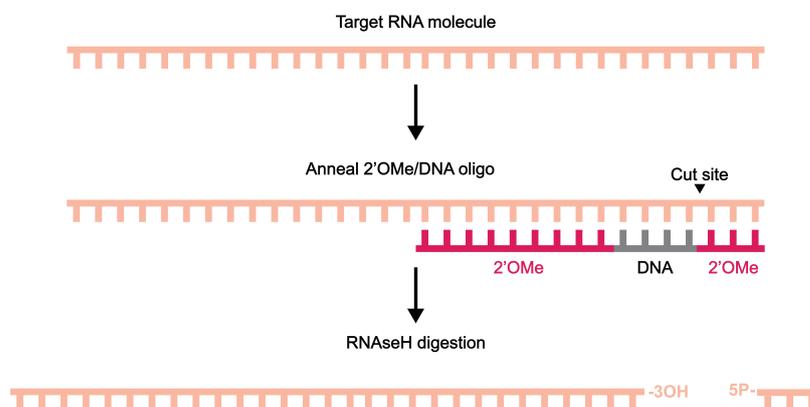


Figure C.3: Site-specific RNaseH cleavage directed by 2'Ome/DNA hybrids - A 2'Ome/DNA hybrid oligonucleotide is annealed to a target RNA. RNaseH cleaves at a specific position – in the RNA strand at the 5' junction of the DNA/2'Ome nucleotides in the complementary oligonucleotide – leaving a 3'OH and a 5'P on the newly generated RNA fragments.

C.4 Splint ligation

To test RNaseH cleavage we *in vitro* transcribed two RNA molecules (RNA5 and RNA8) and designed 2'Ome/DNA hybrid complementary oligonucleotides to direct cleavage to a specific site (Figure C.4A). We first tested 2'Ome/DNA directed RNaseH cleavage under idealized conditions in which we allowed heating and cooling for efficient hybridization of the oligonucleotide to the RNA molecule. RNaseH-mediated cleavage under these conditions was specific and efficient (Figure C.4B). The reaction could be driven to near 100% efficiency by increasing the ratio of oligonucleotide to target RNA (autoreffFigureC-4B).

Because of the limitations of our system, we are unable to heat our samples for hybridization. Therefore, we tested the performance of RNaseH cleavage after different hybridization conditions. We tested a range of hybridization temperatures from 37°C to 95°C and observed only slight changes, qualitatively, in the efficiency at the lower hybridization temperatures (Figure C.4C). The binding of 2'Ome bases to RNA bases is considerably stronger than RNA:RNA binding or DNA:RNA binding. Therefore, the requirements for efficient hybridization may be reached even at physiological temperatures (i.e. 37°C). It is possible to compensate for the reduced efficiency at lower temperatures by the addition of higher concentrations of oligo (Figure C.4D).

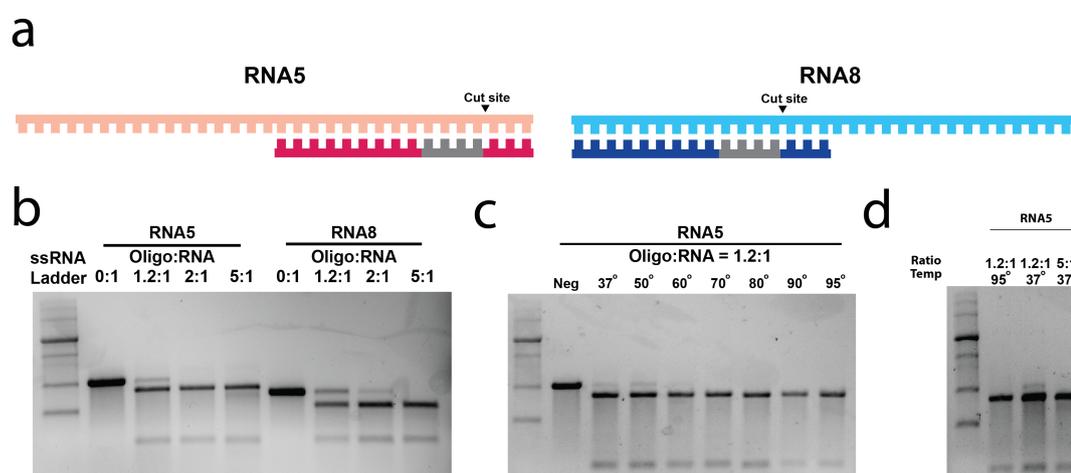


Figure C.4: RNaseH-mediated cleavage of *in vitro* transcribed RNA - (a) Two RNAs, RNA5 and RNA8 are targeted via complementary 2'Ome/DNA hybrid oligonucleotides. (b) Different ratios of oligo:target were tested for cleavage efficiency. (c) A range of hybridization temperatures was tested to determine the effects of heating on proper hybridization and subsequent cleavage. (d) Oligos were added at different ratios at different hybridization temperatures.

C.4.2.2 Ligation

RNaseH mediated cleavage is efficient and specific and results in RNA fragments with known end sequences that have the appropriate 3'OH and 5'P that are necessary for subsequent ligation. Following RNA cleavage, an oligo that bridges the pre-synaptic and post-synaptic RNA barcodes must be annealed for ligation. Therefore, after RNaseH cleavage, we purified the RNA by phenol:chloroform extraction and then performed splint ligation with a splint that bridges the cut RNA5 and RNA8. Unfortunately, we found that the targeting oligo remains bound at physiological temperatures via the extensive 2'Ome:target-RNA base-pairing (Figure C.5A). When the splint oligo is added and the mixture is heated for ideal-case annealing, the 2'O methyl oligo (which was used for cutting but which is still present) competes with the splint oligo and there is no ligation across the two RNA. The splint oligo in turn competes with the 2'O methyl for the re-ligation of the original molecule, and thus here we observe little re-ligation (Figure C.5B). However, if the splint oligo is added without any heating, 2'O methyl oligo remains bound and directs the re-ligation of the original target RNA (Figure C.5B).

Again, in our system, heat denaturation to remove the 2'Ome/DNA oligo from the cleavage reaction is not possible because of potential dissociation of the RNA/protein complexes. Moreover, heat denaturation and cooling in the presence of both the splint and cleavage-directing oligos causes competition for annealing and lowers the splint efficiency – higher concentrations of the splint will inhibit the reaction due to each RNA molecule annealing to a separate splint (Figure C.5B).

This problem is primarily an issue for the acceptor (RNA5) because here, the splint will not be able to displace the 2'Ome oligo. In the case of the donor (RNA8), the splint should easily displace the 2'Ome oligo, as the 2'Ome only binds to 4 bases where the splint will bind. Therefore, we wondered if RNaseH digestion could be performed on the acceptor RNA with an oligo that would serve to direct both the digestion and the subsequent ligation (Figure C.6A,B,C). The addition of the overhang to the oligo resulted in ~50% efficiency of cutting. The loss of efficiency could not be compensated for easily with additional oligo. We therefore pursued a different method, based on strand replacement.

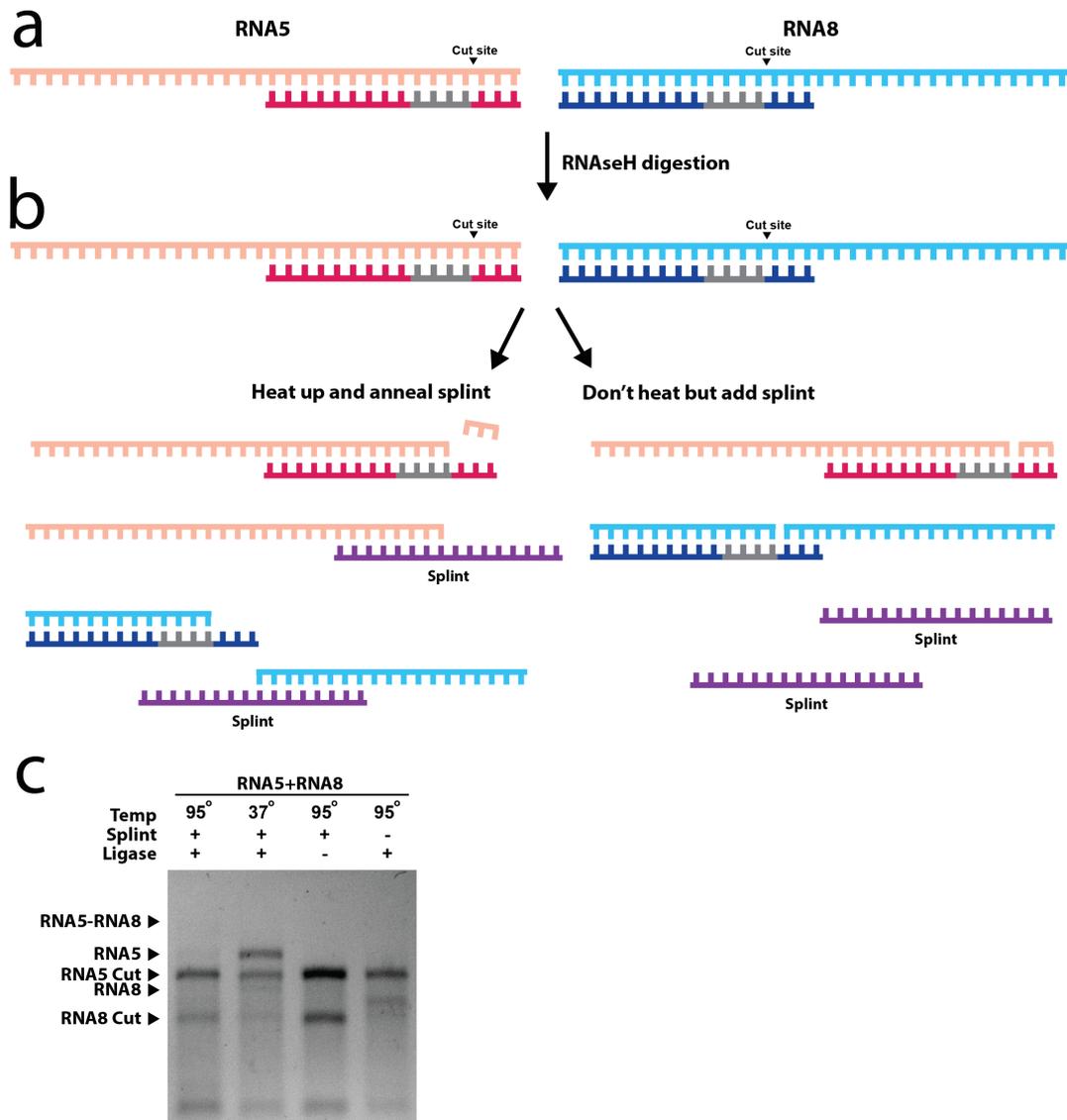


Figure C.5: Competition of targeting oligo and splint - (a) The target RNAs are hybridized to the RNaseH-directing 2'Ome/DNA hybrid oligonucleotides. (b) RNaseH nicks the RNA strand at a single dinucleotide. Heating in the presence of splint results in competition between the splint and the 2'Ome oligo for binding to the target RNA. Without heating, the splint cannot compete with the strong binding of the 2'Ome oligo and re-ligation is possible. (c) Religation occurs in the presence of splint without heating.

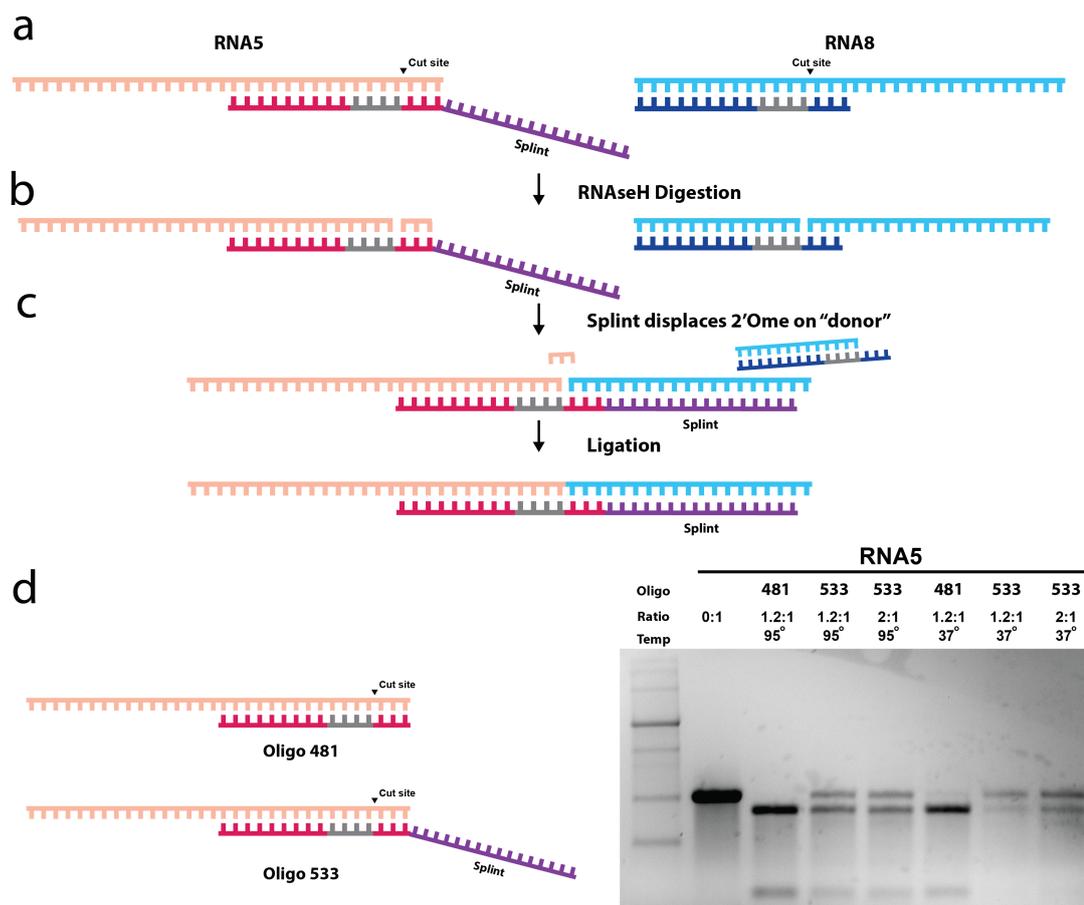


Figure C.6: Dual cleavage and splint oligo - (a) 2'Ome/DNA hybrid oligos are annealed to target RNAs to direct cleavage. To the acceptor RNA (RNA5), an oligo which serves to direct cleavage and subsequent ligation (via an overhang) is annealed). (b) RNAseH digests both target RNAs. (c) The overhang from the acceptor (RNA5) targeting oligo displaces the 2'Ome on the donor RNA (RNA8), splinting the two target RNAs for ligation. (d) Comparison of the cutting efficiency using a normal targeting oligo (Oligo 481) and a targeting oligo with overhang (Oligo 533).

C.4.2.3 Strand replacement

We took advantage of strand-replacement to replace the 2'Ome/DNA cleavage oligos with a 2'Ome splint oligo. Strand replacement is a phenomena by which one oligonucleotide will replace another oligonucleotide on a target strand if it is thermodynamically favored (Figure C.7A). In this case, the splint oligo must also be 2'Ome because of the thermodynamics associated with strand replacement (2'Ome binds to RNA stronger than DNA binds to RNA). To make the reaction thermodynamically favorable, the splint oligo binds to each RNA strand with higher binding energy than the cleavage oligo. This is achieved by additional base-pairing, known as a toehold (Figure C.7B) [Srinivas et al. 2013]. The splint will then replace the original oligo through a random walk process (Figure C.7C,D). Once the splint oligonucleotide is in place, the RNA strands can be ligated with either T4DNL or T4RNL2. We tested strand displacement using a 2'Ome oligo (Figure C.7E,F). Briefly, an Alexa-488 labeled oligo was annealed to RNA5 or RNA8 via heating to 95°C and slow cooling to 37°C. Next, an Alexa-594 labeled oligo was added to the reaction and incubated at 37°C for 0, 6, or 16 hours (Figure C.7E,F). We found the reaction to be slow, but efficient, with nearly all of the Alexa-488 oligo displaced by the Alexa-594 oligo after 16 hours.

C.4.2.4 Strand replacement and ligation

Finally, after performing RNaseH digestion of the acceptor (RNA5) and donor (RNA8) molecules, and strand replacement with the splint oligonucleotide, we attempted RNA ligation. The entire reaction took place isothermally, at 37°C so as not to disturb the RNA/protein interaction. While we do observe a ligation product when the splint and ligase are added, the reaction was inefficient Figure C.8. Moreover, because of the slow reaction times associated with RNaseH digestion (4 hours), strand replacement (16 hours), and ligation (4 hours), we worried about the stability of the complexes on the beads. In addition, the presence of the 2'Ome oligo poses additional problems on reverse transcription – the 2'Ome oligo cannot be displaced efficiently by reverse transcriptase and cannot be enzymatically degraded. For these reasons, this method was therefore ruled out.

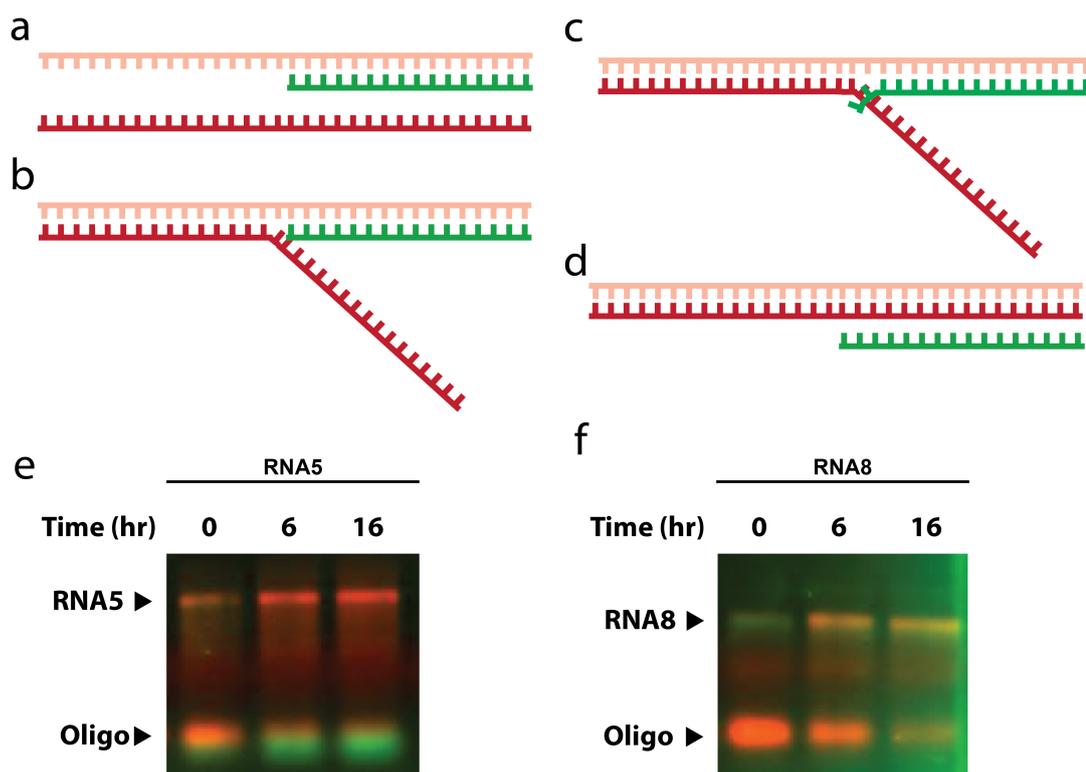


Figure C.7: Monitoring strand replacement with fluorescent oligos - (a) A green oligo (Alexa488 labeled) is annealed to a target RNA strand (pink). A longer red oligo (Alex594 labeled) is added to the tube. (b) The 594-oligo anneals to the unoccupied bases (toehold). (c) The 488-oligo is replaced via a random walk process, initially a single base at a time until (d) the 594-oligo has completely replaced the 488-oligo. (e) Time course of strand replacement of a 488-oligo annealed to target RNA5, displaced by 594-oligo. (f) Time course of strand replacement of a 488-oligo annealed to target RNA8, displaced by 594-oligo.

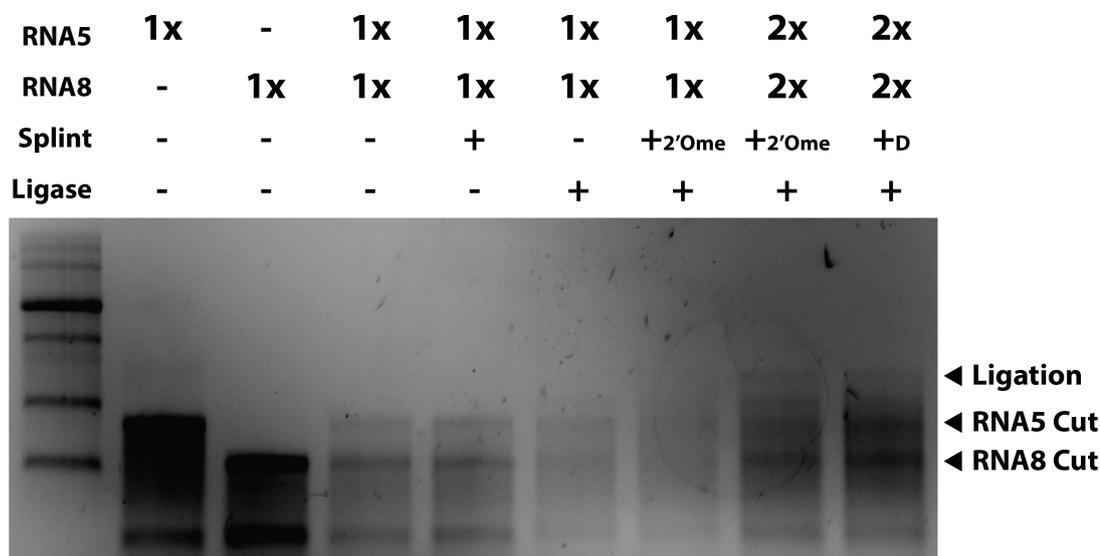


Figure C.8: Strand replacement and ligation - (a) After digestion with RNaseH, the 2'Ome targeting oligos were replaced with the splint and the target RNAs were ligated together.

C.5 Catalytic nucleic acids

C.5.1 Ribozymes

Next, we tested Ribozymes – catalytic RNA molecules – for the cleavage/ligation of the two RNA molecules. The L21 ribozyme from *Tetrahymena* [Zaug et al. 1988] has been shown to mediate 3' replacement through a cleavage/ligation reaction. This allows for ligation of a "donor" RNA transcript onto an acceptor RNA molecule by cleavage of the acceptor and ligation at its 3' end. Recently, a 5' replacement ribozyme has been described [Alexander et al. 2005] which performs cleavage of an acceptor RNA molecule and subsequent ligation at its 5' end. Generally, these ribozymes carry the donor transcript in *cis*. We wondered if it would be possible to generate, via *in vitro* transcription, a "bifunctional" ribozyme by concatenating the 3' replacement and 5' replacement ribozymes, that would ligate two RNA molecules when added in *trans* (Figure C.9).

Because stoichiometry is crucial (see subsection C.2.2) the ribozyme must be added in equimolar concentration to the two target RNAs to be joined. Unfortunately, how-

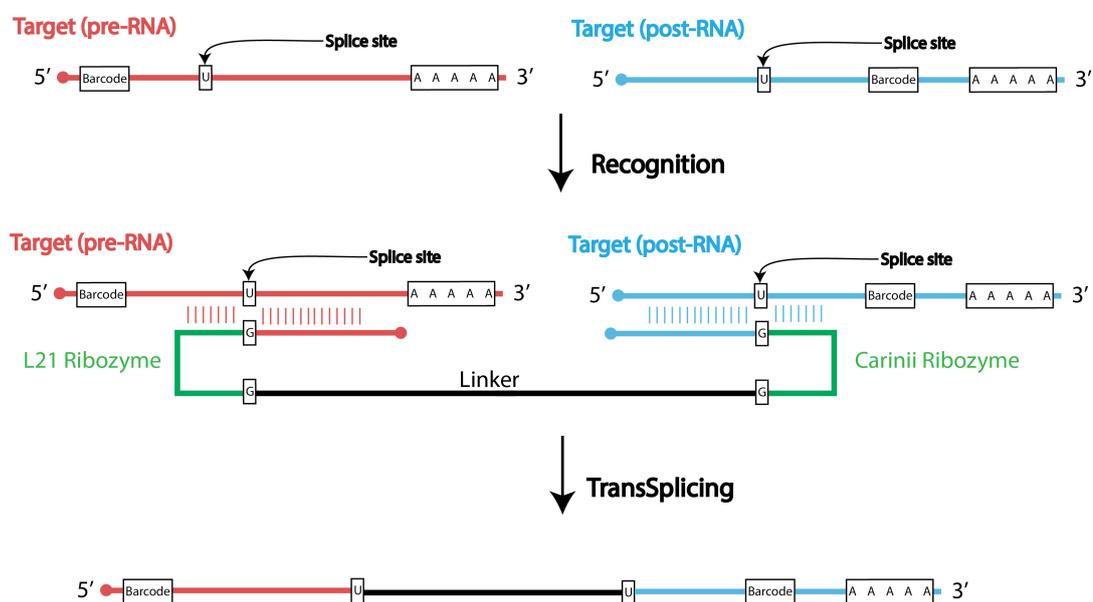


Figure C.9: Bifunctional ribozymes for ligation - A bifunctional ribozyme composed of the concatenation of a 3' L21 ribozyme and a mutant 5' replacement pCarinii Ribozyme, could in theory direct two transplicing reactions resulting in the ligation of two target RNA molecules.

ever, these ribozymes are inefficient. Successful splicing (even under best conditions) requires many fold excess of ribozyme relative to the target. In addition, the ribozymes require a high concentration of Mg for catalytic activity. Mg is a potent non-specific RNA endonuclease and may cause problems when dealing with small amounts of RNA. For these reasons we have decided not to further pursue this method.

C.5.2 Deoxyribozymes

We then tested Deoxyribozymes, catalytic ssDNA molecules that mediate cleavage or ligation reactions. We reasoned that we could build deoxyribozymes with binding energies sufficient to cleave the RNA transcripts, but insufficient to remain bound to either of the cleaved halves, such that the need for strand replacement would be circumvented. We tested deoxyribozymes for cutting and ligating RNA and found these reactions to be very inefficient compared to similar methods based on protein-mediated RNA cutting (RNaseH) and ligation (T4 RNA ligase, T4 RNA ligase 2, T4 DNA ligase). In addition, as with ribozymes, deoxyribozymes require high concentrations of divalent cations (Mg, Mn, Zn, etc) that cause nonspecific cleavage of the target RNA. Therefore we have not continued with this approach.

C.6 Joining by reverse transcription

C.6.1 Overlap extension reverse transcription

Two RNAs with complementary 3' ends can prime RT off of each other resulting in a joined RNA/DNA hybrid molecule (Figure C.10)A. This can then be amplified via PCR resulting in a joined barcode pair. We observed that this reaction can indeed join two RNA molecules *in vitro* (Figure C.10B). Interestingly we observed that RT can proceed in the absence of a primer, when either of the RNA molecules is left unpaired (Figure C.10B).

The method, however, has several important limitations. First, if the complementary region is not the extreme 3' end of the RNA molecule (as is the case with PolIII transcripts containing polyA sequences), this cannot work. Precise trimming of the sequence to the point of complementarity is difficult to achieve in practice – we tried using Phi29 polymerase [Lagunavicius et al. 2009], ExoT, and a few other exonucleases to achieve this, but none could trim to the exact spot of complementarity in a way that

would allow subsequent RT. In addition, this method requires heating for the proper annealing of the two RNAs – which will likely cause disruption of the protein:RNA complex.

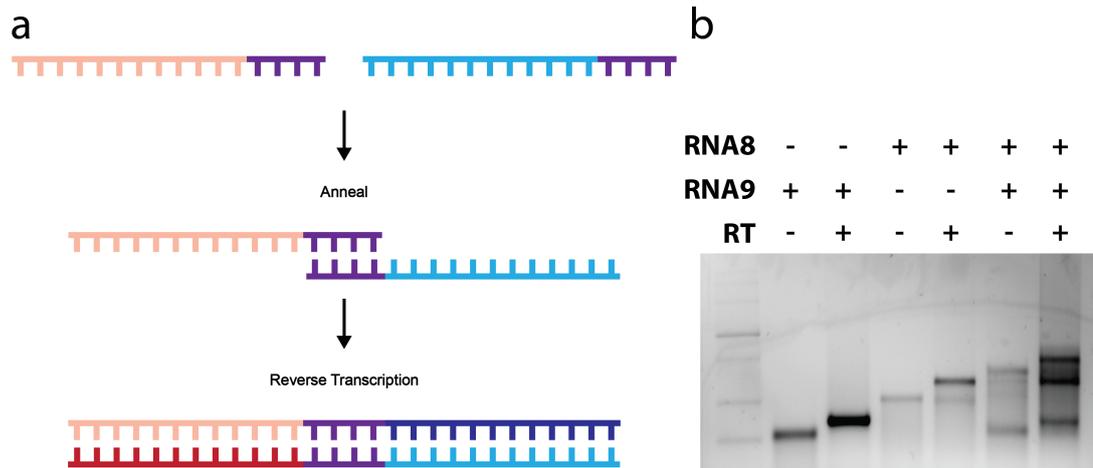


Figure C.10: Overlap reverse transcription - (a) Schematic of overlap RT joining. (b) Overlap RT joins molecules *in vitro*.

C.6.2 dsDNA primed reverse transcription

We tested whether we could perform reverse transcription of both molecules from RT primers that shared a complementary region and had been pre-annealed. In this case, RT proceeds from each ssDNA region and thus joins the two barcodes. RNA can then be degraded and the DNA barcode pair can be double stranded via 2nd strand synthesis – sealing the nicks with ligase (conceptually similar to Gibson assembly [Gibson et al. 2009]).

Unfortunately, we found that the secondary structure of the RNA allows reverse-transcription even in the absence of any primers, consistent with previous reports [Feng et al. 2012] This could perhaps be ameliorated by addition of an excess of RT primer – as is done in typical reverse transcription reactions, or heating of the sample to reduce secondary structure and increase primer binding. However, we are limited by the stoichiometry of the molecules (see subsection C.2.2, and the requirement to maintain the protein-RNA interaction. The addition of a DNA primer at an equimolar concentration to the RNA is insufficient to prime reverse transcription preferentially from the DNA

primer (as compared to from the RNA secondary structure) [Feng et al. 2012]. Finally, this method may pose a higher risk of "false positives" because the joined barcode pairs rely only on the short stretch of complementarity at their center from the pre-annealed DNA primer to maintain the correct association.

C.7 Discussion

We tried many unsuccessful RNA joining methods. However, with further optimization, it may be possible to employ one of these methods for RNA barcode joining. RNaseH digestion followed by splint ligation was the most promising of the methods. It may be possible to engineer additional DNA nucleotides into the 2'Ome splint at strategic locations to allow partial enzymatic digestion of the oligonucleotide, allowing release from the target after cleavage – thereby circumventing the need for the slow strand displacement reaction. For example, the introduction of dU, dispersed throughout the 2'Ome oligo should allow for cleavage of the 2'Ome oligo by Uracil-DNA Glycosylase. However, the RNaseH cleavage reaction will need to be tested with this type of an oligo to ensure that specificity is retained. Similar tricks could be employed in the splint for its degradation following ligation, to allow uninhibited reverse transcription to take place.

Additional methods which rely on overlap PCR may also allow efficient joining. For example, reverse transcription using biotinylated primers, or in the presence of a low-concentration of biotinylated nucleotides, would allow reverse transcription of the RNA molecules such that they would be bound to the streptavidin beads close to their original complex. Bridge overlap-PCR could then be used to fuse the two molecules into a single nucleic acid string representing the barcode pair. If this were done under dilute conditions on beads, it should allow for high-efficiency joining with low cross-over (i.e. false negative) rates.

In the limit, this method is equivalent to the emulsion PCR reaction which we have developed subsection 3.5.3.3, but is simpler in practice. Emulsion PCR requires the formation of high volumes of uniform droplets – a time consuming, and limiting process in practice. Methods that would permit the joining of RNA molecules without require emulsification would drastically increase the throughput of the connectome sequencing technique.

C.8 Methods

C.8.1 Splint ligation

For splint ligation, oligonucleotides were annealed to RNA to be ligated by first mixing at a ratio of 1:1.2:1.3 (Acceptor:Donor:Splint). EDTA was added to a final concentration of .1mM, and NaCl to a final concentration of 50mM. H_2O was added to reach a final volume of $5\mu L$. The reaction was placed in a thermocycler and heated to $95^\circ C$ for 5 min and then cooled slowly at a rate of $-.1^\circ C/s$ until reaching $35^\circ C$. To the annealed product, $1\mu L$ of 10X ligase buffer (NEB), $.5\mu L$ of RNAsin (Promega, $40U/\mu L$), $1.5\mu L$ H_2O , and $2\mu L$ Ligase (NEB – T4RNL2 or T4DNL) was added. The reaction was incubated at 37° for 1 hour.

C.8.2 RNaseH digestion

The target RNA (3pmol) is incubated with 3.3pmol of 2'Ome oligonucleotide in the presence of 100mM KCl and .1mM EDTA. H_2O is added to a final volume of $13\mu L$. The mixture is heated in a thermocycler to the desired annealing temperature (ranging from 37° to 95°) for 5 minutes and the cooled slowly at a rate of $-.1^\circ C/s$ until reaching $35^\circ C$. To the annealed product, $5\mu L$ of RNaseH buffer (NEB), $1.5\mu L$ of RNAsin (Promega $40U/\mu L$), $.5\mu L$ DTT (100mM), $15\mu L$ of RNaseH (NEB, $5U/\mu L$), and $10\mu L$ H_2O is added. The reaction is incubated at $37^\circ C$ for 4 hours.

C.8.3 Reverse transcription

All of the reverse transcription reactions were performed using SuperScript-III (Life Technologies) according to the manufacturer's instructions.

C.9 Acknowledgements

Scott Silverman (UIUC) was extremely helpful in the design of various RNA ligation schemes – I thank him for his support and patience. I also want to thank Dario Bressan (CSHL), and Ed Boyden (MIT) and various members of his lab (Fei Chen, Steve Bates) for great discussion about various RNA joining methods. Diana Gizatullina was instrumental in testing all of the approaches.

Appendix D

Non-transgenic, cell-type specific expression of Cre recombinase

D.1 Introduction

A central challenge in neuroscience is to understand how behavior arises from neural circuits. Circuits in the brain consist of diverse neuronal subtypes and subpopulations. Neurons have classically been distinguished on the basis of many characteristics, including morphology, receptor expression, neurotransmitter, and connectivity. More recently, neuronal subtypes have been defined by the selective expression of specific genes. The ability to target gene expression to genetically-defined neuronal subtypes provides a powerful tool for dissecting neural circuits. In *Drosophila melanogaster*, thousands of cell-type specific transgenic lines are available, and have been invaluable in unraveling the neuronal mechanisms underlying learning, development and behavior [Jenett et al. 2012]. A growing number of cell-type specific mouse lines have also become available. For example, the somatostatin (SOM) and parvalbumin (PV) transcripts define functionally different subclasses of interneurons, and these interneuron subpopulations may play different roles in circuit computation [Kvitsiani et al. 2013]. Our goal is to develop a broadly applicable method for conveniently manipulating the expression of transgenes, including fluorescent proteins and optogenetic tools, in genetically defined neuronal populations in the brains of rats, primates and other mammalian species.

We propose to develop a novel strategy for the expression of reporter genes in specific

neuronal subtypes. Our strategy is based on the specific expression of either Cre or Flp recombinase in specific neuronal cell-types. The expression of transgenes can then be restricted to neurons that express the recombinase. However, unlike previous strategies in which cell-type specific recombinase expression is achieved through the generation of transgenic knock-in mice, our method can deliver both the recombinase and the transgene via recombinant viruses. Thus our method can be applied in organisms in which the potential for genetic manipulation is limited.

D.1.1 Current approaches

There are two main strategies for achieving cell-type specific gene expression in mammals. It is important to note that we use the term “cell-type” to refer to any collection of neurons defined by their shared expression of a particular gene. We do not mean to imply that a cell-type defined by shared expression of a single gene corresponds to a functional cell-type, although in some cases gene expression does correlate well with neuronal function. The preferred strategy in mice is to make a “knock-in” – a mouse in which the transgene is placed directly at the locus of an endogenous gene such as SOM. With this approach, the expression of the transgene is driven directly from the promoter of the endogenous gene, and therefore in principle – and often in practice – recapitulates the expression pattern of the endogenous gene. However, there are at least two important limitations of the knock-in approach. First, the only mammalian species in which knock-ins are made routinely is the mouse. Although transgenic animals have been generated in a variety of other species, including rats and even marmoset monkeys, in these animals the transgene inserts into a random place in the genome. New methods, such as CRISPR [Cong et al. 2013, Mali et al. 2013], should allow for targeted insertions across a variety of species, but this method has not yet been employed widely. The second limitation is practical. Even in mice the generation of such targeted knock-ins is expensive, labor-intensive and slow. Furthermore, the maintenance of large mouse colonies is expensive, and breeding up the crossed mice can take several generations. Nevertheless, in spite of the challenges associated with generating and maintaining knock-in mice, their tremendous utility has driven a growing number of laboratories to switch at least partially to mice as their model organism.

The second strategy for driving transgene expression uses some reduced or “minimal” form of the endogenous promoter. These strategies have the advantage that the

transgene can be delivered virally, and therefore have the potential to work well in mammalian species other than mice. However, mammalian promoters are often very large, and the rules governing mammalian gene expression – i.e. how gene expression depends on promoters remain poorly understood. One strategy has been to express a transgene under the control of a “minimal” endogenous promoter – a promoter short enough (<10 kb) to fit into a virus such as adeno-associated virus (AAV) or lentivirus. Unfortunately, with some notable exceptions (e.g. hypocretin), this strategy typically does not recapitulate the expression pattern of the endogenous gene. Another clever strategy involves the use of short promoters from the puffer fish (*Takifugu rubripes*), an organism with a much more compact genome in which the regulatory sequences are shorter. Unfortunately, fugu promoters also failed to recapitulate mammalian expression patterns. Thus at present there is no general viral strategy for delivering transgenes to genetically defined neuronal populations in mammals.

D.1.2 Ribozyme mediated trans-splicing

The ideal approach to achieving cell-type specific expression would combine the specificity of knock-in transgenics with the convenience of viral delivery. We propose to develop such a technique based on ribozyme-mediated RNA trans-splicing. The key to our approach is that instead of coupling expression of the transgene to the promoter driving the endogenous gene (e.g. SOM) of interest, we move a step downstream, and couple the translation of the transgene directly to the mRNA transcript encoding the endogenous gene.

The core idea underlying our approach is to use ribozyme-mediated trans-splicing to couple an mRNA encoding a recombinase such as Cre or Flp into the mRNA of an endogenous gene such as somatostatin (SOM) (Figure D.1). The expression of the recombinase can then be used to switch on expression of an exogenous transgene such as GFP.

The trans-splicing ribozymes we propose to use are derived from the group I intron from *Tetrahymena thermophile*, a catalytic RNA (or ribozyme) with the ability to perform a cleavage-ligation reaction in the absence of proteins. Though its normal activity is to splice itself out of an mRNA transcript (joining the preceding and following segments into an uninterrupted transcript), it is possible to engineer the ribozyme to trans-splice (join part of one “donor” transcript to another “target” transcript). The

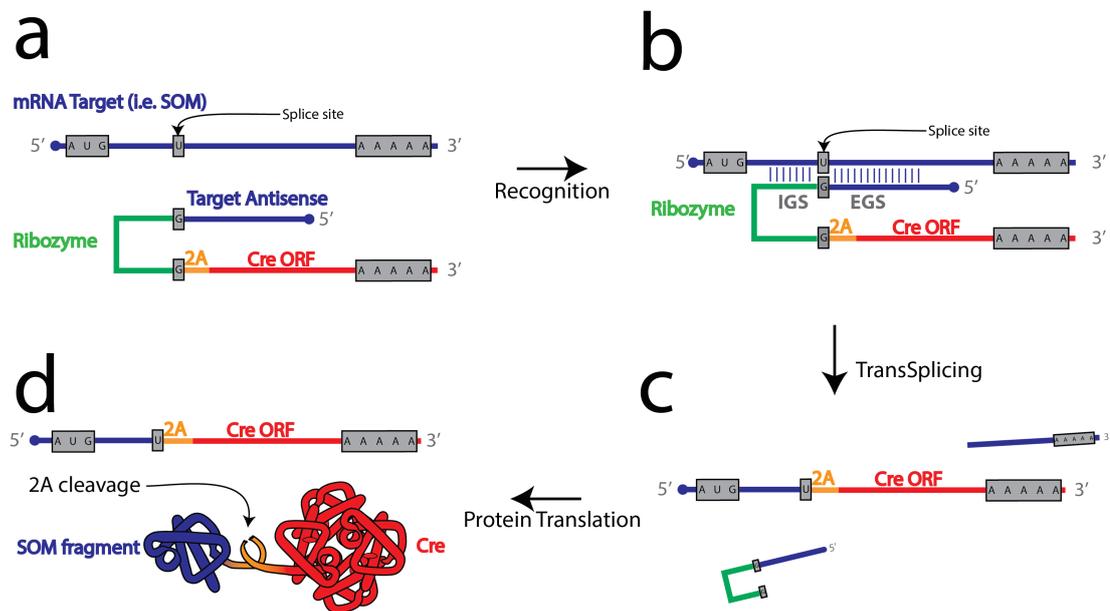


Figure D.1: Cell-type specific expression via trans-splicing - (a) A ribozyme is constructed that consists of the Cre open reading frame (1 kb), the 1A sequence (66 bp), the trans-splicing ribozyme (400 bp), and the target antisense (100 bp). (b) The ribozyme recognizes its target mRNA by complementary base pairing via the internal and external guide sequences (IGS/EGS). (c) Trans-splicing occurs at a conserved G:U wobble pair, ligating the donor transcript (Cre) to the acceptor target (SOM). (d) During protein translation, the 2A peptide sequence causes cleavage of the protein, resulting in a functional Cre protein and a cleaved SOM fragment.

tetrahymena group I intron recognizes its target via a 6bp complementary sequence known as the internal guide sequence (IGS). In the case of trans-splicing, the ribozyme is engineered to contain an IGS complementary to the target gene at the desired splice position. The only absolute requirement for splicing is an available uracil U in the target mRNA. Additional specificity, and efficiency, is achieved by adding a second complementary region (complementary to the target mRNA) known as the extended guide sequence (EGS). After recognizing its target, the ribozyme cleaves the target and ligates its cargo (the donor transcript, i.e. Cre) into the target transcript allowing translation of donor mRNA into a functional protein.

Figure D.1 summarizes our strategy for using ribozyme-mediated trans-splicing to couple an mRNA encoding a trans-activator into the mRNA of an endogenous target gene such as the D2R receptor. To limit expression specifically to cells which express a specific gene (i.e. SOM), we engineer the IGS and EGS sequences to target complementary sequences in the target mRNA transcript. The engineered ribozyme contains all of the necessary coding sequence of Cre, but is missing a translational start codon. In the absence of its target the ribozyme is expressed but its cargo is not translated; only upon trans-splicing into the appropriate target does it acquire the start codon necessary for translation. The resultant transcript includes a portion of the target gene, but this is co-translationally cleaved at a virally-derived cis-acting hydrolase element (CHYSEL) 2a sequence included in the ribozyme. The result is that the expression of functional trans-activator protein is conditional on the presence of a target mRNA.

Our strategy exploits the high specificity of ribozyme-mediated trans-splicing. The specificity of trans-splicing has been demonstrated in experiments with Diphtheria toxin A (DTA), a potent cytotoxin. When a trans-splicing ribozyme encoding DTA is targeted to a particular target mRNA, cells expressing the target are killed with high efficiency. This demonstrates the very high specificity of the trans-splicing reaction, since even a very low level of off-target DTA trans-splicing would reduce cell viability. Similarly, trans-splicing ribozymes have been engineered to correct mutations in beta globulin transcripts responsible for sickle cell anemia and were able to discriminate mRNAs differing by only a single base (wt vs. mutant). Thus trans-splicing has the potential to provide the specificity needed to target a trans-activator in a cell-type specific manner.

Using a trans-activator allows breaking the problem of cell-type specific expression of any gene into two components – cell-type specific trans-activator expression and activator-dependent transgene expression. This has two advantages compared with expressing the transgene directly from the locus of the endogenous gene (e.g. expressing GFP directly under the control of the SOM promoter). First, because the trans-activator acts as a switch, the expression level of the transgene is decoupled from the expression level of the endogenous gene (e.g. the D2R receptor) for which it is a marker; expression of the trans-activator need only surpass the threshold sufficient to activate the switch, and the expression of the transgene can be driven by a strong promoter. Thus robust expression of a transgene coupled to a particular promoter can be achieved, even in a neuron for which that promoter is only weakly active. The second advantage is combinatorial: there is no need to generate and maintain a separate transgenic mouse line for each combination of expression pattern and transgene, since novel combinations can be produced by a single generation of breeding. Thus N trans-activator-dependent transgenes and K recombinase knock-in mouse lines can yield potentially $N \times K$ distinct mice. A substantial fraction of the $> 10^4$ gene expression patterns mapped in the Allen Brain Atlas are of potential interest, and there are a growing number of useful transgenes, including variants of channelrhodopsin (ChR2), halorhodopsin, GFP, etc; generating a new mouse for each possible combination is not feasible. The use of trans-activators reduces the number of mice needed to at most $N + K$ instead of $N \times K$.

Here, we employ recombinases as transactivators. Recombinases are enzymes which catalyze the recombination of DNA between pairs of specific DNA sequences called recombination sites. Several recombinases have been used in mammalian neurons, including Cre and Flp. In both Cre and Flp, the recombination sites consist of specific sequences of 34 nucleotides (called loxP and FRT sites, for Cre and Flp, respectively). FRT and loxP sequences are very different; Cre does not act at FRT sites, nor does Flp act at loxP sites. Recombinases can be used to render the expression of transgenes of interest such as GFP or ChR2 conditional upon their presence by means of a transcriptional “stop” cassette flanked by recombination sites placed between the promoter and the transgene. The stop cassette prevents transgene expression unless it is excised by the recombinase, in which case the transgene is expressed. Conditional expression can also be achieved by flip-excision (FLEX). Transgene expression thus depends on

the logical and of the recombinase and the appropriately engineered transgene with properly placed recombination sites.

This approach has at least three potential advantages over the conventional transgenic mouse approach in which Cre is knocked in to a given genetic locus (e.g. the D1R locus). First, generating a viral trans-splicing ribozyme is both faster (weeks vs. months/years) and less expensive (<\$3K vs. >\$20K) than generating a knock-in mouse. Second, with the viral approach, different subpopulations of neurons can be targeted with different constructs in the same animal. For example, SOM expressing neurons could be targeted with Cre/ChR2, whereas PV expressing neurons could be targeted with Flp/halorhodopsin in the same animal, thereby enabling bi-directional control of neuronal activity within a single brain region. Finally, this approach is not limited to mice. Indeed, this strategy has the potential to transform research on model organisms such as rats and primates by making it possible to target reporter genes to specific neuronal subtypes in these preparations, and when combined with optogenetic approaches may even be useful in the treatment of human neuropsychiatric diseases.

D.2 Results

D.2.1 Cre recombinase

Our approach requires that the Cre ORF not be translated until the Cre transcript acquires a start codon (ATG) through trans-splicing. Unexpectedly, initial experiments revealed that functional Cre protein was produced even in the absence of the first ATG. We hypothesized that Cre translation was initiating downstream of the first ATG, perhaps at the second ATG. To test this hypothesis we made a construct termed CreM2S (Cre initiating on the second Methionine), which was an N-terminal truncation of Cre, starting at the second in frame ATG. This construct failed to express a functional Cre. Thus we reasoned that Cre translation is initiating somewhere between the first and second ATG sequence. Using a binomial search algorithm, we designed a series of truncated Cre sequences to identify a sequence in which the expression of functional Cre required the addition of an ATG start codon (Figure D.2A). We co-transfected the Cre variants into HEK293 cells with a reporter plasmid (Figure D.2B). The reporter contains a constitutively active mCherry expression cassette, and a GFP expression cassette that was dependent on the activity of Cre. We employed flow cytometry to

measure the activity of Cre from transfected cells – measuring the fraction of mCherry+ cells that were also GFP+. Two truncations, CreM1.5 and CreM1.75, showed little or no activity in the absence of a start codon but were fully active upon addition of an in-frame ATG – CreM1.5S and CreM1.75S (Figure D.2C,D). We selected CreM1.75 for all subsequent experiments.

D.2.2 Construction and testing of a SOM-targeting Ribozyme

We first attempted to target the somatostatin transcript with a ribozyme carrying the CreM1.75 ORF as its 3' donor transcript. A uridine towards the 5' end of the SOM transcript, in frame with the protein coding sequence, was chosen at random – 10U. A 9 nucleotide IGS sequence was designed flanking the target U, 10U, by 5 nucleotides in the 5' direction and 3 nucleotides in the 3' direction. In general, for detectable activity in mammalian cells, ribozymes also require a 5-terminal extension – the extended guide sequence (EGS) [Sullenger and Cech 1994]. Therefore, we added 105bp of sequence complementary to the SOM mRNA at the 5' end of the ribozyme IGS.

We tested this ribozyme construct via transient transfection into HEK293 cells, with or without the target transcript. 48 hours after transfection, RNA was isolated and subjected to RT-PCR to detect the spliced product (Figure D.3A). We detected the trans-spliced transcript, Som-Cre, in samples where the target and ribozyme had been co-transfected (Figure D.3B). Importantly, no product was detected in samples that had been separately transfected with the Ribozyme or target and then mixed during lysis (Figure D.3B), indicating that the trans-splicing reaction took place *in vivo* and not during the biochemical manipulations performed *in vitro* for detection.

Having detected successful trans-splicing *in vivo* we next tested if the trans-spliced Cre could mediate recombination on a co-transfected reporter (Figure D.2B). Unfortunately, we found no difference in the Cre-dependent reporter expression with or without the target RNA transcript. Moreover, the overall levels of Cre activity were extremely low, perhaps suggesting inefficient splicing.

It is known that the specific uridine chosen on the target can influence splicing efficiency. Although every uridine within the target RNA is a potential splice site, the accessibility of the uridine, in theory, determine the efficiency of trans-splicing [Ayre et al. 2002, Jones et al. 1996, Jones and Sullenger 1997]. Therefore, we set out to map the most accesible uridines on the SOM transcript via an exhaustive *in vitro* screen.

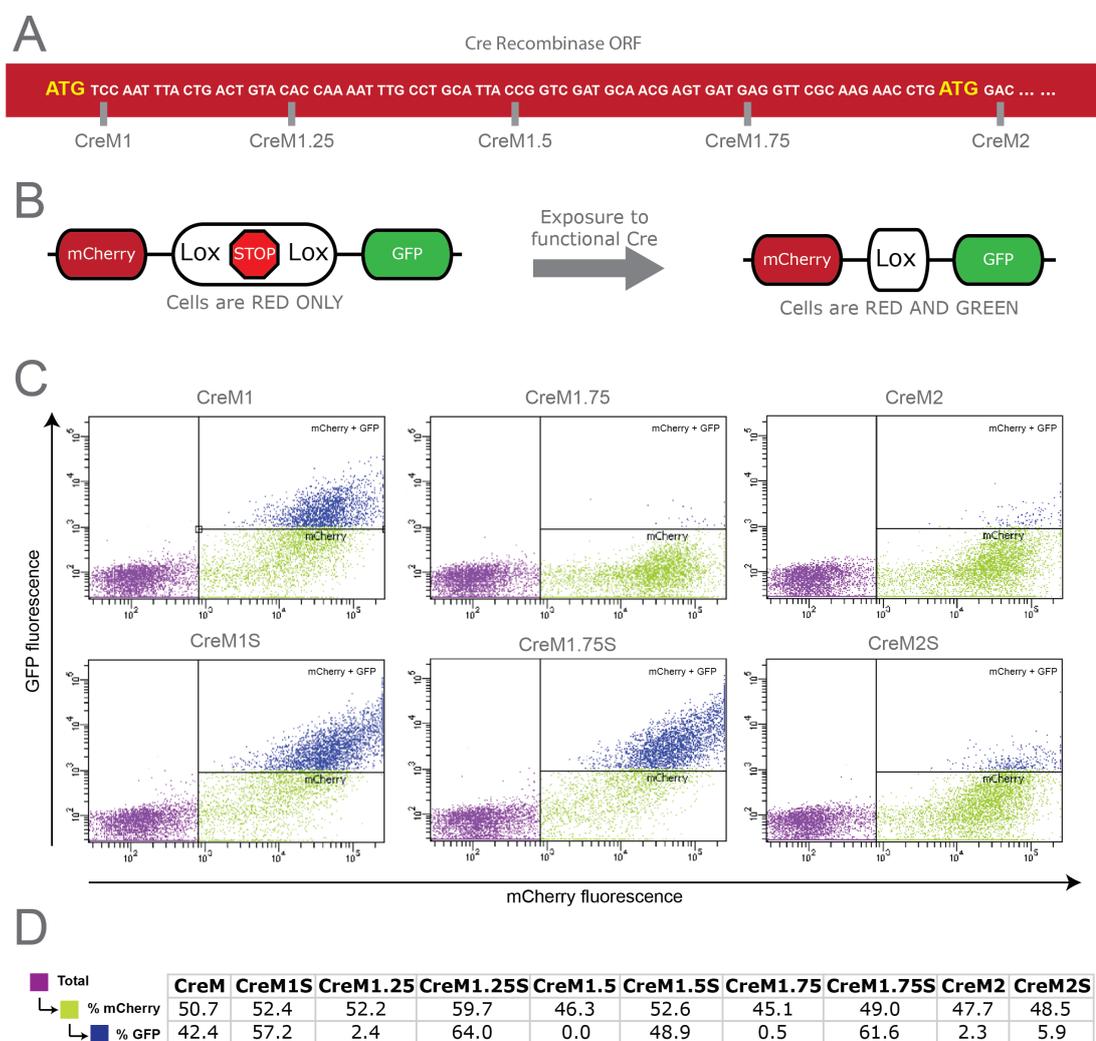


Figure D.2: Function of Cre truncations - (a) The first 87bp of Cre coding sequence are shown, with the positions of the truncations denoted. (b) A reporter plasmid was constructed such that transfection resulted in mCherry expressing cells. Exposure of the plasmid removed a stop cassette between the mCherry open reading frame and the GFP open reading frame. (c, d) Flow cytometry results for co-transfection of the reporter plasmid and the Cre variant plasmids. (S) denotes addition of a Kozak-ATG.

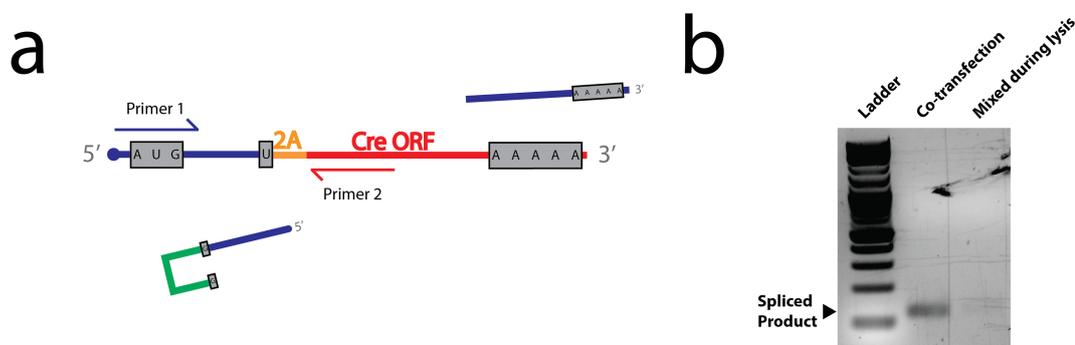


Figure D.3: Trans-splicing *in vivo* - (a) The ribozyme ligates the Cre transcript onto the Som target transcript. The spliced product can be detected by RT-PCR with primers that bind in the target (SOM) and cargo (Cre). (b) Trans-splicing is detected when target and ribozyme are co-transfected.

D.2.3 Identification of efficient trans-splicing sites on target mRNA

In order to determine which uridines on the target transcripts (i.e. SOM, PV) are most accessible for splicing, we used a uridine mapping strategy [Lan et al. 1998]. Briefly, total RNA isolated from mouse brain was incubated with a ribozyme library containing randomized internal guide sequence (IGS). Because the IGS is only 6 nucleotides, the entire space – $4^6 = 4096$ combinations – can be explored easily via DNA synthesis. The trans-splicing products are amplified by RT-PCR with primers specific for the 5' end of somatostatin and the end for CreM1.75 located at 3' end of ribozyme. The amplified splicing products were then Topo-cloned and sequenced to identify accessible uridines. The uridine at position 446 of SOM mRNA appeared to be the most accessible, as 5 of 11 clones sequenced contained this splice site. Therefore we engineered the ribozyme, aSOM-446Rz-CreM1.75 to contain the corresponding IGS to target the 446U on the SOM mRNA.

D.2.4 Addition of an extended guide sequence

The IGS sets the region of the SOM transcript that can be targeted for additional binding via an extended guide sequence. This sequence, the EGS, is important for stabilizing the binding between the mRNA target and the ribozyme to permit efficient

trans-splicing, and is particularly necessary *in vivo* [Kohler et al. 1999]. While the position of the EGS is set by the IGS, the optimal length must be empirically determined. A series of ribozymes containing different lengths of the EGS (56bp, 96bp, 145bp) were generated to identify the optimal length, but minimal differences in efficiency were detected. It has previously been postulated that a longer EGS can result in non-specific cleavage [Herschlag 1991] and may stimulate dsRNA mediated degradation pathways. Therefore, in the absence of a detectable difference in efficiency, we choose the shortest EGS sequence tested and constructed the ribozyme aSOM-446Rz56E-CreM1.75 for further testing.

D.2.5 Trans-splicing in stable cell-lines

After the optimal ribozymes were identified, we next examined the efficiency and specificity of these ribozymes *in vivo* in mammalian cells (HEK293). Because the target mRNA expression level could affect the efficiency of trans-splicing, we established stable cell lines expressing SOM to keep the expression at a constant level. To examine the efficiency and specificity of trans-splicing, ribozymes were co-transfected with the Cre-dependent reporter into either SOM- HEK cells or stable cell lines expressing SOM (SOM+). The efficiency of ribozymes was determined by the percentage of cells expressing reporter genes in SOM+ cells, whereas the specificity of ribozymes was determined by comparing the percentage difference of reporter gene expression between SOM+ and SOM- cells. The percentage of cells expressing reporter genes was quantified by flow cytometry. This SOM targeting ribozyme induced expression in 0.4

D.3 Discussion

Our results are preliminary, but highlight the possibility of using a new approach to induce reporter gene expression in specific cell types expressing a particular gene via ribozyme-mediated trans-splicing. Further work will be needed to increase the efficiency of the reaction, without sacrificing specificity.

Simple theoretical considerations clarify the challenge of achieving specificity and efficiency. Trans-splicing is a bimolecular reaction between the ribozyme and a target mRNA, the result of which is the trans-spliced product (in our case a recombinase with a start codon) and an RNA fragment. To first order, the amount of trans-spliced

product depends on (1) the concentration (i.e. expression level) of ribozyme; (2) the concentration (i.e. expression level) of mRNA; and (3) the probability with which splicing occurs when the ribozyme encounters a given mRNA – ignoring here the other potentially important determinants, such as the cellular localization of RNA. Specificity is defined as the concentration of trans-spliced product (recombinase) resulting from encounters with the target mRNA, compared with the concentration of recombinase resulting from encounters with all other (off-target) mRNAs in the cell. Efficiency is defined by the fraction of ribozyme that successfully undergoes trans-splicing. Since we have no control over the concentration of either the target mRNA or of all other potential off-target mRNAs – these are defined by their expression levels in each neuronal subtype – the opportunities for engineering are limited to (1) the expression level of the ribozyme and (3) the probability of splicing given an encounter between the ribozyme and an mRNA.

The expression level of the ribozyme, like that of any expressed RNA, is determined in large part by the promoter. As suggested by the simple model above, increasing the concentration of the ribozyme leads to increased trans-splicing, but may come at the cost of higher off-target (non-specific) splicing. Using strong promoters such as CAGS, it is possible to achieve high levels of expression in the brain with AAV. Moreover, titration of expression levels can be achieved with inducible systems such as the Tet-off system in which doxycycline delivered intraperitoneally is used to reduce the expression of ribozyme. In this way the expression of the ribozyme can be tuned to a particular target without making new constructs under different promoters.

Manipulation of the probability of splicing is achieved by engineering of the IGS and EGS. We optimized the IGS sequence using an *in vitro* screen, which may highlight different accessibility than exists *in vivo*. Repetition of this screen *in vivo* should aid in determining the most accessible uridines on the mRNA target in a more relevant context.

Finally, efficiency is likely a product of expression levels and time. Given more time, the ribozyme should convert more target mRNA to spliced product, resulting in accumulated protein levels of Cre recombinase, and thus reporter gene. In our experiments in HEK cells, it was difficult to propagate the cells for more than 72 or 96 hours. Perhaps longer expression times will increase the efficiency of the reaction.

This method builds on established ribozyme trans-splicing technology to achieve cell-specific expression of transgenes without the need to generate a knock-in transgenic. Improvements to the technique may allow for multiplexed cell-type specific targeting across a variety of organisms with the simplicity and cost-effectiveness of a viral infection.

D.4 Materials and Methods

D.4.1 Trans-splicing

In vitro trans-splicing was performed as described [Jones et al. 1997].

D.4.2 Flow Cytometry

Immediately before flow cytometry, cells are washed 2X with PBS, trypsinized, resuspended in DMEM and spun down for collection. The cell pellet is washed once by resuspending in 1mL of PBS and spun down for collection. The pellet is resuspended in 4% PFA and fixed at 25°C for 15 min. Cells are spun down, resuspended in PBS, and filtered through a cell strainer. Flow cytometry was performed on the LSR-II (BD).

D.5 Acknowledgements

Huiqing Zhan and Diana Gizatullina, helped tremendously with this project. Diana helped with all of the work in HEK cells – from cloning to transfection to cell counting. Huiqing Zhan improved the initial designs to improve efficiency (in vitro screens) and performed all of the work on stable cell lines.

Appendix E

Sparse transcriptome profiling via nonspecific trans-splicing

E.1 Introduction

Heterogeneity plays an important role in the proper and abnormal functioning of many biological systems from neural computation to tumor development. Methods of probing heterogeneity in cellular expression have been limited by technical challenges (i.e. sorting and separating individual cells, isolation of sufficient quantity and quality RNA from individual cells, etc). A method for rapidly profiling large numbers of cells, while retaining identities of individuals, within a population would be a major advance. Here we describe a novel method for tagging transcripts within individual cells of a heterogeneous population by hijacking the cellular splicing machinery.

Previously, researchers have attempted to use the spliceosome to splice RNA molecules in trans. This technique, known as Spliceosome Mediated RNA Trans-splicing (SMRT), has the potential to target individual RNA molecules with the purpose of repairing mutated transcripts and/or delivering targeted therapeutics (i.e. cytotoxins to kill cancerous cells). However, the technique is plagued by promiscuity of the splicing events – expression of an intron sequence followed by an exon causes transsplicing to occur between the artificial exon and many cellular transcripts [Kikumori et al. 2001]. We reasoned that, combined with cellular barcoding this lack of specificity could be exploited to stochastically tag actively transcribed mRNA via SMRT.

In the mammalian genome, many mRNA molecules are initially transcribed as pre-mRNA with the coding sequence interrupted by one or more intron (Figure E.1A). Overexpression of an intron-Barcode (Figure E.1B) causes replacement of the 3' exon of a target mRNA molecule with a unique cell-identifying barcode via trans-splicing (Figure E.1C). In this way, sparse expression profiles of individual cells within a population can be identified by high-throughput sequencing.

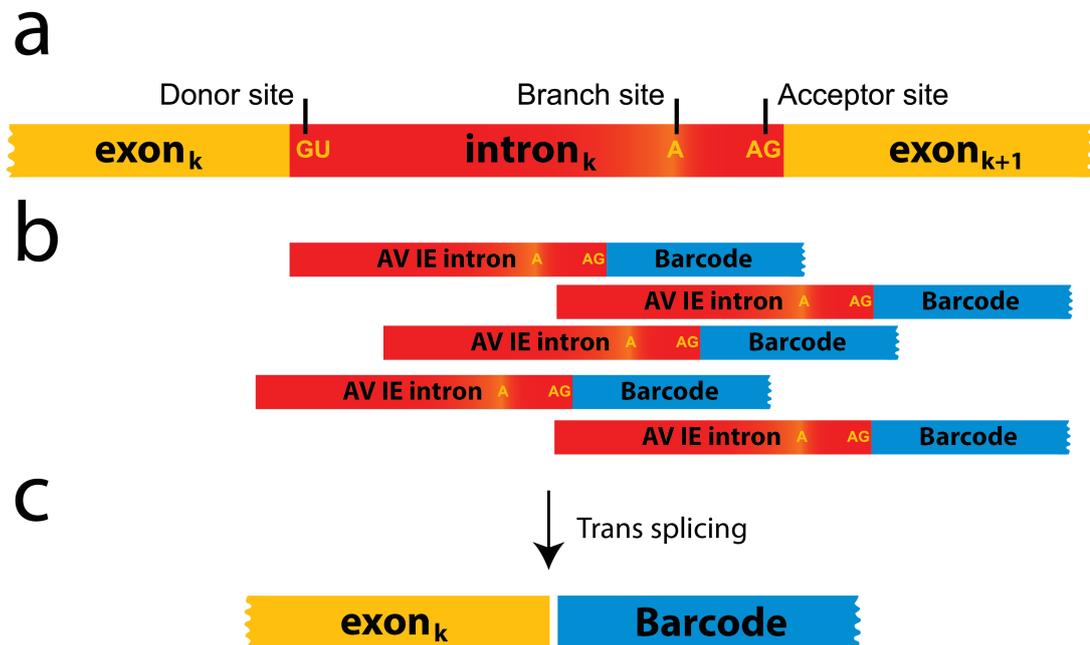


Figure E.1: Tagging cellular transcripts via promiscuous trans-splicing - (a) A typical pre-mRNA containing several exons and introns. (b) Adenovirus IE gene introns fused directly to a barcode are exogenously expressed at high levels. (c) The spliceosome catalyzes the trans-splicing of the barcode into endogenous transcripts, thereby tagging endogenous transcripts with a cell-identifying barcode.

E.2 Results

We tested whether SMRT-mediated barcode tagging of mRNA transcripts can occur *in vivo* in mammalian cells. We separately co-transfected HEK293 cells with INTRON-BC1 and GFP (with an intron) or INTRON-BC2 and mouse SST (with an intron) and harvested cells 48 hours after transfection. The cells were mixed together before lysis

to emulate a heterogenous population. RNA was then Trizol extracted, reverse transcribed into cDNA, and PCR amplified adding adapters for Illumina high-throughput sequencing. We expected to find many instances of the GFP transcript tagged with BC1 and SST tagged with BC2 but not the opposite (GFP tagged with BC2 or SST tagged with BC1). Using bowtie2 we mapped reads to GFP and SST and found that 4046 reads of BC1 aligned to GFP, while only 10 reads aligned to SST (as expected). For BC2 we found 5170 reads map to SST and only 3 reads to GFP (Figure E.2) Additional analysis showed that splicing occurred to many types of RNA, the largest target of which was mRNA (Figure E.3a). The splicing occurred to transcripts throughout the genome (Figure E.3b) and, in many cases, occurred at sites near an intron (Figure E.3c). There was a high correspondence between cellular transcripts tagged in BC1 and BC2 and, encouragingly, among the most abundant transcripts were some that were cell-type specific (Table 1).

E.3 Discussion

The results shown here are preliminary, but highlight the possibility of using this simple approach for sparse sampling of many cellular transcription profiles, with single cell resolution.

At present, it is unclear if the sparse transcript profiles obtained via this method will be sufficient to identify cell-types. Typical cells contain as many as 100,000 mRNA molecules (Islam et al. 2011), but most of these transcripts are not distinct. Fortunately, genes that are differentially expressed between neuron types are generally among the most abundant transcripts (Mellen et al. 2012). That is not to suggest that the transcripts that are most informative for cell-type identification are necessarily the same transcripts that cause differentiation. For example, transcription factors which drive gene expression and largely define the cell's transcriptome are expressed at low levels. While it may not be possible to identify cell-types *de novo* in this way, it should be possible to use this sparse information to "call" cell-types based on a reference set of known transcription profiles.

In the context of the connectome project (see chapter 3), cellular transcripts tagged with the same barcodes used for circuit mapping may allow integration of cell-type information into the connectivity matrix.

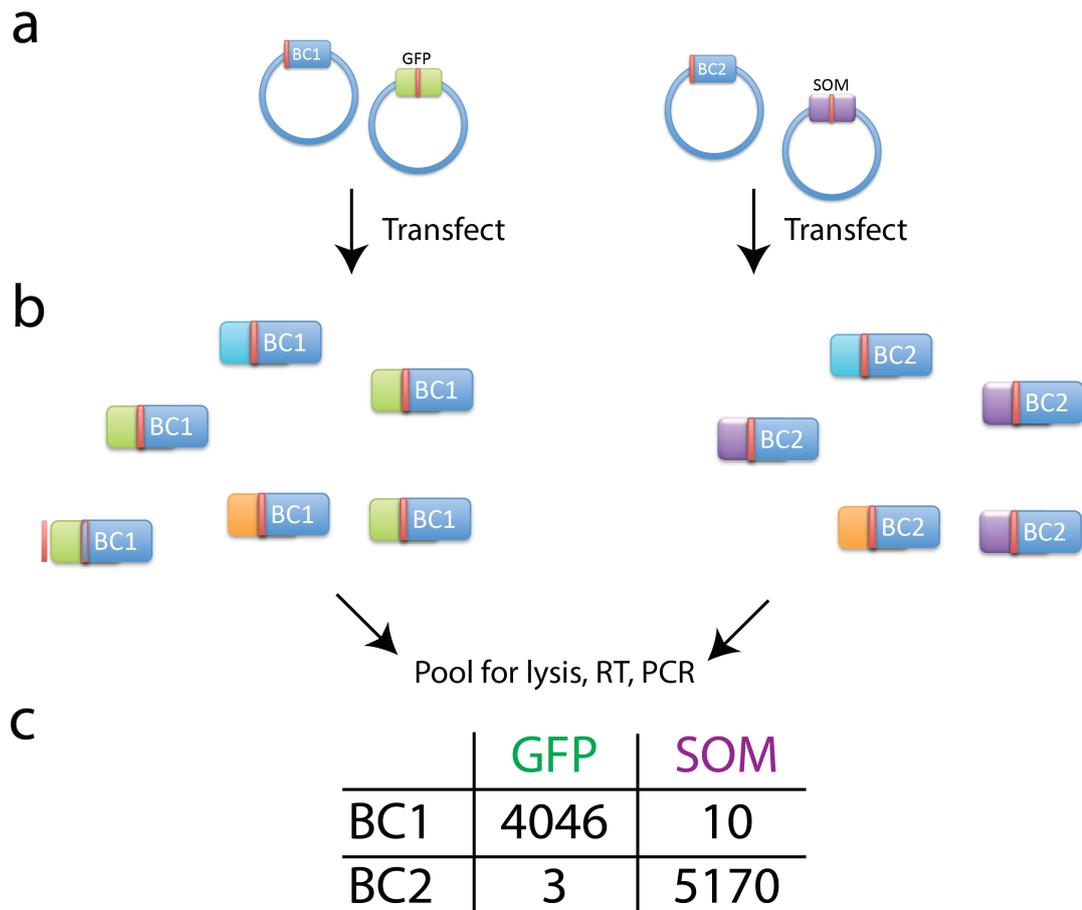


Figure E.2: Barcode trans-splicing to defined transcripts - (a) A splicing cassette (Intron-BC1) was co-transfected with GFP-INTRON-GFP in one population of cells and a splicing cassette (Intron-BC2) was co-transfected with SOM-Intron-SOM in a separate population of cells. (b) Trans-splicing labels random transcripts with BC1 or BC2, including the co-transfected transcripts GFP and SOM, respectively. (c) The cells are pooled and lysed for RT-PCR and the barcode splicing results to the control transcripts are shown.

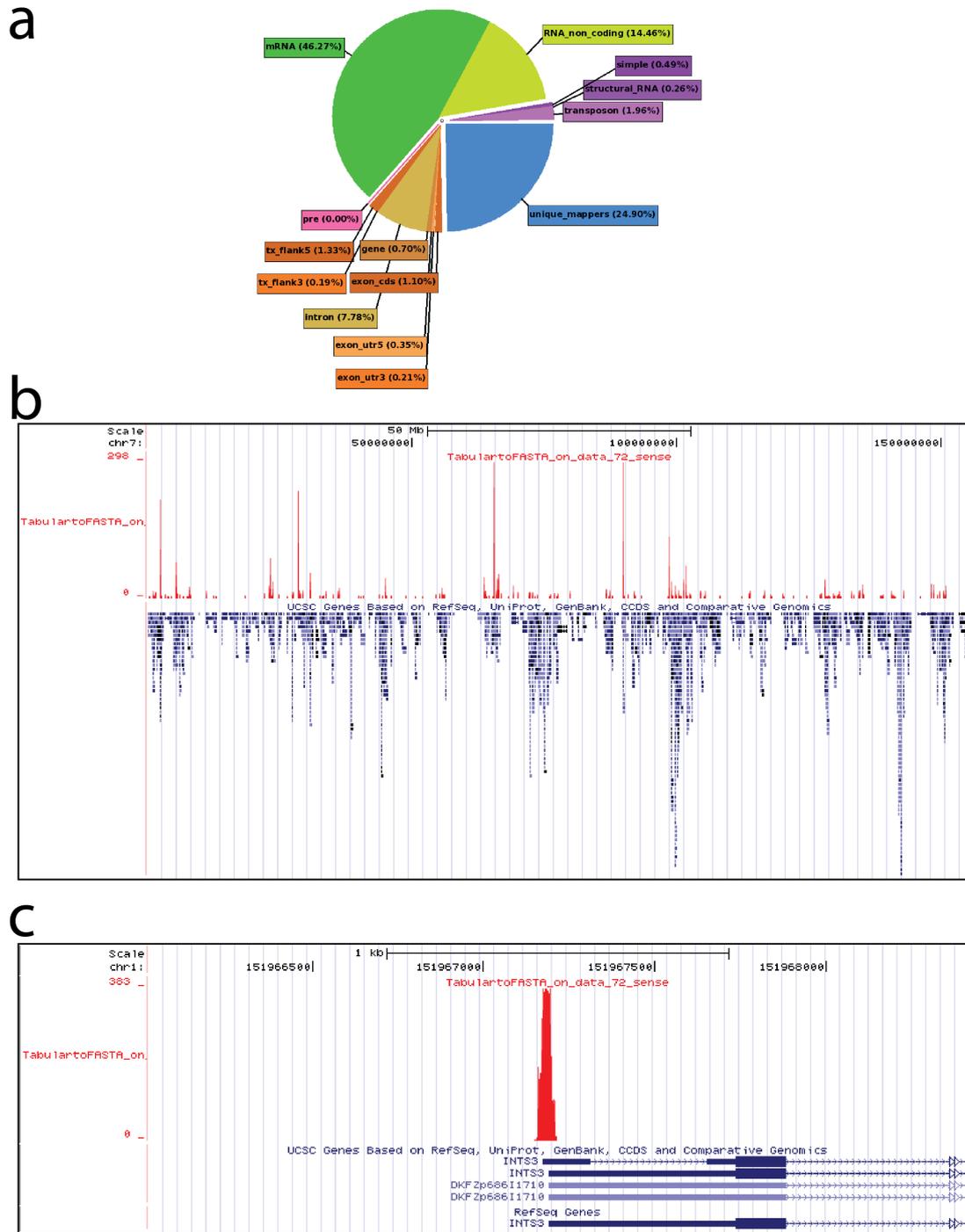


Figure E.3: Barcode trans-splicing genome wide - (a) The barcode transplices into a variety of RNA molecules (b) across the genome. (c) Many times, the splicing event occurs near a defined intron.

E.4 Materials and Methods

E.4.1 Design of splicing cassette

The splicing donor consists of the Adeno Virus immediate early intron followed by a 3 splice site (5'-CAG-3'), although it is expected that any intron will work for this purpose. The sequence of the intron is given below. Immediately following the 3 splice site is a unique barcode sequence. This barcode sequence can be a static single barcode, or can consist of a high-diversity barcode library generated by a variety of means including but not limited to shotgun cloning of oligonucleotides, in vivo barcode generation by a recombinase, etc. Following the barcode, additional known sequence elements are present to aid in reverse transcription, and high-throughput sequencing library preparation.

ADENO VIRUS IE Intron: GGCCTACTTATCCTGTCCCTTTTTTTTCCACAG

E.4.2 Mapping and analysis

Mapping to the genome or to the artificial transcripts GFP and SOM (which were separately transfected) was performed using Bowtie2. Additional analysis was performed in Galaxy. All genome browser plots were generated from the UCSC genome browser.

E.5 Acknowledgements

Thank you to Peter Znamenskiy and Kaja Wasik, who both worked with me on this project. And thank you to Huiqing Zhan, who started this project by first employing SMRT for targeted trans-splicing.

Appendix F

Color by number Brainbow

F.1 Abstract

Traditional large-scale projection mapping techniques group large populations of neurons together as a single functional unit. Heterogeneity of projection targets, however, exists even within highly spatially restricted cell populations. Brain-wide mapping of projection patterns with single-neuron resolution will allow for the stratification of cells based on their downstream targets – a characteristic which is likely to correlate with other functional properties. We have developed a method that marries DNA barcoding of individual neurons (BOINC) with readout via fluorescent in situ sequencing (FISSEQ) to trace the projections of several hundred neurons with single-neuron resolution. At scale, this method – effectively an infinite-color, or digital, Brainbow – will allow brain-wide, multiplexed tracing of millions of neurons across dozens of injection sites, within a single brain. High-resolution reconstructions of single-neuron projections will have a profound impact on our ability to dissect neural circuits.

F.2 Introduction

The cortex has a striking laminar organization. To a first approximation, excitatory cells in each layer have characteristic morphological and physiological properties. They also make specific local and long-range connections, suggesting that excitatory neurons within a layer constitute a class – inhibitory neurons are classified separately. According to the “canonical cortical circuit” hypothesis [Douglas and Martin 2004, Oviedo et al. 2010], information from the thalamus flows from cortical layer 4 to 2/3 to 5 and thence to other cortical and subcortical targets, with thalamic feedback arising in layer 6. This hypothesis, though useful, is based on a highly simplified model of neocortex as discrete modules defined by layer identity. In reality, cellular heterogeneity within each layer may be critical in establishing the functional diversity of computation both within and across cortical areas. For example, two classes of projection neurons in mouse layer 5b can be distinguished by distinct target regions (bilateral striatum only vs. ipsilateral striatum and subcortical structures), distinct physiology (single vs. bursting action potentials) and distinct morphology (small vs. large cell bodies) [Hooks et al. 2013]. However, these cells are not yet separable in most tracing experiments due to their juxtaposition. Similarly, some neurons in mouse barrel cortex project to primary motor cortex, whereas others project to secondary somatosensory cortex [Yamashita et al. 2013] likely with profound functional consequences.

Traditional projection tracing techniques rely on viral expression of fluorescent proteins within a defined group of cells (spatially, genetically, or both). These techniques have revealed bulk projection patterns, which have been important in forming our current understanding of neural circuit function [Oh et al. 2014]. However, these techniques lack single-cell resolution, which is - in many cases [Hooks et al. 2013, Yamashita et al. 2013] – critical for understanding the functional differences between juxtaposed cells. Brainbow [Livet et al. 2007], permits the tracing of single-cells with relatively high-throughput via the combinatorial expression of a set of fluorophores – yielding a palette of about a hundred unique colors. Here, we expand the potential of Brainbow for neuronal tracing, by employing unique nucleic acid sequences, rather than unique colors. The result is an infinite expansion of Brainbow. A string of 30 nucleotides yields $4^{30} = 10^{18}$ unique barcode sequences – far more than enough to label every one of the roughly 10^8 neurons in the mouse brain [Roth and Dicke 2005].

F.3 Results

The method we have developed, fluorescent *in situ* barcoding of individual neural connections (FIBOINC), utilizes fluorescent in situ sequencing (FISSEQ) [Lee et al. 2014] to read individual RNA barcode molecules at sites distant from the cell soma. Briefly, RNA barcodes are tethered within the axonal compartment via modifications to the pre-synaptic protein Neurexin1B (Figure F.1A). The brain is thinly sliced and subjected to FISSEQ, which involves reverse transcription of the mRNA barcode into cDNA, circularization of the cDNA molecules via single-strand DNA ligation, and amplification of the cDNA by rolling circle amplification to yield rolling circle colonies – rlonies (Figure F.1B). The resulting rlonies are sequenced via the sequential imaging of oligonucleotide-conjugated fluorophores during sequencing-by-ligation (Figure F.1C) to reconstruct cell-identifying barcodes for axonal tracing (Figure F.1D) [Lee et al. 2014].

In order to ensure high-concentrations of RNA barcodes in distant processes, we employed the pre-synaptic protein Neurexin1B (Nrx1B), which has been modified to bind RNA barcodes for tethering in the axonal compartment (Figure 2A). This was accomplished via the fusion of the 22 amino acid RNA binding protein λ N to the cytoplasmic domain of Nrx1B, and the engineering of RNA barcodes to contain four repeats of the λ N cognate RNA binding motif, the 15 nucleotide boxB hairpin. This interaction was previously shown, in HEK cells, to be specific and efficient (see chapter 3, Figure 3.9A,D). During the first step of FISSEQ, any fluorescence is quenched, and does not interfere with the subsequent imaging required for sequencing.

We packaged the two components (protein and RNA barcode) into a single Sindbis vector, which allows high expression levels and rapid expression timing. A barcode library [Gerlach et al. 2013, Golden et al. 1995, Lu et al. 2011], reaching diversities of $> 1M$ – was produced to allow the unique labeling of individual cells. We introduced the barcoded virus into the left auditory cortex. The virus expresses a unique barcode in each cell as well as the modified RNA-binding Nrx1B protein. Forty-eight hours after infection, the brain was fixed before thinly sectioning (10 μ m) on a cryostat using the tape-transfer (Cryojane, Leica) method (Figure F.2A). Slices were placed individually on cover-glass for FISSEQ. Imaging of rlonies detected the barcode RNA

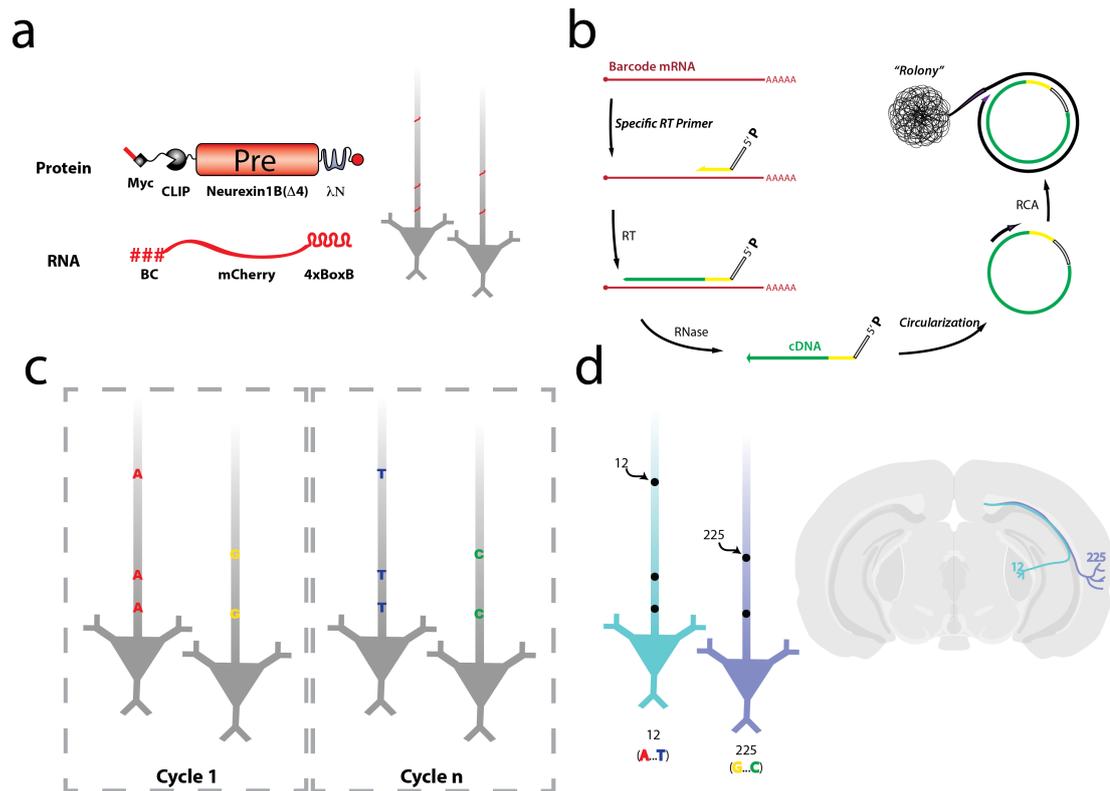


Figure F.1: Overview of the FISSEQ neuronal tracing - (a) An RNA barcode transcript and an engineered pre-synaptic protein are expressed via viral transduction. The presynaptic protein has been modified to bind to the RNA barcode, thereby preferentially localizing the barcode transcripts to the axonal compartment. (b) The barcode mRNA is reverse transcribed into cDNA, the cDNA is circularized, and the circularized cDNA is amplified by rolling circle to form a FISSEQ "rolony" *in situ*. (c) Rolonies are sequenced via sequential imaging of oligo-conjugated fluorophores during sequencing-by-ligation. (d) Barcodes are identified by sequence, allowing axons to be unambiguously traced throughout the slice irregardless of the distance from the soma.

(Figure F.2B), which had traveled throughout the slice, with many rolonies found distant from the injection site (Figure F.2C). Little labeling was detected in a non-injected slice (Figure F.3). The labeling of rolonies probing for circularized DNA is equivalent to tracing of a single fluorophore. We are currently in the process of sequencing and analyzing data from several slices for the long-range tracing of neuronal processes.

In addition to barcode sequencing, FISSEQ has the capability to capture and sequence endogenous transcripts within a cell. This capability, when combined with FIBOINC, will allow for simultaneous tracing of projections (or even connectivity, see subsection 4.4.1) and cell-type identification. We sliced and sequenced a wild-type brain slice to determine the transcription profile in a region of the auditory cortex. Briefly, transcripts were fixed, reverse transcribed, and amplified to form rolonies (Figure F.4A) which were sequenced by sequential cycles of ligation (Figure F.4B). After mapping to the mouse genome, we observed $\sim 60,000$ rolonies – of which 20,000 mapped to an mRNA sequence (the remaining derived from rRNA) – throughout the slice. Importantly, many of the most abundant genes are **neuron-specific**. We are currently analyzing the data to determine the extent to which gradients of expression exist in this region of the brain.

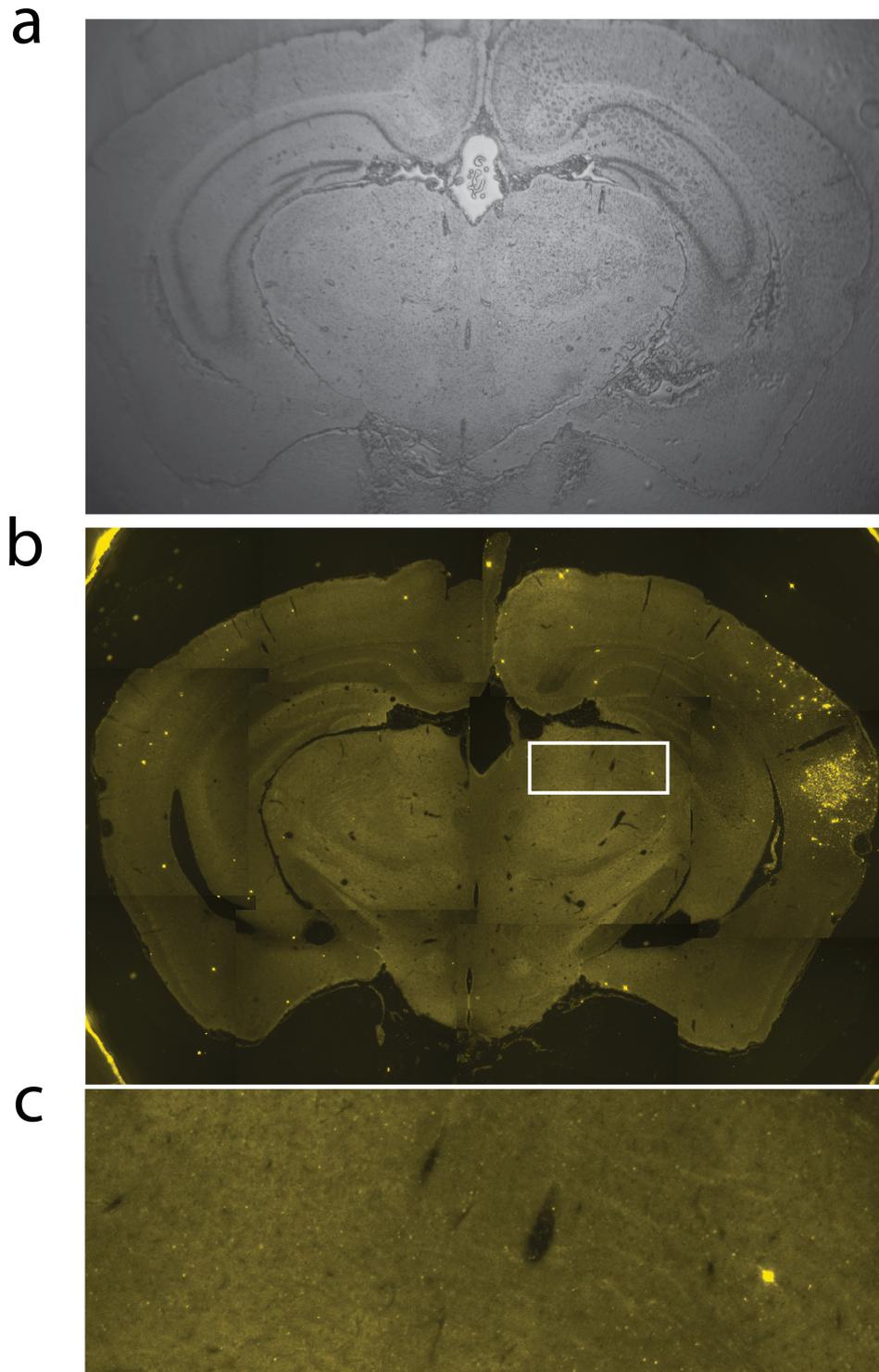


Figure F.2: Barcode FISSEQ in a coronal brain slice - (a) A typical cryo-slice using the tape-transfer system. (b) Rolony imaging in a single slice after rolling circle amplification of barcoded RNA transcripts. (c) Zoom-in on rolonies distant from the injection site (subcortical).

Table F.1: FISSEQ gene hits from coronal mouse brain sample

| Gene | Read Count | Gene | Read Count |
|--------------|------------|--------------|------------|
| Rn28s1 | 9588 | Camk2a | 68 |
| Rn45s | 6495 | Cmip | 68 |
| Rn45s | 846 | Nf1 | 66 |
| Lars2 | 556 | Lars2 | 61 |
| Malat1 | 467 | Calm2 | 60 |
| Gm20594 | 381 | Calcoco2 | 59 |
| Chrna4 | 210 | Map1b | 59 |
| C2cd5 | 181 | Pnck | 59 |
| Habp2 | 151 | Ncf2 | 58 |
| Rn28s1 | 151 | Synj2bp | 57 |
| Meg3 | 141 | Itsn1 | 56 |
| Rn18s | 138 | LOC102635337 | 53 |
| Mgmt | 117 | Sparcl1 | 53 |
| Rn18s | 117 | Atp1b1 | 52 |
| LOC102638644 | 102 | Ghrl | 51 |
| Rn45s | 94 | Top1mt | 51 |
| Esrrg | 93 | Kcnq1ot1 | 49 |
| Fut11 | 91 | LOC102639444 | 47 |
| Slc1a2 | 86 | Calca | 45 |
| Atp1a3 | 73 | Calm1 | 44 |

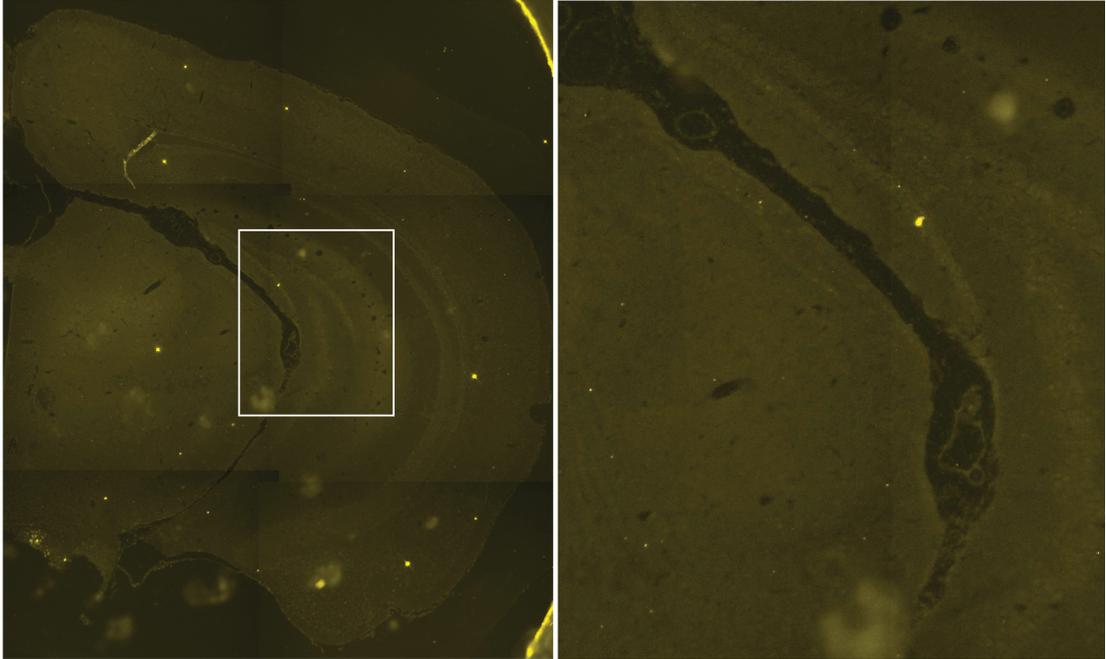


Figure F.3: Negative control - unbarcoded slice after FISSEQ -

F.4 Discussion

This method, in its current form, can be applied to several specific brain regions to track innervations from a relatively small (~ 100 s) population of barcoded cells. However, the method can be easily scaled to trace the projection patterns of orders of magnitude more cells throughout the entire brain while maintaining single-cell resolution. Transgenic techniques, combined with an *in vivo* barcoding scheme [Peikon et al. 2014] will allow for scaling this technology to mapping the projection patterns of all cells within a single brain.

The projection map alone will likely be insufficient for a full understanding of neuronal circuit architecture and/or function. However, a projection map will provide a framework for disentangling the heterogeneity of brain areas, testing the roles of important upstream/downstream brain areas for a given region of interest, and/or integrating knowledge of other cellular measures such as physiological properties and/or transcription profile (Figure F.4) [Marblestone et al. 2014] within a single specimen. Improvements to FISSEQ will increase the speed and affordability of this technique, as has been the case for other DNA sequencing platforms.

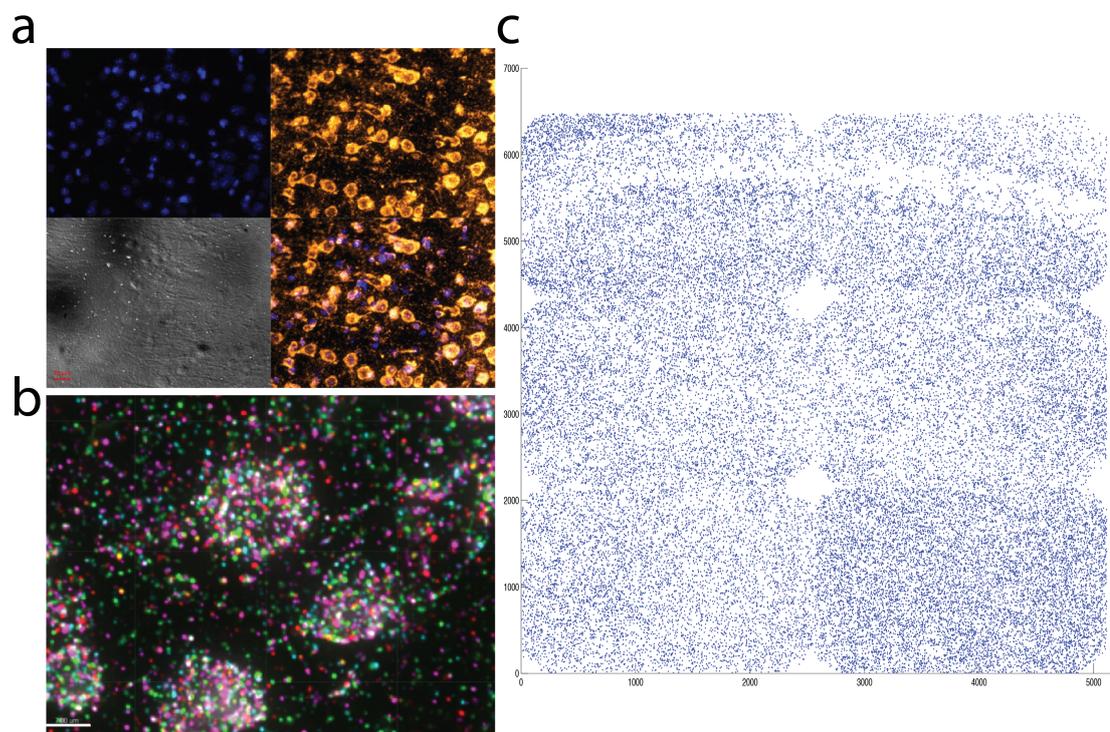


Figure F.4: FISSEQ of endogenous transcripts in a coronal slice - (a) Rolonies probed with a fluorescent oligonucleotide. (b) Rolonies after one cycle of sequencing-by-ligation. Each color in colorspace is later converted to bases. (c) Rolonies which mapped to the mouse genome.

F.5 Supplementary Materials

F.5.1 Injections

Stereotactic injections of barcoded Sindbis virus were performed to introduce the components into the brain. All injections were performed in 5-week old male CBA/CaJ mice in left auditory cortex or bilaterally under anesthesia (ketamine medetomidine). Following injections, mice were returned to their cages for 48 hours before sacrificing.

F.5.2 Cryoslicing

Brains were prepared for cryoslicing as previously described [Pinskiy et al. 2013]. Mice were sacrificed 48 hours post-infection and the brain was fixed via intracardial perfusion with cold saline for 5 minutes followed by cold 4% paraformaldehyde (PFA) for 10 minutes. Brains were removed and post-fixed overnight at 4°C in 4% PFA with 10% sucrose. The following day, the brain was placed in PBS + 20% sucrose for 24 hours at 4°C. Brains were flash frozen into cubes of NEG50 (Richard-Allan Scientific) and then sliced on a cryostat (Leica) at a thickness of 10 μ m. Slices were transferred to cover-glass coated with Solution A and Solution B (Leica CryoJane) by a tape-transfer system (Leica) and affixed to the slide by brief UV curing (Leica).

F.5.3 FISSEQ

FISSEQ was performed as previously described [Lee et al. 2014].

F.5.3.1 FISSEQ of endogenous transcripts

A 200 μ L mixture containing 4,000U M-MuLV reverse transcriptase (Enzymatics), 250 μ M dNTP (Enzymatics), 40 μ M aminoallyl dUTP (Anaspec), 50U RNase inhibitor (Enzymatics), and 100 pmol tagged random hexamers (/5PhosTCTCGGGAACGCT-GAAGANNNNNN), prepared on ice is added to cells at 25°C for 10 minutes. The concentration of aminoallyl dUTP can vary depending the cell type and the application. Generally, a high incorporation rate of aminoallyl dUTP results in better cross-linking and reduced cDNA diffusion but a lower amplicon density. The sample is then incubated overnight in a humidified 37°C chamber. The sample is washed using 1X PBS and cross-linked using BS(PEG)9 (Thermo Scientific), diluted to 50mM in PBS, for 1

hour at 25°C. 1M Tris (G Biosciences) is added to quench the reaction for 30 minutes at 25°C. A mixture of DNase-free RNases (Roche Diagnostics) and RNase H (Enzymatics) is added to degrade residual RNA for 1 hour at 37°C. A 100µL circularization reaction mixture (1X reaction buffer, 2.5mM MnCl₂, 1M Betaine and 5µL CircLigase II from Illumina/Epicentre) is then added to the sample well and incubated at 60°C for 2 hours. After circularization, the sample is washed using H₂O and incubated with a 200µL mixture containing 0.1µM rolling circle amplification primer (TCTTCAGCGTTCCCGA*G*A from IDT) in 2X SSC and 30% formamide for 15 minutes at 60°C. The sample is washed using 2X SSC, and a 200µL amplification mixture containing 500U Phi29 DNA polymerase (Enzymatics), 250µM dNTP, and 40µM aminoallyl dUTP is added. The sample is incubated in a dry 30°C chamber overnight and cross-linked using BS(PEG)9 diluted to 50mM in PBS for 1 hour at 25°C. After a rinse with PBS, 1M Tris is added to quench the reaction for 30 minutes. At this point, the sample can be stored in nuclease-free 1X PBS at 4°C.

Sequencing reactions, imaging, and analysis was performed as described [Lee et al. 2014].

F.5.3.2 FISSEQ of barcode transcripts

The FISSEQ reaction is identical to the above, except that a gene specific primer (/5PhosTCTCGGGAACGCTGAAGA-GGAAAGTTGGTATAAGACAAAAGTG) is used for reverse transcription.

F.5.3.3 FISSEQ Barcode RNA sequence

The FISSEQ RNA coding sequence is flanked on either side by a **NotI** restriction site and a **MluI** restriction site for barcode cloning. The intervening sequence contains: **5' FISSEQ Universal Sequence-Barcode-3' FISSEQ Universal.**

DNA SEQUENCE:

**ACGCGTGCCTGGAGCAATTCCACAACACNNNNNNNNNNNNNNNG
TGCAATCACTTTTGTCTTATACCAACTTTCGCGGCCGC**

F.6 Acknowledgements

Partha Mitra's lab helped with cryoslicing. Thank you to Vadim Pinskiy, Alexander Tolpygo, and Neil Franciotti for all of their help. Reza Kalhor (Church lab, Harvard University) performed all of the FISSEQ including imaging. I performed all injections and cloned all constructs. Justus Keschull packaged the Sindbis viruses. Thank you to Jay Lee and Evan Daugharthy (Church lab, Harvard) for helpful conversations.